T.C.
DOKUZ EYLÜL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
EKONOMETRİ ANABİLİM DALI
EKONOMETRİ PROGRAMI
DOKTORA TEZİ

MAKİNE ÖĞRENMESİNE DAYALI EKSİK VERİ TAMAMLAMA YÖNTEMLERİNİN İSTATİSTİKSEL PERFORMANSLARININ KARŞILAŞTIRILMASI ÜZERİNE BİR ARAŞTIRMA

Şemsettin ERKEN

Danışman
Prof. Dr. Rabia Ece OMAY

2024

T.C.
DOKUZ EYLÜL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
EKONOMETRİ ANABİLİM DALI
EKONOMETRİ PROGRAMI
DOKTORA TEZİ

MAKİNE ÖĞRENMESİNE DAYALI EKSİK VERİ TAMAMLAMA YÖNTEMLERİNİN İSTATİSTİKSEL PERFORMANSLARININ KARŞILAŞTIRILMASI ÜZERİNE BİR ARAŞTIRMA

Şemsettin ERKEN

Danışman
Prof. Dr. Rabia Ece OMAY

İZMİR - 2024

TEZ ONAY SAYFASI

YEMİN METNİ

Doktora Tezi olarak sunduğum “Makine Öğrenmesine Dayalı Eksik Veri Tamamlama Yöntemlerinin İstatistiksel Performanslarının Karşılaştırılması Üzerine Bir Araştırma” adlı çalışmanın, tarafımdan, akademik kurallara ve etik değerlere uygun olarak yazıldığını ve yararlandığım eserlerin kaynakçada gösterilenlerden oluştuğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve bunu onurumla doğrularım.

05.01.2024

Şemsettin ERKEN

ÖZET

Doktora Tezi

Makine Öğrenmesine Dayalı Eksik Veri Tamamlama Yöntemlerinin İstatistiksel Performanslarının Karşılaştırılması Üzerine Bir Araştırma

Şemsettin ERKEN

Dokuz Eylül Üniversitesi
Sosyal Bilimler Enstitüsü
Ekonometri Anabilim Dalı
Ekonometri Programı

Veri kavramı, insanlık tarihi boyunca yaşam için gerekli olan ihtiyaçlar kadar önemli olan bir kavramdır. İnsanlar, topladığı veriler aracılığıyla bilgi elde ederek içinde bulundukları zaman dilimindeki problemlerini çözmek amacı taşımışlardır. Aynı zamanda insanlar, verilerin arasındaki örüntü ve ilişkileri tespit ederek geleceğe yönelik çıkarımlar yapma amacı taşımaktadır. Bu noktada, bilgiye erişim sürecinde veriler, verilerin nitelikleri ve verilerin sahip olduğu özellikler anahtar niteliktedir.

Veri topluluklarından direkt olarak bilgi elde etmek çok zordur. Bilginin elde edilmesi için kullanılacak olan verilerin gerekli analize hazır olması gerekmektedir. Veri setlerinde, eksik değerlerin olması sık rastlanan problemlerden bir tanesidir. Bu problemin çözülmesi, yapılacak analizden anlamlı bilgiler ve sonuçlar üretebilmek için oldukça önemlidir. Eksik değerlerin göz ardı edildiği analizlerin sonucunda elde edilen bilgiler, gerçeği yansıtmayan ve sağlıksız bilgiler olacağından, eksik değerler uygun şartlar çerçevesinde tamamlanmalıdır. Eksik değerlerin, veri setinin istatistiksel özelliklerini yansıtan ve bu şartlara uygun bir yöntemle veya yöntemlerle tamamlanması özellikle dikkat edilmesi gereken bir durumdur.

Makine öğrenmesi, verilerden eğitim yoluyla elde edilen modeller yoluyla gelecekte karşılaşılabilecek olan durumlara yönelik tahminler üreten bilgisayar teknolojisi ve yapay zekâya dayalı bir disiplindir. Makine öğrenmesi

algoritmaları kullanılarak sınıflandırma, regresyon ve kümeleme gibi problemlere çözüm bulunmaktadır.

Tez çalışmasında, Hitters veri setindeki veriler, manipüle edilerek %5, %10 ve %15 oranlarında rastgele eksiltilmiştir. Eksiltilen değerler, temel eksik eksik veri tamamlama yöntemleri ve makine öğrenmesi algoritmaları ile tamamlanmıştır. Temel eksik veri tamamlama yöntemi olarak Liste Boyunca Silme, Son Gözlemi İleri Taşıma ve Ortalama İle Tamamlama yöntemleri tercih edilmiştir. Diğer taraftan, makine öğrenmesi algoritmalarından En Yakın k-Komşu algoritması, Rassal Orman algoritması, Stokastik Regresyon ve Amelia algoritması kullanılmıştır. Böylece, Hitters veri setinden, bir anlamda kullanılan yöntem sayısı kadar yeni veri seti türetilmiştir. Bu yeni veri setlerinin ve verinin orijinal halinin, Naive Bayes algoritması ile sınıflandırılmasının ardından sınıflandırma sonuçları, performans değerlendirme ölçütlerine göre karşılaştırılmıştır.

Belirtilen şekilde gerçekleştirilen uygulama sonucunda, makine öğrenmesi algoritmalarının temel eksik veri tamamlamam yöntemlerine göre üstün bir eksik veri tamamlama ve sınıflandırma performansı gösterdiği sonucu elde edilmiştir. Özellikle, En Yakın k-Komşu algoritması ve Rassal Orman algoritmalarının dikkat çekici performanslar gösterdiği tespit edilmiştir.

Anahtar Kelimeler: Eksik Veri, Eksik Veri Tamamlama, Makine Öğrenmesi, Sınıflandırma, Veri Analizi.

ABSTRACT

Doctoral Thesis

Doctor of Philosophy(PhD)

A Research On Comparison Of Statistical Performance Of Missing Data Imputation Methods Based On Machine Learning

Şemsettin ERKEN

Dokuz Eylül University

Graduate School of Social Sciences

Department of Econometrics

Econometrics Program

The concept of data has been as important as the necessities of life throughout human history. People have used data to solve problems and make inferences about the future by identifying patterns and relationships. At this point, data, data qualities, and characteristics are key in the information access process.

Obtaining information directly from data can be challenging. The data must be properly prepared to perform the required analysis in order to obtain information. A common problem with data sets is missing values, which must be addressed in order to produce meaningful results. It is crucial to solve this problem in order to obtain accurate information and results from the analysis. Missing values should be imputed under appropriate conditions, since the information obtained as a result of analyses in which missing values are ignored will be unrealistic and unhealthy information. It is important to pay particular attention to imputing missing values using a method or methods that reflect the statistical characteristics of the data set and are appropriate for these conditions.

Machine learning is a discipline of computer technology and artificial intelligence that uses models trained on data to make predictions about future situations. It offers solutions to problems such as classification, regression, and clustering by using machine learning algorithms.

In this thesis, the data in the Hitters dataset were manipulated and randomly removed by 5%, 10% and 15% of its data. Missing values are imputed using basic missing data imputation methods and machine learning algorithms. Listwise Deletion, Last Observation Carried Forward, and Mean Imputation are the preferred basic missing data imputation methods. On the other hand, various machine learning algorithms were employed, including the k-Nearest Neighbor, Random Forest, Stochastic Regression, and Amelia algorithms. Thus, new datasets were derived from the Hitters dataset, in a sense, as many as the number of methods used. After classifying these new datasets and the original data by Naive Bayes algorithm, the classification results were compared and evaluated based on performance criterias.

As a result of the application, it was concluded that machine learning algorithms outperformed basic missing data imputation methods in terms of missing data imputation and classification performance. In particular, the k-Nearest Neighbor and Random Forest algorithms have shown noteworthy performance.

Keywords: Missing Data, Missing Data Imputation, Machine Learning, Classification, Data Analysis.

MAKİNE ÖĞRENMESİNE DAYALI EKSİK VERİ TAMAMLAMA YÖNTEMLERİNİN İSTATİSTİKSEL PERFORMANSLARININ KARŞILAŞTIRILMASI ÜZERİNE BİR ARAŞTIRMA

İÇİNDEKİLER

İKİNCİ BÖLÜM
MAKİNE ÖĞRENMESİ VE MAKİNE ÖĞRENMESİ ALGORİTMALARININ EKSİK VERİ TAMAMLAMADA KULLANIMI

ÜÇÜNCÜ BÖLÜM
PERFORMANS ÖLÇÜTLERİ VE PERFORMANSLARIN DEĞERLENDİRİLMESİ

DÖRDÜNCÜ BÖLÜM
UYGULAMA

KISALTMALAR

ARFF	Nitelik İlişkisi Dosya Formatı(Attribute Relationship File Format)
CRAN	Kapsamlı R Arşiv Ağı (Comprehensive R Archive Network)
FN	Yanlış Negatif (False Negative)
FP	Yanlış Pozitif (False Positive)
KNN	En Yakın k-Komşu Algoritması (K- Nearest Neighbors Algorithm)
MAP	En Büyük Sonrasal Sınıflandırma (Maximum A Posteriori Classification)
MAR	Rassal Eksiklik (Missing At Random)
MCAR	Tamamen Rassal Eksiklik (Missing Completely At Random)
MNAR	Rassal Olmayan Eksiklik (Missing Not At Random)
RF	Rassal Orman Algoritması (Random Forest Algorithm)
SR	Stokastik Regresyon (Stochastic Regression)
TN	Gerçek Negatif (True Negative)
TP	Gerçek Pozitif (True Positive)
UCI	California Üniversitesi, Irvine (University of California, Irvine)
WEKA	Bilgi Analizi İçin Waikato Ortamı (Waikato Environment for Knowledge Analysis)

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

GİRİŞ

Veri kavramı, bilginin kaynağı olması sebebiyle oldukça önemli bir kavramdır. İnsanlık, doğuşundan itibaren farklı şekil ve türlere sahip verilerden elde ettiği bilgileri, yaşamına entegre ederek hayatta kalmış, ihtiyaçlarını karşılamış ve geleceğe yönelik olay ve durumlara ilişkin yeni bilgilere ulaşmaya çalışmıştır. Veri denildiğinde akla ilk başta gelen, bilim çerçevesinde ve bilimsel nitelikli çalışmalarda ilgilenilen konuya ilişkin çekirdek pozisyonunda yer alan kaynaktır. Fakat veri kavramı, hayatın her alanında yer alan, bilgi sağlayan ve hayatın devam etmesi sürecindeki bir sonraki ana yönelik tahminlere imkân veren daha geniş bir çerçeveye sahip ve bu şekilde değerlendirilmesi gereken bir kavramdır.

Veriler, yapısı gereği direkt ve doğrudan bilgiyi sağlayan bir kavram değildir. Bilginin verilerden elde edilmesi diğer bir ifadeyle ortaya çıkarılmalıdır. Bu anlamda, veri kaydedilen, depolanan sembol ya da sinyal okumalarıdır şeklinde en temel haliyle kavramsal olarak tanımlanabilir (Liew, 2007: 1-3). Verinin bilgiye dönüşmesi bir anlamlandırmaya dayalı bir süreçtir. Veriden bilgi elde etme sürecinde bazı kavramlardan yararlanılmaktadır. Bunlardan bir tanesi, tez çalışması içerisinde de yararlanılacak olan makine öğrenmesi kavramıdır. Makine öğrenmesi, veriler arasındaki örüntüden ve ilişkilerden yararlanarak verilerin taşıdığı bilgiyi ortaya çıkarmak adına çeşitli yöntem, algoritma ve yazılımlardan yararlanan bir disiplindir (Gollapudi, 2016: 3). Veriden bilgi çıkarımı için oldukça büyük öneme sahip veri bilimi kavramından da yararlanılmaktadır.

Veri ile ilgili işleri kapsayan işlemlere tarihsel bir süreç olarak bakıldığında, gözler önüne 17. yüzyılda, İngiltere'de yapılmış olan veri çalışmaları gelse de modern anlamda değerlendirildiğinde 19. yüzyılın son çeyreğine rastlanmaktadır. Bu tarihlerin ön plana çıkmasının nedeni yeni birtakım teknik ve teknolojik adımların ortaya çıkmasıdır. Bunlardan belki de en önemlisi bilgisayarın icadıdır (Gürsakal, 2019: 19). Bilgisayarın icadı ve en önemlisi hayatın fark alanlarına doğru bilgisayar kullanımın artması, toplumların yaşamını şekillendirdiği gibi yeni formların da ortaya çıkmasını sağlamıştır. Bu değişim veri bilimi adına da yaşanmıştır.

Gelişim ve değişim kavramının doğasındaki en önemli özellik durağan olmamasıdır. Bilgisayar kavramının diğer alanda yarattığı etki, bilgisayar

teknolojisinin kendisinde de aynı etkiyi göstermiştir. Teknolojinin ve sunduğu imkânların gelişmesi devam ederken, veri bilimi alanında yaşanan gelişimin sürmesi elbette beklenen bir sonuç olarak karşımıza çıkmaktadır. Özellikle internet kavramının ortaya çıkışı ile bu gelişim hızlanmıştır. Günümüzde de yapay zekâ, elektronik ortamlar ve cloud (bulut) teknolojileri sayesinde bu gelişimin etkileri hala sürmektedir. Bu gelişimin asla durmayacağı konusu bizzat teknolojinin özellikle kendi yapısı gereği olağandır. Bahsedilen tarihsel süreçteki gelişim, kendisinden bilgi üretilecek olan verilerin depolanması, iletilmesi, korunması ve verilere erişim anlamında çok büyük olanaklar sağmıştır (Rao ve Selvamani, 2015: 204).

Bahsedilen gelişim ve teknoloji bütünleşmesinin sağladığı olanaklar ve avantajlar hayatımızı oldukça kolaylaştırmaktadır. Hayatımızın en basit ve gündelik durumlarından en detaylı yapısal durumlara kadar her alanında bir destek unsuru olarak yer almaktadır. Diğer taraftan, artan dünya nüfusu, var olan ve yeni ortaya çıkan ihtiyaçlar, endüstriyel ya da sosyal hayata ilişkin milyarca detay ve her geçen gün bunlara bağlı artan dokular hayal edilemeyecek büyüklükte veri üretimini beraberinde getirmektedir. Bu ilk bakışta, olumsuz bir durum olarak görünmemektedir. Çünkü olumsuz olan tarafı veri üretimi kısmı değil, mevcut teknolojik olanaklarla birlikte saklanan ya da elde edilen verinin bilgiye dönüşmesi yani anlamlandırılmasıdır. Bu durumda, elde edilen verilerin belki de çok ama çok küçük bir kısmının anlaşılabildiği ortaya çıkmaktadır. Tıpkı tarihsel süreç olarak ele alınan teknolojinin gelişimine bağlı olarak ortaya çıkan veri bilimi atılımları gibi veri üretim hızı da her geçen katlanmaktadır. Dolayısıyla, veri üretim hızı ile bilgi elde edebilme hızı veya potansiyeli arasındaki farkın her geçen gün daha da arttığı gerçeği açık bir şekilde ortaya çıkmaktadır. Bu konuyu, üretilen verileri okyanusa benzetecek olursak bahsedilen hız farkının okyanustaki dalgaları izlemeye benzediği şeklinde ifade edebiliriz (Pyle, 1999: 2).

Bilimsel araştırmalarda, veriler analiz edilerek elde edilen bilgiyle amaçlanan operasyonlar gerçekleştirilmektedir. Sağlıklı analiz yapılabilmesi için öncelikle verinin analize hazır olması gereklidir. Analize hazır halde olmayan verilerden elde edilen bilgiler, çarpık ve yanlış niteliğe sahip bilgiler olduğundan, analiz adımından önce verilerde var olan problemler ortadan kaldırılmalıdır (Famili ve diğerleri, 1997: 4). Veri önişleme adımı verilen bu kısımda bilgi çıkarımı yapılacak veriler üzerinde

bir dizi işlem yaparak veriler üzerindeki problemlerin ortaya kaldırılması gerekmektedir ve hatta literatürde, veri önişleme işlemlerinin, sonradan yapılacak olan operasyonları etkilediği konusunda çalışmalar mevcuttur (Zhu ve Gao, 2016: 56).

Veri topluluklarında sık olarak gözlenen sorunlardan birisi, veri setlerinde eksik değerlerin bulunmasıdır. Bahsedilen eksik değerler, elbette ki farklı durumlardan kaynaklı farklı şekillerde meydana gelmektedir. Bu durumda, eksik değerler, tamamen rassal, rassal ve rassal olmayan eksik veri tipi olarak kategorilere ayrılmaktadır (Van der Heijden ve diğerleri, 2006: 1103). Veri setlerinde, farklı sebeplerden kaynaklı ortaya çıkabilen eksik değer sorununun giderilmesi eksikliğin türüne göre fark şekilde ele alınmaktadır. Farklı durumlara göre eksik değer sorunu, farklı yöntemlerle tamamlanarak veya eksikliği giderecek şekilde yeniden veri toplayarak giderilmektedir. Eksik verilerin tamamlanması, gerçekçi ve anlamlı sonuçların elde edilmesi için anahtar niteliktedir.

Tamamlanırken eksik değerlerin var olduğu veri setinin karakteristik ve istatistiksel özelliklerini yansıtacak şekilde tamamlanması oldukça önemlidir. Çünkü bu durum hem veri setinin analiz için hazır hale gelmesinde hem de bu veriler yoluyla elde edilecek bilginin gerçek ve nitelikli bilgi olmasında en etkili aşama içerisinde yer almaktadır. Bu nedenle, eksik verilerin hangi değerle nasıl tamamlanacağı ve eksik verinin oranı ya da niteliği çok önemlidir. Eksik olan veriler tamamlanırken tamamlanan değerlerin, veri setinin karakteristiğini bozmayan ve veri bütünlüğüyle uyum sağlayan değerler olması ve ifade edilen kriterlerin göz ardı edilmemesi gerekmektedir (Aittokallio, 2010: 253-254). Bu düşüncenin dışında bir mantıkla eksik veriler tamamlanırsa, bahsedilen uyum ve veri karakteristiğinden uzak bir yapı söz konusu olmaktadır. Adeta mevcut veri setinden farklı özelliklere sahip başka bir veri seti ile çalışılması benzeri bir durum ile karşılaşılabilir. Bu sorunun çözümlenmesi için birçok yöntem bulunmaktadır. Eksik verilerin tamamlanmasında, temel yöntemlerin yanı sıra bilgi ve bilgisayar teknolojisinin gelişmesiyle birlikte her geçen gün daha geniş bir kullanım alanına sahip olan makine öğrenmesi yöntemleri de kullanılmaktadır. Bahsedilen yöntemlerle eksik verilerin tamamlanmasının ardından yapılan tamamlama operasyonunun başarısını diğer deyişle performansını değerlendirmek, gelecekte böyle bir sorun ile karşılaşıldığında, hangi kısıt ve

durumlarda nasıl bir eksik veri tamamlama yaklaşımı ve yöntemi kullanılmasının tespiti açısından önemli olan bir diğer konudur (Bertsimas ve diğerleri, 2018: 1-2).

Bu kısımda bahsedilen veri, veri bilimi ve bilgi çıkarımı konusunda ifade edilen durumlarda olduğu gibi insanlık tarihi boyunca gelişen ve değişen sayısız durum ve unsur olsa da değişmeyen az sayıdaki şeylerden biri de insanların bilme isteği ve ihtiyacıdır. İnsanlık tarihinin başlangıcından bu yana, insanlık her zaman geleceği bilmek veya bu konuda öngörüye sahip olmak istemiştir. Tarihte yer almış özellikle kabile toplumlarına bakıldığında, hala günümüzde az sayıda da olsa bu toplumlardan hala var olanlar söz konusudur, bu toplumların mutlaka bir kâhin veya bilgin ismini verdikleri ve kendi toplumları içerisinde önemli bir statü sağlayıp saygı duydukları kişilerin olması durumu gözlenmiştir. Bunun sebebi bu kişilerin, geleceğe yönelik haberler vermesi yani doğruluğu konusunda hiç fikirleri olmamasıyla birlikte bilgi sağlamasıdır.

İlkel modelde olduğu gibi günümüzde de bilgiye olan istek ve ihtiyaç aynı şekildedir. Meselenin bilimsel açısı gereği bilginin, veri bilimi ve teknolojik imkânlar vasıtasıyla elde edilmesi, yakın geçmişin ve günümüzün en büyük farkıdır. Fakat görüldüğü gibi bilgiye erişim iştahı her zaman aynı şekildedir. Yakın geçmişten bu yana bilgi çıkarımı için başvurulan en önemli yöntemlerin başına makine öğrenmesi gelmektedir. Daha önce ifade edildiği ve çalışma kapsamında detaylı olarak ele alınacağı gibi teknolojik destek unsurlarından biri olan makine öğrenmesi, özellikle bilgi ve bilgisayar teknolojilerinin gelişimi ile birlikte bilgi çıkarımı konusunda oldukça ön plana çıkan çok önemli bir kavramdır. Bu çalışma da görüleceği gibi, bilgi çıkarımı konusunda olduğu gibi veri ön işleme kısmında ifade edildiği gibi eksik verilerin tamamlanmasında da kullanılmaktadır. Makine öğrenmesi türleri kendi içinde denetimli, denetimsiz, yarı denetimli ve takviyeli öğrenme şeklinde dörde ayrılmaktadır. Denetimli, denetimsiz ve yarı denetimli öğrenmeyi birbirinden ayıran farklılıklar eğitim veri seti olarak ifade edilen veri topluluğundaki verilerin etiketli yani hedef nitelik/değişken değerinin bilinip bilinmemesinden kaynaklıdır. Takviyeli öğrenmede ise durum tecrübe etme, geri bildirim ve ödül-ceza mekanizması ile ilişkilidir (Mohri ve diğerleri, 2018: 6). Denetimli öğrenme yöntemlerinden sınıflandırma kavramı da çalışma kapsamında açıklanacak ve uygulama kısmında da sınıflandırma işlemi yer alacaktır.

Tez çalışmasının amacı, veri setlerinde çok sık yaşanan eksik değer problemine yönelik kullanılan temel yöntemler ve makine öğrenmesi algoritmalarının eksik değer tamamlama performanslarının kıyaslanmasını ve sınıflandırma operasyonuna olan dolaylı etkilerinin de ortaya koyulmaya çalışılmasıdır. Bu kıyaslama, yöntemlerin kendi arasındaki karşılaştırmanın yanı sıra yöntem ve algoritmalar arasında bir karşılaştırmayı da içermektedir. Bunun yanı sıra, farklı eksiklik oranlarında, yöntem ve algoritmaların gösterdiği performansların ve reaksiyonların ortaya koyulması da amaçlanmıştır.

Çalışmanın ilk bölümünde, eksik veri kavramı ele alınmıştır. Bu tanım ışığında, eksik veri türlerine ve bu türlerin özelliklerine ilişkin bilgiler verilmiştir. Aynı bölümde, temel eksik veri tamamlama yöntemlerinin işleyişi ve detaylı bir şekilde ele alınmıştır.

İkinci bölümde ise makine öğrenmesi kavramına yer verilmiştir. Bu bölümde, makine öğrenmesine ilişkin tanımlar, makine öğrenmesindeki öğrenme türlerine değinildikten sonra makine öğrenmesi yöntemlerinden olan sınıflandırma kavramı açıklanmıştır. Aynı bölümde, çalışma kapsamında kullanılacak olan makine öğrenmesi algoritmalarına yer verilmiş ve ilgili algoritmalar detaylı bir şekilde açıklanmıştır.

Çalışmanın üçüncü bölümünde, eksik veri tamamlama yöntemlerinin ve buna bağlı uygulamaların başarı derecelerini tespit etmek ve aralarında kıyaslama yapmak için kullanılan performans değerlendirme kriterleri açıklanmıştır.

Uygulamanın gerçekleştirildiği bölüm olan dördüncü bölümde, ifade edilen amaçlar doğrultusunda, makine öğrenmesi ve veri analizi konularında oldukça sık bir kullanıma ve bilinirliğe sahip olan Hitters veri seti kullanılmıştır. Hitters veri seti, manipüle edilerek farklı oranlarda rastgele olacak şekilde eksiltilmiş ve ifade edilen amaçlar yönünde eksik veriler tamamlanmıştır. Verilerin tamamlanmasıyla elde edilen veri setleri sınıflandırmaya tabi tutularak elde edilen sonuçlar, performans değerlendirme kriterleri nispetinde değerlendirilmiştir.

LİTERATÜR TARAMASI

Makine öğrenmesi, teknolojik gelişmelerle birlikte kullanım alanı oldukça genişleyen bir kavramdır. Belirli bir problemin çözümü için modelleme veya var olan modellerin geliştirilmesi, sınıflandırma ve kümeleme analizleri, görüntü analizleri vb. konularda makine öğrenmesi kullanılarak başarılı sonuçlar elde edilmiştir.

Makine öğrenmesinin kullanıldığı durumlardan birisi de veri setlerinde var olan eksik değerlerin tamamlanmasıdır. Veri setlerinde eksik değerlerin olması ya analizin gerçekleştirilememesine neden olan ya da analiz sonuçlarına direkt olarak etki bırakan bir sorun olduğundan ele alınması çok önemli bir durumdur. Bu nedenle, eksik değerlerin tamamlanması sadece bir operasyon olarak değil, elde edilen sonucun bir anlamda kalitesini ortaya koyan bir süreç olarak değerlendirilmelidir. Eksik değerlerin tamamlanması konusunda yapılan çalışmalar, bilimsel tarih içerisinde özellik geçmişte durma noktasına kadar gelmişse de, bu alana yapılan katkılar neticesinde canlanmış ve teknolojinin gelişmesine paralel olarak gelişim göstermiştir. Bu alanda, ortaya koyulan yöntemler veya var olan yöntemlerin geliştirilmesi de oldukça etkili olurken bilgi ve bilgisayar teknolojisinde yaşanan hızlı gelişim ve yapay zekâ kavramının, her geçen gün hayatın hemen her noktasına etki etmesiyle birlikte eksik değer problemlerine karşın üretilen çözümler derin etkisini ve kapsayıcılığını arttırmaktadır.

Literatürde, sık karşılaşılan bir sorun olan bu konuda yapılan çalışmalarda, makine öğrenmesinin etkin ve yüksek doğruluğa sahip sonuçlar ortaya koyduğu gösterilmiştir. Bazı çalışmalarda, bahsedilen iyi performans ve etkileri ve performans kıyaslamaları gibi analizler söz konusu iken bazı çalışmalarda ise eksik değer sorunu çözecek ve hatta bu sorunun çözümüne bağlı olarak dolaylı yoldan diğer durumları da olumlu yönde etkileyecek yeni ve entegre yöntemler önerilmektedir. Literatürde bu konuda ele alınan bazı çalışmalar, genel hatlarıyla aşağıda ifade edilmiştir.

Abidin ve diğerleri (2018), eksik verilerin tamamlanmasında makine öğrenmesi yöntemleri kullanmıştır. Farklı veri setleri kullanarak, veri setlerindeki eksik değerler makine öğrenmesi yöntemleri tamamlanmış ve bu yöntemlerden özellikle bayes ağların etkin sonuçlar elde ettiğini göstermiştir.

Veri setlerinde, eksik veri oranı, bazı durumlarda yüksek kabul edilebilecek seviyede olmaktadır. Alamoodi ve diğerleri (2021), yüksek oranda eksik veri

setlerinde, makine öğrenmesi yöntemleri ile eksik verilerin tamamlanmasında etkin sonuçlar elde edildiğini ifade etmiştir.

Hastalık tanısının koyulabilmesi için de makine öğrenmesi yöntemlerinden faydalanılmaktadır. Jerez ve diğerleri (2010), ortaya koydukları çalışmada, eksik veri tamamlanmasında makine öğrenmesi yöntemlerinin en çok uyum gösteren sonuçları ürettiğini ve tanı doğruluğunu artırdığını tespit etmişlerdir.

Thomas ve Rajabi (2021), yaptıkları çalışmada, doğru tahmin oranı ve hata kareler ortalaması değerleri ışığında, makine öğrenmesi yöntemlerinden kümeleme ve mesafe tabanlı algoritmaların eksik veri tamamlanmasında kullanılmasını önermişlerdir.

Eksik veri tamamlama operasyonlarında, veri setindeki eksik verilerin hangi oranda olduğu da etkilidir. Emmanuel ve diğerleri (2021), veri setinin %20'sine varan oranlarda eksik veri içerdiği durumlarda da, en yakın k- komşu, MissForest ve Random Forest algoritmalarının eksik değer tamamlama performanslarının ütün olduğunu göstermişlerdir.

Takviyeli öğrenme, makine öğrenmesi içerisindeki öğrenme türlerindendir. Awan ve diğerleri (2022), çalışmalarında kullandıkları takviyeli öğrenme tabanlı eksik veri tamamlama yönteminin, diğer yöntemlere göre oldukça iyi sonuçlar verdiğini raporlamışlardır.

Biyolojik mesafeleri analiz edildiği çalışmada, eksik olan verileri ile ilgili değerlerin tamamlandığı yeni değerler arasındaki farkı en düşük şekilde veren algoritmanın, en yakın k-komşu algoritması olduğu, Kenyhercz ve Passalacqua (2016) tarafından gösterilmiştir.

Raja ve Thangavel (2020), UCI veri deposundan elde ettikleri, Dermatology, Pima, Winconsin ve Yeast veri setleri üzerinde yapmış oldukları makine öğrenmesi yöntemleri eksik veri tamamlama çalışmasında, K-ortalamalar, Bulanık C-ortalamalar ve önerdikleri Rough K-ortalamalar yöntemleri ile eksik değerleri tamamlamışlardır. Çalışma sonucunda elde edilen sonuçlarda, yöntemlerin çok iyi sonuçlar verdiği görülmekle beraber önermiş oldukları Rough K-ortalamalar yönteminin en iyi performansı çarpıcı şekilde ortaya koyduğunu saptamışlardır.

Eksik veri tamamlamada yöntemler kadar, yöntemlerin uygulanmasına olanak tanıyan, ilgili yöntemlerin başka uyumlu yöntemlere entegre edilebildiği yazılımlar da

önemlidir. Yapısal olmayan metin verileri de dahil olmak üzere homojen olmayan veri türlerine sahip veri tablolarına uygulanabilen ve eksik değer tamamlamada robust ve ölçeklenebilir sonuçlar üretmede etmede yardımcı paket program olan DataWig, Biessmann ve diğerleri (2019) tarafından ortaya koyulmuştur. Bu yazılımın bir diğer özelliği de derin öğrenme özellik çıkarımı operatörleri ile otomatik hiperparametre ayarlamayı entegre etmesidir.

Sürdürülebilirlik ve temiz çevre konuları dünya çapında önemli olan ve tüm dünya toplumlarının geleceğinde çok önemli bir yer teşkil eden konulardandır. Rodriguez ve diğerleri (2021), yaptıkları çalışmada su kalitesi üzerine bir analiz gerçekleştirmişlerdir. Eksiklik oranı %50 ile %70 arasında değişen eksik değerlere ve yüksek değişkenliğe sahip niteliklerin varlığının çalışmanın en zorlu kısıtı olduğu, araştırmacılar tarafından belirtilmiştir. Çalışma kapsamında, Adaboost, En Yakın k-Komşu Algoritması, Rassal Orman, Ridge ve Bayezyen Ridge, Ters mesafe Ağırlıklandırma ve destek vektör makinaları kullanılmıştır. Elden sonuçlara göre, tamamlanan değerlerin %76'sından fazlasının tatmin edici olduğu ve en iyi performans sahip modelin Ters Mesafe Ağırlıklandırma ve ardından Rassal Orman yaklaşımına ait olduğu ifade edilmiştir.

Gajawada ve Toshniwal (2012), kümeleme ve en yakın komşuluk tabanlı bir eksik değer tamamlama yöntemi önerdikleri çalışmada, kavramsal bir yaklaşım olarak özellikle, eksik değerlerin yüksek yani eksik değerlerin gözlemlenmiş değerlerden fazla olduğu durumlarda, tamamlanan değerlerin sonraki durumlarda karşılaşılabilecek eksik değer durumlarında gözlemlenmiş bir değer olarak kullanılması konusu üzerine eğilmişlerdir. UCI veri deposu platformundan elde ettikleri klinik veri setleri üzerinde yaptıkları analizler sonucunda, önerdikleri yöntemin diğer yöntemlere göre daha iyi sonuçlar ürettiğini ortaya koymuşlardır. Bu çalışmada, önerdikleri yöntemin daha iyi sonuçlar vermesine rağmen gerçek değerlerle olan sapmaları dikkat edilmesi gerektiğini vurgulamışlardır. Bunun nedeni, nispeten iyi sonuçlar üretilmesinde rağmen sapmaların da fazla olduğu durumlarda karşılaşılabilecek yeni eksik değer problemlerinde yeni eksik değerlerin, tamamlanmış olan değerler ışığında tamamlanacak olmasıdır.

Tsai ve Hu (2022), çalışan veri setlerinde eksik değerlerin bulunması sorununun sık karşılaşılan bir sorun olmasından kaynaklı, bu probleme etkili çözüm

üreten yaklaşımların tespiti için yaptıkları çalışmada, Çok Katmanlı Algılayıcı Sinir Ağı, Naive Bayes, Sınıflandırma ve Regresyon Ağaçları ve Destek Vektör Makinaları yöntemlerini eksik değerlerin tamamlanmasında kullanmışlardır. Çalışma sonucunda, kategorik veriler için en iyi sonucu Sınıflandırma ve Regresyon Ağaçlarının ortaya koyduğunu, diğer taraftan nümerik ve karışık veri tipleri için ise Çok Katmanlı Algılayıcı Sinir Ağının en iyi performans gösterdiğini ortaya koymuşlardır.

Silva-Ramirez ve diğerlerinin (2011), tamamen rassal olan eksik veri tipleri üzerine yaptıkları çalışmada, 47'den 1389 adetlik kayıtlarına varan bir çeşitliliğe sahip 15 adet gerçek ve simüle edilmiş veri seti kullanılmışlardır. Eksik değer oranının %5 olarak ayarlandığı çalışmalarında araştırmacılar, çeşitli Çok Katmanlı Algılayıcı Sinir Ağı yaklaşımını test edip 3 adet klasik eksik veri tamamlama yapısı olan regresyon, mod/medyan atama ve Hot-Deck atama ile kıyaslamışlardır. Çalışma neticesinde elde edilen bulgulara dayanarak kategorik değişkenlere sahip veri setlerinde, Çok Katmanlı Algılayıcı Sinir Ağının oldukça iyi bir performans sergilediğini ve eksik değerlere sahip ver setinin kalitesini arttırdığını saptamışlardır.

Poulos ve Valle (2018), eksik değer tamamlama yöntemlerinin, eksik veri karışıklığı durumunda tahmin doğruluğunu artırabildiğini ve sınıflandırıcıyı düzenleyerek tahmin doğruluğunu gözle görülür şekilde artırabildiğini ifade ettikleri çalışmada makine öğrenmesi çalışmalarında oldukça geniş bir kullanıma ve bilinirliğe sahip "Adult" ve "CRV" veri seti üzerine bir analiz gerçekleştirmişlerdir. En Yakın k-komşu algoritmasının kullanılan diğer yöntem olan Yapay Sinir Ağından daha iyi sonuçlar ürettiğini göstermişlerdir.

BİRİNCİ BÖLÜM
EKSİK VERİ KAVRAMI VE EKSİK VERİ TAMAMLAMADA KULLANILAN TEMEL YÖNTEMLER

1.1. EKSİK VERİ TANIMI

Eksik veri, istatistiksel analizlerde özellikle incelenmesi gereken bir kavramdır. Eksik veriler, analizlerin sonucunda elde edilecek çıktılarda önemli etkilere sebep olabilmektedir. Eksik veri, kayıp veri veya başka bir ifade ile eksik değer, bir veri setindeki değişkenlere ait belirli gözlemlerin veya kayıtların, bir değere sahip olmaması veya boş olduğu durum olarak tanımlanabilir. Doğal olarak, buradaki eksikliklerin analiz sonuçlarına da etkileyeceğini öngörmek oldukça berraktır (Newman, 2014: 372).

Eksik veriler, farklı sebeplerle ortaya çıkmaktadır. Örneğin, eksik veriler rastgele ortaya çıkabilmektedir. Tamamen rassal bir sürecin sonucunda ilgili kayıt ya da kayıtlarla ilgili değerler eksik olabilir. Eksik veriler, verilerin toplanması sürecindeki işlemler veya hatalardan kaynaklı da ortaya çıkabilir. Bunun dışında, hatalı veya eksik veri girişi, veri toplanması için kullanılan ekipmanlardan kaynaklı sorunlar veya kayıp birtakım dosyalar nedeniyle de eksik veriler ortaya çıkabilir. Farklı bir sebep olarak, var olan değerlerin anlamsız veya belirsiz olması da eksik veri kaynağı olarak değerlendirilebilir (Garcia ve diğerleri, 2015: 59). Bu sebeplerin dışında bazı özel sebeplerden dolayı eksik veri ortaya çıkmaktadır. Siyasi, ekonomik veya sosyal nitelik taşıyan bazı kritik verilerin bazı özel kuruluşlarca veya devlet kurumları tarafından bilinç olarak yayınlanmaması veya raporlanmaması bu duruma örnek olarak gösterilebilir (Acock, 2005: 1012).

1.2. EKSİK VERİ KAVRAMININ TARİHSEL GELİŞİMİ

Eksik veri kavramı, istatistik ve veri bilimi literatüründe özellikle incelenmesi gereken kavramlardandır. İstatistik ve veri bilimi literatürü incelendiğinde, eksik verinin dikkat çekici şekilde ele alınmasının 1970'li yıllara rastladığı gözlenmektedir. Eksik veri ile ilgili tartışmaların, istatistik dünyası içerisinde diğer gelişmelere göre

daha geç başlamış olması, birtakım problemler de doğurmuştur. Eksik veri kavramının ele alınmasında yaşanan bu zamansal olgu, bu kavramın mevcut istatistik yazılımlarına da geç entegre olması sonucunu doğurmuştur. Böylece, eksik veri ve eksik veri yapılarına dayanan yöntemlerin beklenilen seviyede olmaması gerçeğinin nedenine ilişkin bir kanıt niteliğindedir (İnan, 2019: 7).

Eksik veri çalışmalarına bakıldığında, Rubin (1976)'in eksik veri yapılarını kategorize etmesi oldukça dikkat çekmiştir. Bu çalışma, ilgili konuda bir kilometre taşı olarak değerlendirilmektedir. Ardından Little ve Rubin (1983) yaptıkları eksik veri çalışmasıyla, eksik veri yapılarının incelendiği konularda oldukça önemli bir kaynak niteliğine kavuşmuşlardır. Bu dönemi izleyen yılların ardından, bir anlamda eksik değer konusuna öncelik eden bu çalışmaları izleyen birçok çalışma bilimsel literatüre kazandırılmıştır. Yine o dönemki çalışmalarla başlayan sürece ilişkin özellikle temel teşkil eden Little ve Rubin (2002) gibi teorik çalışmalar yakın geçmiş ve günümüzde de güncellenmektedir.

İlk defa ilkesel anlamda eksik verinin ele alınması, Orchard ve Woodbury (1972)'nin tüm veri setini, tam ve tamamlanmamış olarak ayırarak ele aldıkları çalışmada gözlenmektedir. Bu çalışmada, hesaplanacak parametreler için eşitlikler ortaya koyulmuş ve yanı sıra eksik değerlerin tamamlanması için de olabilirlik fonksiyonu kullanılmıştır. Böylece, maksimumu verecek noktalara ait değerleri eksik değerler yerine tamamlamışlardır. Bu uygulama için Dempster ve diğerleri (1977), beklenti maksimizasyonu algoritmasının ilk teorisi olduğunu belirtmişlerdir (İnan, 2019: 8).

1.3. EKSİK VERİ TÜRLERİ

Veri setinde eksik değerlerin bulunması ihtimali azımsanmayacak seviyelerdedir. Bazı durumlarda, veri setlerindeki eksik veriler, tüm veri seti içerisinde göz ardı edilebilecek büyüklükte olurken bazen de veri setinin karakteristik özelliklerinin yansıtılmasına ve etkin sonuçlar elde edilmesine engel teşkil edecek boyutta olmaktadır (Chhabra ve diğerleri, 2017: 2). Bu durumda, yapılması gereken analizin sağlıklı ve etkin sonuçlar üretmesi adına, veri setlerinde yer alan eksik veri sorununun ortadan kaldırılmasıdır. Eksik veriler daha önce ifade edildiği gibi farklı

nedenlerle ortaya çıkmaktadır. Böylece eksik veriler üç farklı türler halinde incelenmektedir. Bu türler missing completely at random (MCAR); tamamen rastgele eksik veri, missing at random (MAR); rastgele eksik veri ve missing not at random (MNAR); rastgele olmayan eksik veri olarak literatürde yer almaktadır (Rawal ve diğerleri, 2017: 34).

1.3.1. Tamamen Rastgele Eksik Veri

X ve Y değişkenleri bir veri setinde, cinsiyet ve yaş gibi birbirinden bağımsız değişkenler olsun. X ve Y'nin gözlenmesi ihtimali, birbirini etkilemeyen ve etkilenmeyen değişkenler olduğu durumda, X veya Y değişkeninin eksik veriye sahip olması tamamen rastgele eksik veri durumunu meydana getirmektedir (Donders ve diğerleri, 2006: 1088).

Böyle bir durumda $P(X|Y) = P(X)$, $P(Y|X) = P(Y)$ olur. X değişkenini ele aldığımızda, X değerleri Y'den etkilenmediği için ilgili değişkenin eksik verileri de Y'den etkilenmeyecektir (Kwak ve Kim, 2017: 408). Böylece $P(X_{eksik}| Y) = P(X_{eksik})$, $P(Y_{eksik}| Y) = P(Y_{eksik})$ olur. Bu ifadeye göre, X değişkenine ilişkin bir verinin eksik olma olasılığı X ve Y değişkenlerinin değerlerinden bağımsızdır. Hem değişkenin kendi değerleri ile bir kovaryans yok hem de diğer değişken ile bir ilişkisi yoktur.

1.3.2. Rastgele Eksik Veri

Rastgele eksik, herhangi bir verinin eksik olma olasılığının kaybın bulunduğu değişkenin değerinden bağımsız, fakat araştırma kapsamında ölçülen diğer bir değişkene (ya da değişkenlere) bağımlı olması durumudur (Farhangfar ve diğerleri, 2008: 3695). $P(X|Y) \neq P(X)$ şeklinde ifade edilebilecek olan bu durum, doğal olarak $P(X_{eksik}|Y)$ olasılığı ile X değişkenindeki bir verinin de eksik olma olasılığının Y değişkeninin değerlerine bağlı olduğu anlamına gelmektedir. Burada değişkenin kendi değerleri ile bir kovaryans yoktur fakat diğer değişken ile bir ilişkisi vardır.

1.3.3. Rastgele Olmayan Eksik Veri

Herhangi bir değişkendeki eksik veri ihtimali, değişkenin kendisi ile ilgili olduğu halde diğer değişkenlerle alakalı değilse bu şekildeki eksik veri mekanizmasına rastgele olmayan eksik veri denir. Rastgele olmayan eksik, herhangi bir verinin eksik olma olasılığının kaybın bulunduğu değişkenin değerine bağlı olması durumudur. Bu durum, $P(X|X_{i-k}) \neq 0$, $Cov(X|X_{i-k}) \neq 0$, $\rho(X|X_{i-k}) \neq 0$ şeklinde ifade edilebilir. Bu ifadenin doğal sonucu olarak $P(X_{eksik(i)|}X_{i-k}) = P(X_{eksik}|X)$, X değişkenindeki bir verinin eksik olma olasılığının kendi değerine bağlı olduğu anlamına gelmektedir. Burada ise değişkenin kendi değerleri ile bir kovaryans varken diğer değişken ile bir ilişkisi yoktur (Enders, 2022: 11).

1.4. TEMEL EKSİK VERİ TAMAMLAMA YÖNTEMLERİ

Eksik verilerin tamamlanmasında birçok yöntem kullanılmaktadır. Bu yöntemlerin, eksik veri tamamlama operasyonu içerisinde eksikliği tamamlama mantıkları yöntemlerin farklılıklarını oluşturmaktadır. Literatürde, sayı anlamında, oldukça zengin bir yöntem içeriği mevcuttur. Bu kısımda, sadece bu tez çalışmasında kullanılacak olan yöntemler ele alınacaktır.

1.4.1. Liste Boyunca Silme Yöntemi

Literatürdeki, eksik değer ile karşılaşılan durumlarda çözüm yöntemi olarak en çok ele alnan yöntemlere bakıldığında, eksiklik oluşturan kayıtların silinmesini işaret eden yaklaşımlara rastlanmaktadır (Van Buuren, 2018: 6). Liste boyunca silme yönteminde, veri setinde bulunan eksik değerlerin bulunduğu veri kayıtlarının tamamının veri setinden çıkarılması söz konusudur.

Tablo 1: Örnek Veri Seti ve Eksik Değerler

Kayıtlar	Değişken 1	Değişken 2	Değişken 3
1	12	1	3
2	15	10	6
3	13	-	-
4	-	9	-
5	19	4	5
7	-	-	3
8	12	8	1

Kaynak: Yazar tarafından oluşturulmuştur.

Yöntemin, bu özelliğinden dolayı kullanım açısından bir kolaylığa sahip olduğu ortadadır. Bu nedenle, belli bir kullanım sıklığına sahiptir. Fakat özellikle, veri setinde bulunan eksik değerlerin bulunduğu kayıtların, veri setinde dikkat çekecek bir fazlalıkta ve oranda olduğu durumlarda liste boyunca silme yöntemini kullanmak çeşitli problemlere neden olabilmektedir. Bahsedilen türdeki kayıt sayısının fazla olduğu bir veri setinde, bu yöntemi kullanarak ilgili kayıtların veri setinden silinmesi, veri setinin karakteristik özelliklerini ortadan kaldırabilmekte, sapmalı ve etkin olmayan çıktılar üretebilmektedir (Myers, 2011: 300).

Tablo 2: Liste Boyunca Silme Yöntemi Sonrası Veri Seti

Kayıtlar	Değişken 1	Değişken 2	Değişken 3
1	12	1	3
2	15	10	6
5	19	4	5
8	12	8	1

Kaynak: Yazar tarafından oluşturulmuştur.

1.4.2. Son Gözlemi İleri Taşıma

Bu yöntemde, veri setindeki eksik değerin, kendisinden önceki gözlenmiş olan değerle tamamlanması operasyonu söz konusudur. Uzun bir süreci gerektiren klinik çalışmalarında kolay bir kullanım sunmasının, uygulamada bir karşılığı olmasına rağmen yöntemin ürettiği sonuçların etkinliği konusunda farklı görüşler mevcuttur (Kenward ve Molenberghs, 2009: 875). Örnek veri seti Tablo 3'te kullanılmıştır.

Tablo 3: Son Gözlemi İleri Taşıma Yöntemi Sonrası Veri Seti

Kayıtlar	Değişken 1	Değişken 2	Değişken 3
1	12	1	3
2	15	10	6
3	13	**10**	**6**
4	**13**	9	**6**
5	19	4	5
7	**19**	**4**	3
8	12	8	1

Kaynak: Yazar tarafından oluşturulmuştur.

1.4.3. Ortalama İle Tamamlama

Veri setinde yer alan eksik değerlerin, gözlenmiş değerlerin ortalaması ile tamamlanmasına ortalama ile tamamlama yöntemi denir. Veri setindeki eksik değer nedeniyle, ilgili kayıtların silinmesi gibi bir durumu gerektirmeyen ve dolayısıyla veri kaybına yola açmayan tekli değer atama yöntemlerindendir (Patrician, 2002: 79). Örnek veri seti Tablo 4'te kullanılmıştır.

Tablo 4: Veri Setindeki Eksik Değerlerin Ortalama İle Tamamlanması

Kayıtlar	Değişken 1	Değişken 2	Değişken 3
1	12	1	3
2	15	10	6
3	13	**6,4**	**3,6**
4	**14,2**	9	**3,6**
5	19	4	5
7	**14,2**	**6,4**	3
8	12	8	1

Kaynak: Yazar tarafından oluşturulmuştur.

Değişkenler arasındaki herhangi bir korelasyon ilişkisinden de yararlanılmamaktadır. Burada, tüm eksik veri tamamlama yöntemlerinde olduğu gibi eksik veri miktarı ve kayıt sayısı önemli durumdadır. Fazla sayıda eksik değer söz konusu olursa, hepsinin ortalama ile tamamlanmasından kaynaklı ilgi değişkenin dağılımının şekli değişebilmektedir. Veri setindeki değerlerin ortalamasının alınması, bazı durumlarda ilgili değerlerdeki değişkenliğin sıkıştırılması ile sonuçlanabilmektedir (Jadhav ve diğerleri, 2019: 917).

İKİNCİ BÖLÜM
MAKİNE ÖĞRENMESİ VE MAKİNE ÖĞRENMESİ ALGORİTMALARININ EKSİK VERİ TAMAMLAMADA KULLANIMI

2.1. MAKİNE ÖĞRENMESİ KAVRAMI VE TANIMI

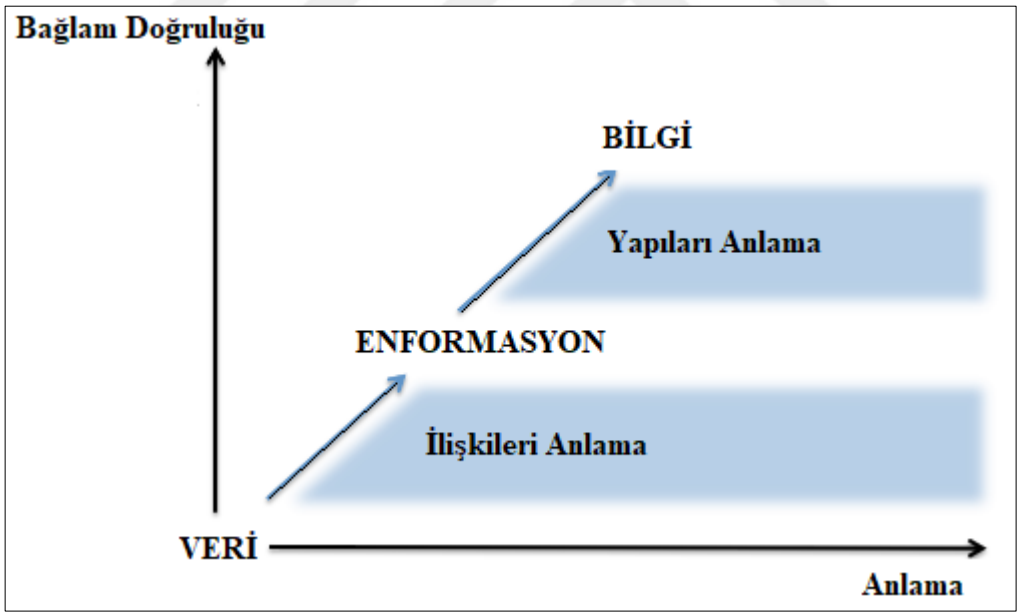
Makine öğrenmesi kavramındaki en önemli kısmın öğrenme olduğu düşüncesi oldukça nitelikli ve bilimsel olarak ele alınması gereken bir olgudur. Öğrenme ifadesi, sözlük anlamı ile değerlendirildiğinde karşımıza aşağıda sıralanan bazı unsurlar ile karşılaşılmaktadır.

- Bilgi edinmek veya bilincine sahip olmak
- Bilgi veya gözlem neticesinde sahip olunan yeni durum ile farkındalık

Buradaki bilgi edinmek ve gözlem neticesinde oluşan farkındalık ifadesinin birkaç cümle ile açılmasına ihtiyaç vardır. Öğrenmenin sözlük anlamı olarak karşımıza çıkan bu ifadelerin, aslında sadece öğrenme sürecini değil öğrenme sürecinin sonrasını da kapsadığı fark edilmesi gereken bir noktadır. Yani bilgiyi elde etme durumu ki bu durumun ortaya çıkması sadece direkt hazır halde bulunan bilginin özümsenmesi değil, bilginin ortaya çıkarılması, bunun için ortaya koyulması gereken çaba veya çalışma, deneyimleme ya da birtakım dış unsurlar tarafından bahsedilen bu edim ve durumların öğreniciye sağlanmasını da içermesidir. Bunun ardından, öğrenme sürecinin bir anlamda durumsal çıktılarından bir olan kazanılmış bilgi ve bilginin hafıza da saklanması durumu yani kalıcılığı da öğrenme kavramının çerçevesi olarak çizilebilir (Dolmans ve diğerleri, 2005: 733; Gürlen, 2011: 221). Öğrenme sonucu elde edilen kazanımların uygulanması ya da uygulamaya koyulması gibi bir gerekliliğin olmadığı sözlük anlamından çıkarabilmek mümkündür. Elbette ki daha sonra ifade edilecek olan bilgi ve bilgisayar teknolojilerinin bu konuya entegre edilmesiyle öğrenme ifadesinin sözlük anlamına ait içerikte eksiklikler ya da tam olarak ihtiyaçların karşılanmasına yönelik etkinlik becerisinde tamamlanmaya gerek duyulacak noktalar ortaya çıksa da öğrenme kavramına ilişkin iskeletin niteliği büyük bir temeldir. Bu temel ışığında, uygulamaya koyulmadan da öğrenmenin kazanımları elde edilebilir (Witten ve diğerleri, 2005: 7). Öğrenme kavramının ele alınmasıyla birlikte odak konusu kavram bilgi kavramı olmaktadır.

Bilgi üretim için en önemli unsurlardan birisi de veri kavramıdır. İlk adımda, genel olarak karıştırılan veri, enformasyon ve bilgi kavramlarının farkının ortaya koyulması önemli ve faydalı olacaktır. Bu konuda karmaşanın nedeninin özellikle yabancı dildeki karşılıklarının bazen tek bir kelime ile ifade edilmesinden kaynaklanmaktadır (Yılmaz, 2009: 96). Bilgi kavramı, temelinde enformasyon ve veri kavramlarını barındırmaktadır. Veri, verici veya gözlem yoluyla elde edilmiş olan kodlanmış değişmez ifadeler olarak tanımlanabilir (Chen ve diğerleri, 2008: 15). Enformasyon ise verilerin anlamlandırılması ile ortaya çıkmış veri bütünleri ve bilgi ifadesi, enformasyonlardan hareketle yapılan yorumlama ve çıkarımlar şeklinde ifade edilebilir (Zins, 2007: 480). Böylece bilgi kavramı ele alınırken veriden bağımsız veya ilişkisiz bir şekilde değerlendirilmesinin olanaksız olduğu bir kez daha gözler önüne serilmektedir.

Şekil 1: Veri-Enformasyon-Bilgi İlişkisi



Kaynak: PennState World Campus MIS204, 25.11.2023

Bilgisayar teknolojilerinde yaşanan gelişmelerle birlikte çok miktarda verinin saklanması ve işlenmesi mümkün hale gelmiştir. Bunun sonucunda, bilgisayar ağları vasıtasıyla, fiziksel olarak farklı lokasyonlardan veriye ulaşım konusunda da

kolaylıklar ve imkânlar ortaya çıkmıştır. Bu süreçle birlikte, veri üretim hızındaki artışa paralel olarak, teknolojik imkânlara olan ihtiyaçlar da artmıştır. Bu noktada, özellikle karmaşık problemlere çözümünde bilgi ve bilgisayar teknolojilerini barından yaklaşım ve yöntemlere yönelim söz konusu olmuştur (Apaydın, 2020: 1).

Verinin elde edilmesi ve sonrasında bilgiye dönüşüm için gerekli olan analiz sürecine hazır hale getirilmesi oldukça önemlidir. İlgili süreç içerisinde de, veri değerlendirilerek bilgisayar teknolojisi ile birlikte çeşitli amaçlar için kullanılmak üzere bilginin üretilmesi sağlanmaktadır. Veriden bilgiye doğru olan bu dönüşüm çeşitli şekillerde ortaya çıkmaktadır. Bunlardan biride daha önce ifade edilmiş olan öğrenme kavramı ile ilgilidir. Sözlük anlamı ile öğrenme kavramı değerlendirildiğinde, bilgi kazanma, tecrübe edinme, çıkarım sonucu elde edilen kalıcı kazanç olarak ortaya çıktığı açıklanmıştı ve bu anlamda, öğrenme ve bilgisayar teknolojileri kavramına ait kesişim elamanlarından biri olan makine öğrenmesi kavramı ile karşılaşılmaktadır.

Makine öğrenmesi kavramı, çok geniş bir tanım ölçeğine sahiptir. Bu nedenle, makine öğrenmesi için literatürde birçok tanım bulunmaktadır. Bu kavramlardan bazıları aşağıdaki gibidir.

İstatistik, bilgisayar bilimleri ve mühendislik alanlarının kesişimi olan bu kavram, geçmiş ait veri ve deneyimlerden faydalanarak elde ettiği bilgiyi, geleceğe yönelik karar verme mekanizması olarak tanımlanabilir (Dangeti, 2017: 8).

Başka bir tanımda, makine öğrenmesi, geleneksel programlama yöntemlerini kullanarak programlamanın zor olduğu karmaşık problemlere, algoritma ve tekniklerle otomatik çözümler üreten bilgisayar bilimi dalı olarak ifade edilmektedir (Rebala ve diğerleri, 2019: 1).

Bilgisayar unsurlarının, gerekli bilgiyi ya da bilgileri öğrenip sonradan karşılaşılması mümkün olaylar üzerinde birtakım örüntü ve ilişkiler ışığında, çeşitli yaklaşım ve modelleri kullanarak çözüm sunması veya karar vericilere destek vermeleri olarak kavramsal olarak açıklanabilir (Bilgin, 2018: 13; Öztemel, 2003: 21).

Programlama, makine öğrenmesi kavramının içinde değerlendirilmesi gereken bir kavramdır. Bu sebeple, programlama kavramının da değerlendirilmesiyle birlikte makine öğrenmesi tanımı da yapılabilir. Bu anlamda, bizzat kodlanmadan,

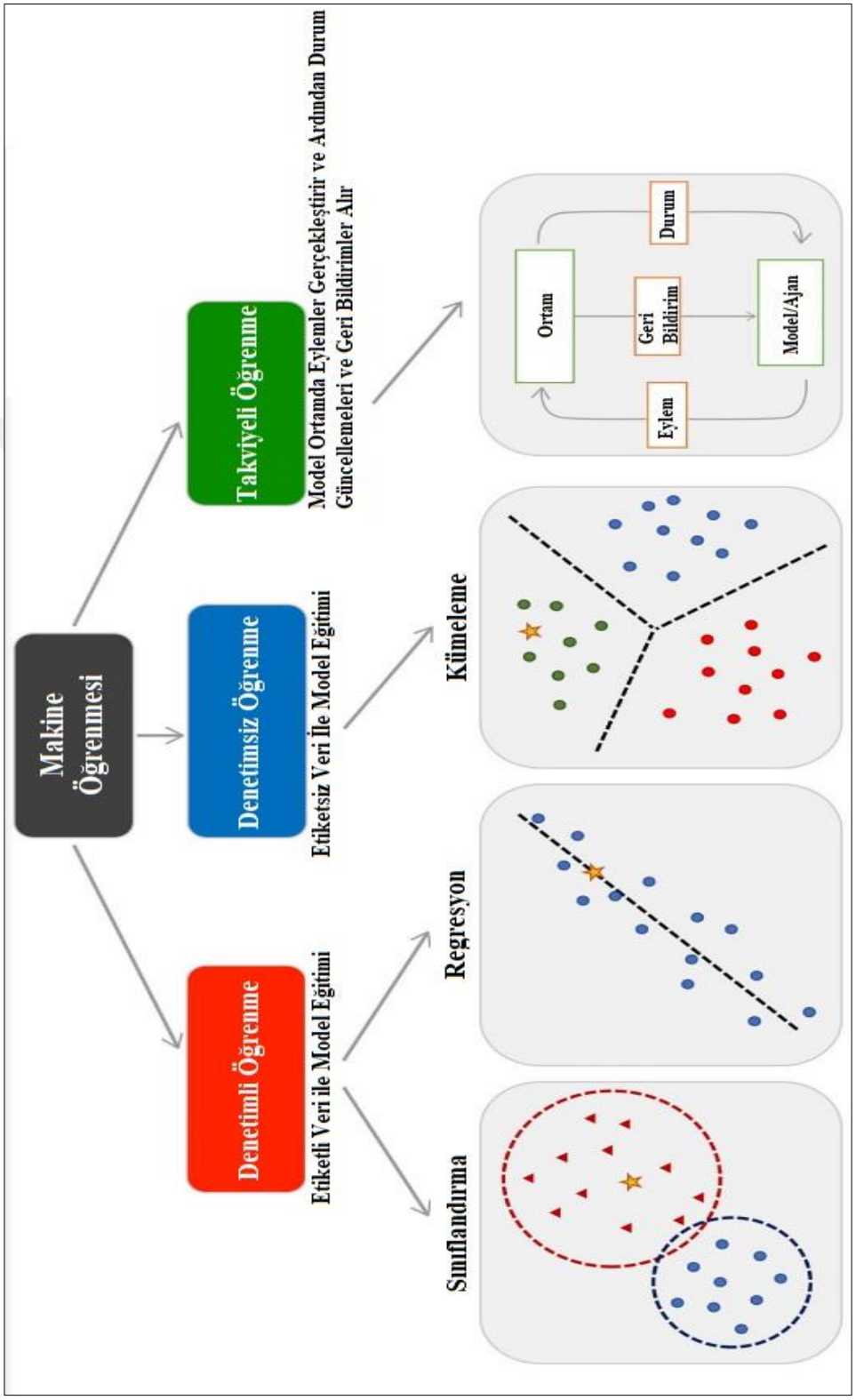
bilgisayarlara öğrenme kabiliyetinin sağlandığı bir saha olarak değerlendirilebilir (Samuel, 1959: 211)

Belirtilen tanımların dışında, birçok farklı bakış açısı ile aynı kavram ele alınabilir. İnsan zekâsı ve müdahalelerinin devre dışı bırakılarak ilgili çevreden elde edilen verileri, algoritmalar yoluyla bilgi haline getiren bir yapay zekâ alt dalı olarak da tanımlanabilmektedir (Bi ve diğerleri, 2019: 2222). Veri bütünü içindeki örüntü ve yapıları analiz edip elde edilen sonuçlar ışığında açıklayıcı, bilgi verici ve karar verme unsuru da ortaya koyan birtakım model ve araçlar üreten sistem makine öğrenmesi olarak açıklanabilir (Brynjolfsson ve Mithchell, 2017: 1530).

2.2. MAKİNE ÖĞRENMESİNDE ÖĞRENME TÜRLERİ

Makine öğrenmesi kavramının bilgi ve çıkarım elde edilmesinde kullanılan en önemli araçlardan birisi olması özelliği, teknolojik gelişmelerle birlikte daha da ön plana çıkmıştır. Bu sayede, makine öğrenmesi oldukça geniş bir kullanım alanına sahip olmuştur. Bu başlık altında ifade edilecek öğrenme türlerinde girdiler ve çıktılara arasındaki ilişkisel durum veya denetleyici birtakım unsurların öğrenme süreci içerisindeki varlığı türlerin farklılığını ortaya koymaktadır. Girdiler ışığında, istatistik tabanlı bir model üzerinden tahmin ve çözüm üretme de denetimli bir mekanizmadan diğer taraftan, girdiler söz konusu iken denetleyici bir yapıda çıktıların olmadığı durumlarda da denetimsiz bir mekanizmadan bahsedebiliriz (James ve diğerleri, 2013: 1). Makine öğrenmesinin anahtar niteliği olan öğrenme yeteneği ele alındığında, öğrenme kavramın 4 başlık altında toplamak mümkündür. Bu başlıklar denetimli öğrenme, denetimsiz öğrenme, yarı denetimli öğrenme ve takviyeli öğrenme şeklinde incelenebilir.

Şekil 2: Öğrenme Türleri

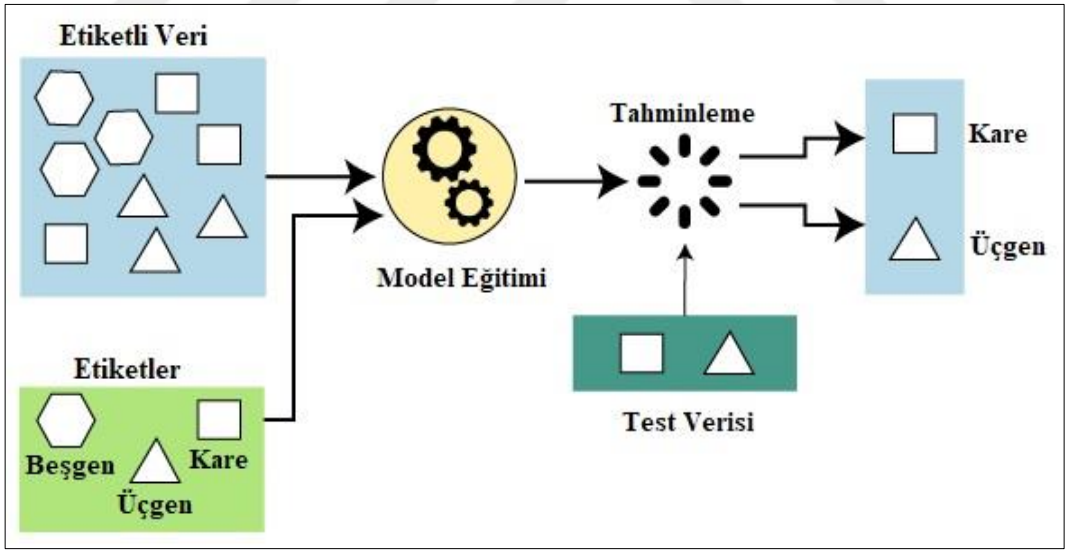


Kaynak: Peng ve diğerleri, 2021: 2.

2.2.1. Denetimli Öğrenme

Denetimli öğrenmede, gerçek bilgilerden oluşan etiketli verilerin kullanılmasıyla öğrenme gerçekleştirilmektedir. Kullanılacak olan girdiler yani veriler etiketli olarak sunulmaktadır. Öğrenme sonucunda, gerekli model ve araç ortaya koyulur (Zebari, 2023: 7; Friedman ve diğerleri, 2001: 9). Böylece, bir dizi örnek veya eğitim setine doğru çıktılar sağlanır ve bu örneklerin ışığında algoritma, üretmiş olduğu çıktısını girdi olarak verilenlerle kıyaslayarak daha doğru sonuç vermeyi öğrenir (Alzubi ve diğerleri, 2018: 6). Görüldüğü gibi öğrenmenin gerçekleşmesi için hem verileri hem de etiketlenmesi sonucu elde edilmiş olan çıktılar sürece dâhil edilmektedir. Böylece, sistem içerisindeki veriler yine verilmiş olan etiketler sayesinde bir anlamda sonuçlar ile eşleştirilerek girdi-çıktı ilişkisi ortaya çıkarılmış olur (Nilsson, 1996: 6).

Şekil 3: Denetimli Öğrenme Süreci



Kaynak: Javatpoint, 25.11.2023

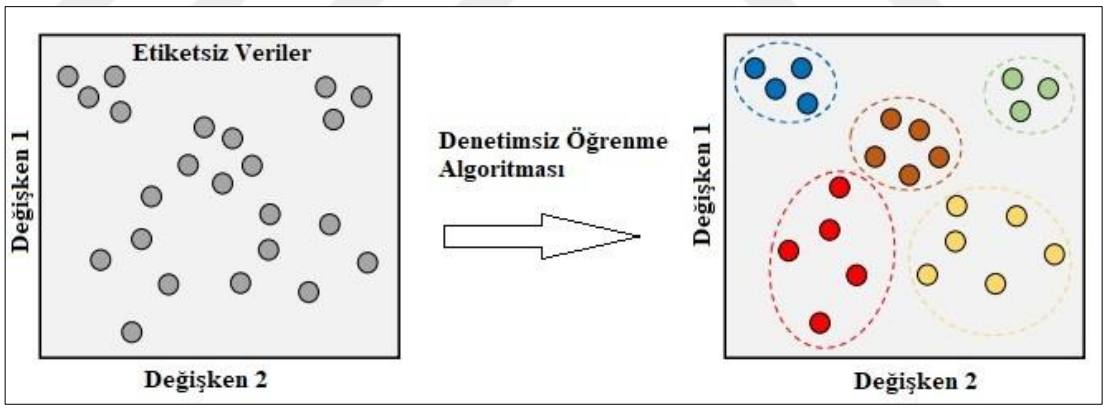
Bu süreç sonrasında, model tarafından öğrenilen ilişkiler ve örüntüler sayesinde test verisi adı verilen sınanmamış girdiler ile karşılaşıldığında yeni girdilerin doğru şekilde etiketlenmesi sağlanmaya çalışılmaktadır.

Şekil 2'de görüldüğü gibi denetimli öğrenme türü kendi içinde iki alt gruba ayrılmaktadır. Bunlar sınıflandırma ve regresyon operasyonları olarak isimlendirilmektedir. Sınıflandırma ve Regresyon alt gruplarının farkı ele alınan değişkenlerin türü ile ilgili bir durumdur. Tahminlenecek olan çıktılar, nitel değişkenler ise sınıflandırma, bu çıktılar eğer nicel değişkenler ise regresyon kullanılmaktadır (Chao, 2011: 9).

2.2.2. Denetimsiz Öğrenme

Denetimli öğrenmeden farklı şekilde öğrenmek için kullanılan girdiler yani veriler etiketsiz şekilde sunulmaktadır. Bu durumda, etiketsiz olarak sunulmuş verilerde, bilgisayar tarafından, verideki birtakım örüntü veya yapılar araştırılıp ortaya çıkarılır. Bu sebeple, denetimsiz öğrenmede kompleks veya öncesinde tahminlenmesi güç sonuçlar ortaya çıkabilmektedir (Erken ve Şenyay, 2023: 55).

Şekil 4: Denetimsiz Öğrenme Süreci



Kaynak: Morimoto ve Ponton, 2021: 2

Denetimsiz öğrenmede denetimli öğrenmeye nazaran zorluklar yaşanabilmektedir. Bunun sebebi, belirtildiği gibi öğrenmenin gerçekleştirileceği veri için daha öncesinde herhangi bir değerlendirme ve açık bir şekilde belirlemeler olmamasıdır. Bundan dolayı, denetimsiz öğrenmede girdi verisinin karakteristiklerini, veriden örüntü ve yapısın yakalayarak çıktıda elde edilmesi anahtar pozisyondadır

(Dayan ve diğerleri, 1999: 859; James ve diğerleri, 2023: 504). Denetimsiz öğrenme, kümeleme, temel bileşenler analizi, görselleştirme ve boyut indirgeme uygulamaları için kullanılmaktadır.

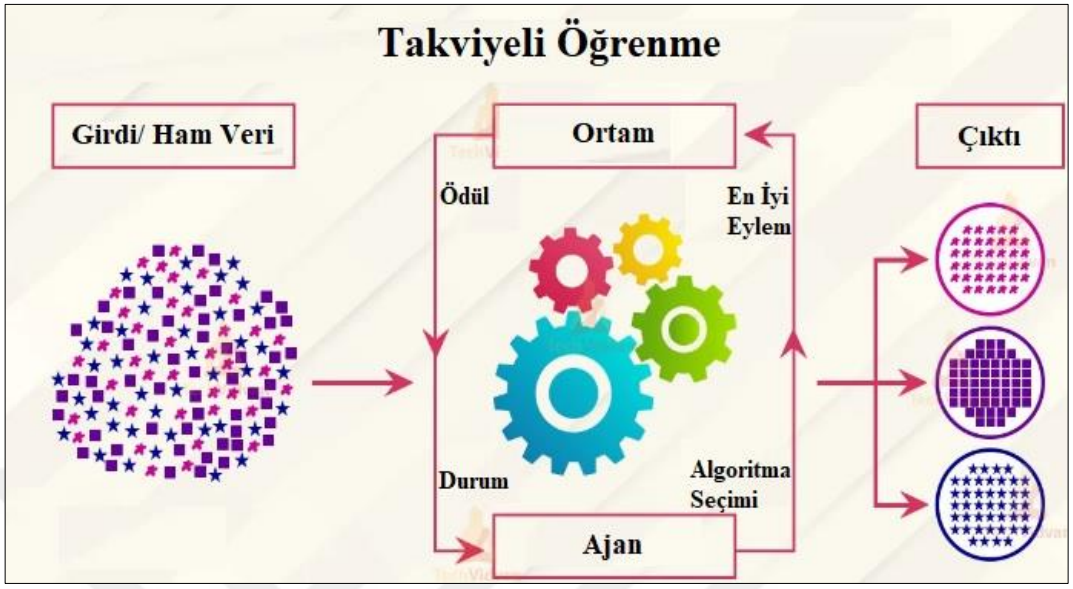
2.2.3. Yarı Denetimli Öğrenme

Yarı denetimli öğrenme, belirli öğrenme görevlerini gerçekleştirmek için etiketli verilerin yanı sıra etiketsiz verilerin de kullanılmasını da içine alan öğrenme türüdür (Van Engelen ve Hoos, 2020: 373). Geleneksel sınıflandırıcılar eğitilmek için etiketli verilere kullanmaktadır fakat etiketli verilere sahip olmak pahalı, zaman açısından sorunlu olabilen ve zor bir durumdur. Etiketsiz verilerin ise daha kolay elde edilmesi durumu mümkün olsa da bu verileri kullanıma sokmanın az sayıda yolu söz konusudur. Yarı denetimli öğrenme, daha iyi sınıflandırıcılar oluşturmak için etiketli verilerle birlikte büyük miktarda etiketsiz veri kullanarak bu sorunu ele alır (Gibson ve diğerleri, 2013: 133). Böylece, daha az insan müdahalesi ve iyi bir doğruluk seviyesi ortaya koyduğu için oldukça ilgi görmektedir (Li ve Zhou, 2015: 175).

2.2.4. Takviyeli Öğrenme

Takviyeli öğrenme ya da diğer bir ifade ile pekiştirmeli öğrenmedeki yaklaşım şu şekilde ifade edilebilir: Doğal ve yapay sistemler, kısaca öğrenici, gerçekleştirilen aksiyonlarla, belli bir durum veya pozisyondan farklı bir durum veya pozisyona götürülmektedir. Bu aksiyonların sonucunda ceza ya da ödüllerin elde edildiği ortamlarda, öğrenicinin, davranışlarının sonuçlarını tahmin etmeyi ve optimize etmeyi amaçladığı bir öğrenme türüdür (Dayan ve Niy, 2008: 185).

Şekil 5: Takviyeli Öğrenme Süreci



Kaynak: TechVidvan, 25.11.2023.

Takviyeli öğrenme, ne yapılacağı, hangi eylemin nasıl gerçekleştirileceği ve bunun sonucunda sayısal bir ödülün optimize edileceğinin araştırmasıdır. Öğrenicinin gerçekleştireceği eylemler kendisine söylenmeyip öğrenici bahsedilen araştırmada, hangi aksiyon ile ödülün en üst düzeye çıkacağını deneyerek keşfetmesi gerekmektedir (Sutton ve Barto, 2018: 1).

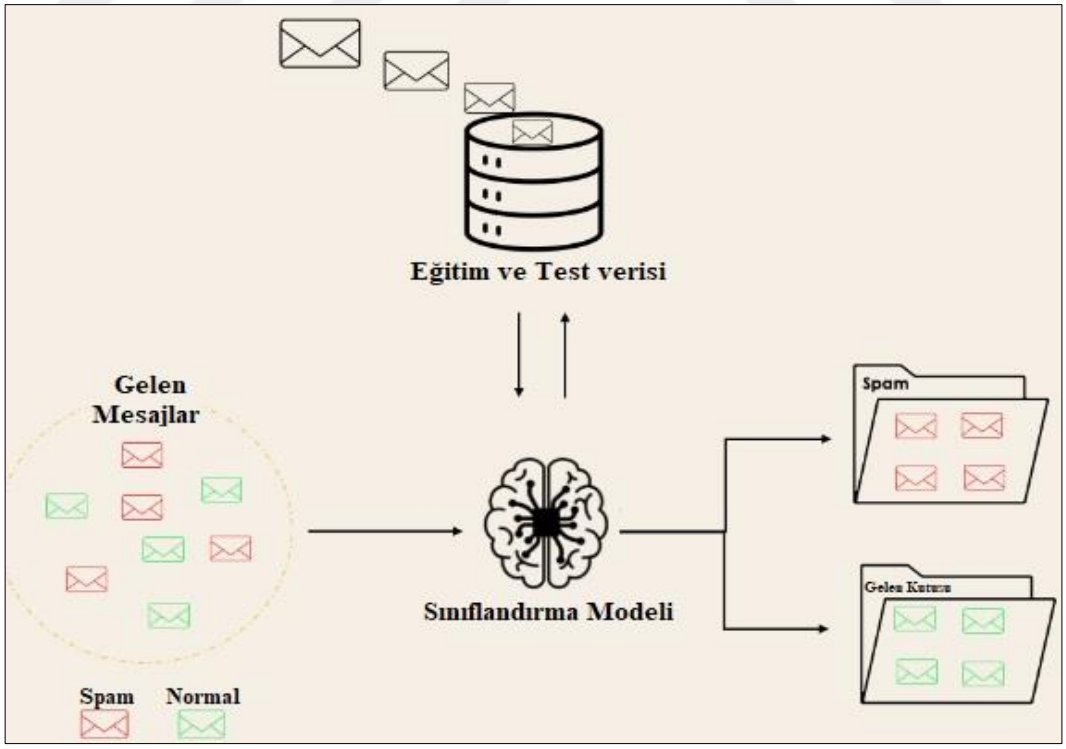
2.3. SINIFLANDIRMA

Sınıflandırma, makine öğrenmesi yöntemleri arasında yaygın bir kullanıma sahip yöntemdir. Tıbbi çalışmalar, görüntü işleme ve görüntü analizi, pazarlama, finansal analizler, spam ve zararlı yazılım tespiti, sigortacılık ve bankacılık gibi araştırmalarda oldukça geniş bir kullanıma sahiptir.

Sınıflandırma işlemi, en basit anlamda sınıf nitelik/değişken değerleri bilinen bir veri kaydından hareketle oluşturulan bir fonksiyon veya model yapısıyla sınıf nitelik/değişken değeri bilinmeyen ve elde edilen modelle sınanmamış veri kayıtlarının sınıf niteliği/değişkeni değerini belirleme operasyonudur şeklinde ifade edilebilir (Özkan, 2008: 51).

Sınıflandırma operasyonu iki ana temel üzerine kurulu şekilde gerçekleştirilmektedir (Aggarwal, 2015: 286-287). Ana temel olarak ifade edilen 2 adım eğitim ve test aşamasıdır. Eğitim aşamasında, sınıf niteliği belli olan diğer bir ifade ile sınıf niteliği etiketli olan veriler kullanılmaktadır. Bu verilerden alınan örneklerden eğitim verisi adı verilen bir veri seti oluşturulur. Oluşturulan veri setindeki nitelikler/değişkenler ve sınıf değerleri arasındaki ilişki ve örüntünün analiz edilmesiyle sınıflandırma modeli şeklinde ifade edilen tahmine dayalı yapı ortaya koyulur. Modelin ortaya çıkması ile birlikte daha sonra karşılaşılacak veya ilk defa gözlenmiş olan veri kayıtlarının sınıf nitelik/değişken değerlerinin belirlenmesi işlemi yani sınıflandırılması gerçekleştirilmektedir. Böylece, yeni veri kayıtlarının da daha yüksek bir doğruluk oranı ile sınıflandırılması olanağı da ortaya çıkacaktır (Silahtaroğlu, 2016: 67). Bunun yanı sıra, veri setindeki veri kayıtlarının ve aralarındaki örüntülerin anlaşılmasını sağlayıp be işlemi kolaylaştıracaktır (Sumathi ve Sivanandam, 2006: 205).

Şekil 6: Sınıflandırma İşleminin Spam Mesaj Yakalama Üzerinden İşleyişi



Kaynak: Datacamp, 25.11.2023

İkinci aşama olan test aşamasında ise eğitim verileri kullanılarak elde edilen sınıflandırma modelin doğruluğu sınanır. Böylece, sınıflandırma modeli kullanılarak daha önce ifade edildiği gibi daha önce gözlenmemiş olan veri kayıtlarının sınıf nitelik/değişken değerlerinin tayin edilmesi işlemi gerçekleştirilmektedir. Bu sistemde hareket etmekle birlikte, farklı yapılarda ortaya çıkmaktadır. Olasılıklara dayalı modeller, karar ağaçlar ve sinir ağları gibi yapılar en çok kullanılan sınıflandırma yapılarındandır.

Sınıflandırma kavramı, çoğu zaman regresyon kavramı ile birlikte kullanılan bir kavramdır. Sınıflandırmada, hedef nitelik/değişken değeri kategorik olan değişkenler olduğunda kullanılırken regresyon uygulamalarına, hedef niteliği numerik olan değişkenler söz konusu olduğunda başvurulmaktadır (Vercellis, 2009: 221).

2.4. EKSİK VERİ TAMAMLAMADA KULLANILAN MAKİNE ÖĞRENMESİ ALGORİTMALARI

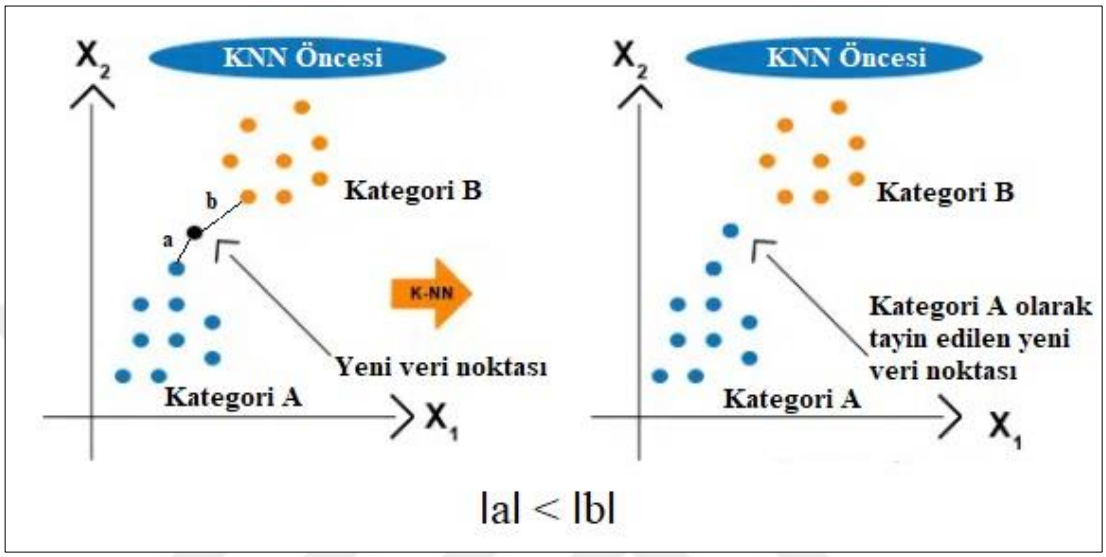
Bu kısımda, çalışma kapsamında kullanılmış olan ve birçok amaçla makine öğrenmesi çalışmalarında kullanıldığı gibi eksik veri tamamlama operasyonlarında da kullanılan makine algoritmaları ele alınacaktır. Tez çalışması çerçevesinde ele alınacak makine öğrenmesi algoritmaları; en yakın k- komşu algoritması (KNN), Random Forest (RF), Stokastik Regresyon (SR), Amelia ve Naive Bayes algoritmalarıdır.

2.4.1. En Yakın K-Komşu Algoritması

En yakın k-komşu algoritması, mesafe tabanlı makine öğrenmesi yöntemlerindendir. Anlama ve kullanım kolaylığı sunan bu yöntem, özellikle regresyon ve sınıflandırma problemlerini çözümlemek için sık başvurulan yöntemler arasındadır (Mahesh, 2020: 385). Algoritmanın çalışma içeriğinde iki adet önemli unsur söz konusudur. Bu unsurlardan birisi uzaklık ve diğeri de komşu sayısını ifade eden k değeridir. Veri setine yeni dâhil edilen bir veri kaydının veya veri vektörünün hedef niteliğine ait değerin atanması işlemi, bahsedilen uzaklık ölçüsü limitlerindeki k adet komşu veri kaydından hareketle benzerlik ataması şeklinde yapılmaktadır. Bu

işleyiş, temel mantığın en yakın mesafe içerisinde bulunan k- adet komşu veriden öğrenme üzerinedir şeklinde özetlenebilir (Aryanto ve diğerleri, 2023: 36).

Şekil 7: En Yakın k-Komşu Algoritması İşleyişi



Kaynak: Datamites, 25.11.2023

Yöntemin işleyişinde ele alınacak olan hangi k adet kaydın ve/veya verinin değerlendirileceği sorusunun cevabı uzaklık ölçüsü üzerinden verilebilmektedir. En yakın k- komşu algoritmasında kullanılan çeşitli uzaklık ölçüleri mevcuttur. Uzaklık ölçütlerinin fonksiyon olarak ifade edilmesinde kullanılan p nitelik veya değişken sayısını, n ise veri veya kayıt sayısını göstermektedir. Bahsedilen uzaklık ölçülerinden birisi Öklid uzaklığıdır. Öklid uzaklığı, Pisagor teoreminin iki boyutlu uzayda bir uygulamasıdır ve ilgili uzayda iki nokta arasındaki uzaklığı ifade etmektedir (Hu ve diğerleri, 2016: 3). Veri noktalarının koordinatları arasındaki farkın karekökü ilgili noktalar arası Öklid uzaklığıdır. Bu uzaklık (1) numaralı eşitlik ile tanımlanmaktadır (Mulak ve Talhar, 2015: 2102).

$$d(i,j) = \sqrt{\sum_{h=1}^{p} (x_{ih} - x_{jh})^2}, \quad i,j=1,2,\ldots,n; \; h=1,2,\ldots,p \qquad (1)$$

Diğer bir uzaklık ölçütü Manhattan uzaklığıdır. Veri noktalarının daha yüksek boyutlara sahip olduğu durumlarda iki nokta arasındaki mesafeyi tanımlamak için kullanılır. Bu durumlarda Öklid uzaklığı yerine Manhattan uzaklığı tercih edilir. Manhattan uzaklığında, bahsedilen uzaklık noktaların kartezyen koordinatları arasındaki farkın mutlak değerlerinin toplamı şeklinde tanımlanmaktadır (Ratnasari, 2023: 99). Bu ifade, (2) numaralı eşitlikte gösterilmiştir.

$$d(i,j) = \sum_{h=1}^{p} \left| x_{ih} - x_{jh} \right| , \quad i,j=1,2,\ldots,n; \; h=1,2,\ldots,p \tag{2}$$

Minkowski uzaklığı, kullanılan uzaklık ölçütlerinden diğeridir. Bu uzaklık, (3) numaralı eşitlik ile ifade edilmektedir.

$$d(i,j) = \left(\sum_{h=1}^{p} \left| x_{ih} - x_{jh} \right|^{m} \right)^{\frac{1}{m}} \tag{3}$$

Minkowski uzaklık ölçütünde, negatif olmayan bir tam sayı olan m değeri ile daha önce açıklanan uzaklık ölçüleri elde edilebilmektedir. m değeri 1 olduğunda Manhattan, 2 olduğunda ise Öklid uzaklık ölçütleri elde edilmektedir (Zhang, 2012: 2543).

2.4.2. Rassal Orman (Random Forest) Algoritması

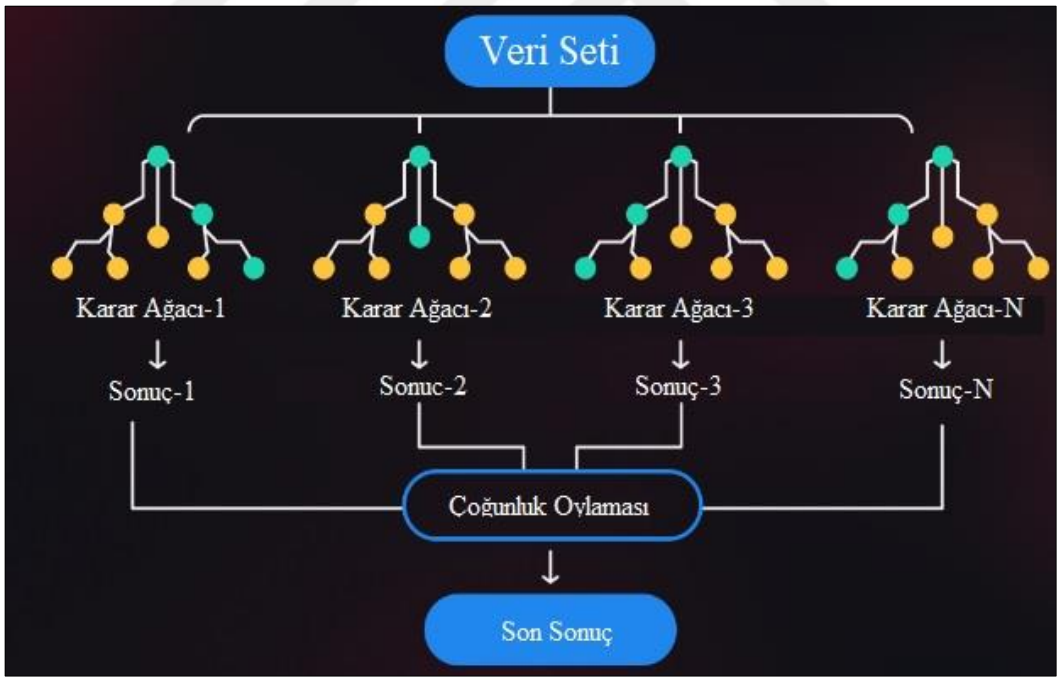
Rassal orman algoritması, sınıflandırma işlemlerinde, karar verme problemlerinde, regresyon uygulamalarında kullanıldığı gibi eksik veri tamamlama operasyonlarında da kullanılan karar ağaçlarından oluşmuş bir karar ormanı yapısı teşkil etmektedir (Akar ve Güngör, 2012: 107). Aynı zamanda, Bagging şeklinde ifade edilen yöntemler arasında kullanımına en çok başvurulan algoritmalardan biridir.

Bilinen klasik karar ağaçlarında, dallanma diye ifade edilen düğümler merkezinde bölümlenme işleminde, düğümler mümkün olan en iyi değişkeni kullanarak dallanmaktadırlar. Rassal Orman algoritmasında ise kullanılan değişkenlerin rassal olarak seçilmesiyle dallanma yapılması söz konusudur. Gini

indeksi, yaygın bir şekilde dallanma ölçütü olarak kullanılmaktadır çünkü burada sınıflandırma ve regresyon ağaçları yönteminden yararlanılmaktadır. Böylelikle, rassal olarak oluşturulan karar ağacı modellerinin eğitimi gerçekleştirilir ve çoğunluk oylaması ile sonuçlar birleştirilmektedir. Bu durum eğer bir regresyon problemine yönelik ise birleştirmede çoğunluk oylaması yerine ortalama kullanılır (Liaw ve Wiener, 2002: 18)..

Belirtilen karar ağaçlarının birbirinden bağımsız olması ile birlikte, ilişkisel anlamda göz ardı edilebilecek seviyede korelasyona sahip karar ağaçları olması da söz konusudur. Söz konusu karar ağaçları, varyansı düşürmek için bootstrap ile veri setinden çekilmiş örneklerden meydana gelmektedir. Böylece, ilgili verinin hedef niteliği ve/veya sınıfı, karar ağaçlarınca oylanarak tayin edilir (Kulkarni ve Sinha, 2013: 1145).

Şekil 8: Random Forest Algoritması İşleyişi



Kaynak: Serokell, 25.11.2023

Rassal orman algoritmasında, her karar ağacı veri setinin farklı alt veri kümesi üzerinde eğitilir. Burada, her karar ağacının farklı bir özelliğe odaklanmasının

sağlanması ve ilgili özellik için ilgili alt veri kümesinden yararlanarak optimum karar ağacı oluşturma amaçlanmaktadır (Rigatti, 2017: 33). Karar ağaçlarının eğitiminin ardından, yeni bir kayıt için karar ağaçları hedef nitelik/değişken değeri belirlemektedir. Belirlenen bu değerler, oylama şeklinde ifade edilen yöntemle birleştirilir. Böylece, doğrulukta yüksek seviyelerin yakalanması, farklı ağaçlarda yer alan verilerden kaynaklı sorun ve kusurların azaltılması ya da mümkün olduğu kadar yok edilmesi ve yine sapmaların azaltılması sağlanmaya çalışılmaktadır (Rigatti, 2017: 33).

Eksik verilerin tamamlanması da, rassal ormanların kullanıldığı uygulamalardan biridir. Algoritma olarak nasıl bir işleyişe sahip olduğu açıklanan bu yöntemin eksik verilerin tamamlanmasında bahsedilen işleyişi nasıl uyguladığı aşağıdaki gibi özetlenebilir (Tang ve Ishwaran, 2017: 364).

- Eksik olan verilerin tamamlanması; ormanı genişletmek, verilerin yakınsamasını kullanarak eksik verilerin tamamlanmasına ilişkin sonuçları güncellemek ve geliştirilmiş sonuçlar için iteratif şekilde devam etmek
- Ormanı genişletirken eş zamanlı olarak eksik verileri tamamlamak ve geliştirilmiş sonuçlar için iteratif şekilde devam etmek
- Eksik değerlere sahip her bir değişkeni kullanılarak bir orman geliştirmek ve bu orman ile eksik verileri tamamlamak ve geliştirilmiş sonuçlar için iteratif şekilde devam etmek

2.4.3. Stokastik Regresyon

Standart bir regresyon ile eksik verilerin tamamlanması gibi farklı bir regresyon versiyonu şeklinde ele alınabilecek stokastik regresyon ile ilgili hesaplamalar sonucunda eksik veriler tamamlanmaktadır. Genel olarak, sürekli değişken tahminleri için, gürültüye sahip ve Gauss Markov varsayımlarının sağlandığı bir durumdaki veriler söz konusu olduğunda kullanılan, çeşitli bağımsız değişkenler ve normal dağılan rassal hata teriminden oluşan stokastik regresyon modeller kullanılmaktadır (Rässler ve diğerleri, 2013: 22).

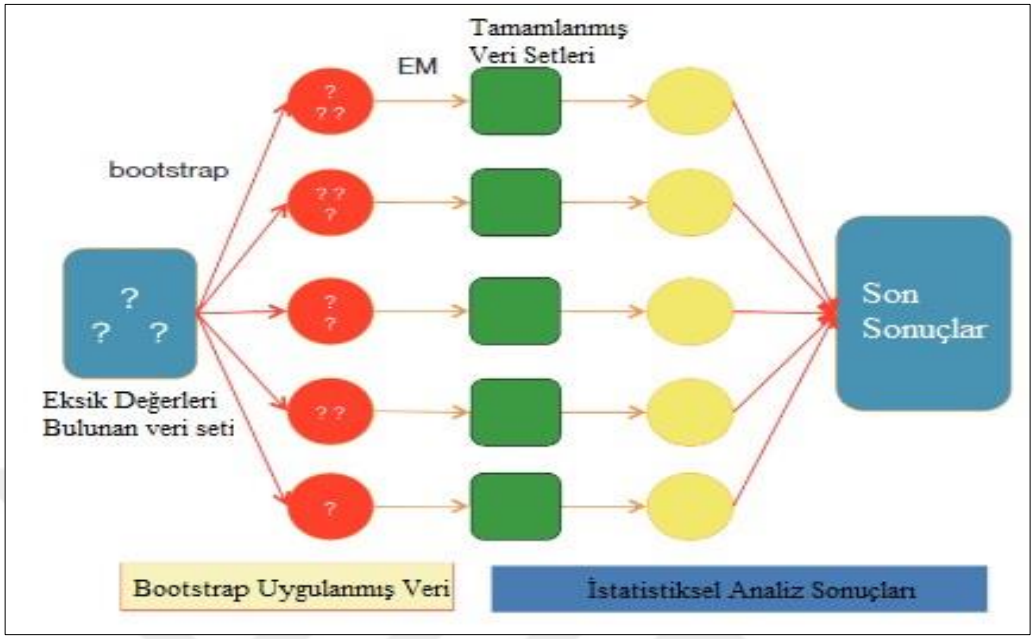
Regresyon eşitlikleri, eksik verileri tamamlamak adına bazı değerler tahminlemektedir. Tahmin edilen değerlerin hesaplanmasının ardından, tahmin edilmiş olan her değere rassal bir hata terimi eklenir. Sonuçta elde edilen bu toplam, eksik değerler yerine kullanılarak eksiklik tamamlanır. Buradaki hata terimleri, birbirinden bağımsız olan ortalaması sıfır ve varyansı önceki regresyon analizinden kalan varyansa eşit olan bir normal dağılıştan elde edilmiş olan rastgele sayılardır (Baraldi ve Enders, 2010: 13).

Eksik değerlerin, doğrusal regresyonla tamamlanması söz konusu olsa da bazı problemleri de beraberinde getirebilmektedir. Doğrusal regresyonla tamamlamada, eksik değerler yerine koyulacak değerler bir doğrudan elde edilmektedir. Bu durumda, verinin istatistiksel dağılımının etkilenmesi ve kovaryansların gerçekte olduğundan daha düşük olarak elde edilmesi gibi bir problem ortaya çıkabilmektedir. Stokastik regresyon, bu sorunların üstesinden geldiği, özellikle tamamen rastgele ve rastgele eksik verilerin varlığında yansız sonuçlar ürettiği literatürdeki çalışmalarda gözlenmektedir.

2.4.4. Amelia Algoritması

İsim ve fonksiyon olarak bir program veya program içinde bir paket olan bu yöntem, istatistiksel modelleme yöntemlerinden biridir. Amelia algoritması, çoklu tamamlama mekanizması ile hareket eden bir yöntemdir. Çoklu tamamlama mekanizması, bootstrap ve beklenti maksimizasyonu yöntemlerini içermektedir (Honaker ve diğerleri, 2011: 2).

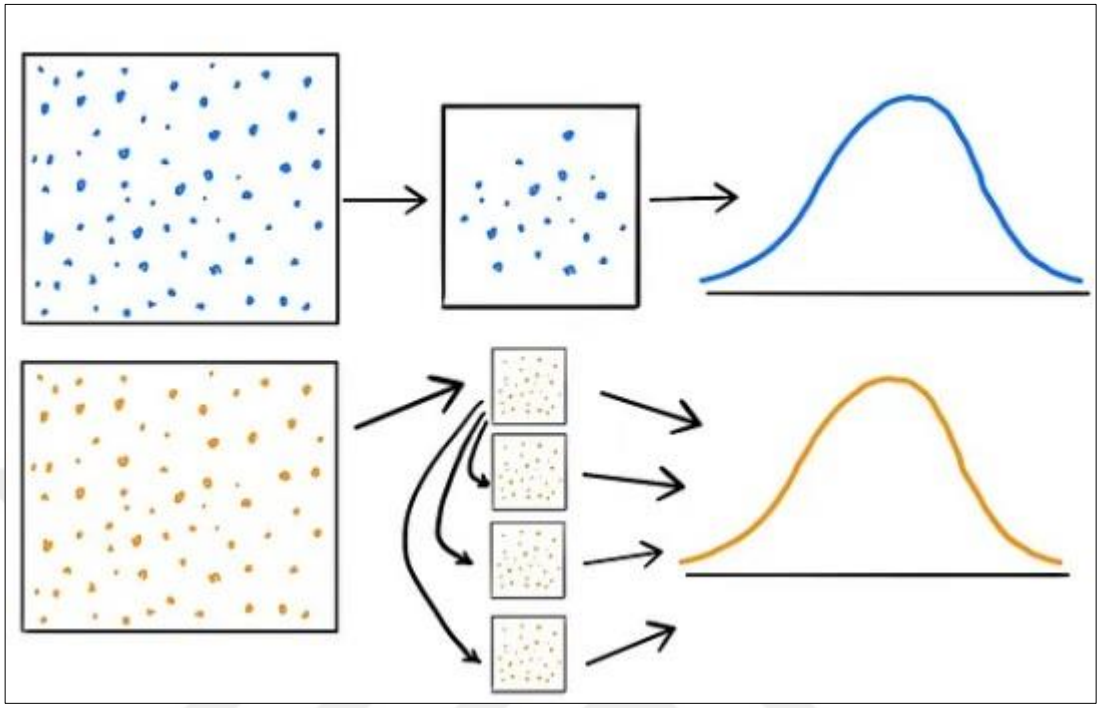
Şekil 9: Amelia Algoritmasının İşleyişi



Kaynak: Zhang, 2016: 2.

Bootstrap, kısaca istatistiksel çıkarım için bir hesaplama aracı şeklinde tanımlanabilir (Zoubir ve Iskandler, 2007: 11). Bootstrap yöntemleri, belli bir tahminci veya istatistiksel öğrenme yöntemiyle ilişkili belirsizliğin ölçülmesi ve değerlendirilmesi için kullanılabilen yeniden örnekleme teknikleridir (Doğan, 2017: 2).

Şekil 10: Bootstrap İşleyişi



Kaynak: Medium, 25.11.2023

Beklenti maksimizasyonu ise eksik verilerin söz konusu olduğu durumlarda, parametrelerin en yüksek olabilirlik tahminlerini elde etmek adına bir kullanıma sahiptir (Dempster ve diğerleri, 1977: 1). Bu yöntemde, farklı örneklemler ile eksik verilere sahip alt veri setleri elde edilmektedir. Birden fazla model kullanılıp her bir modelde, bahsedilen bootstrap ve beklenti maksimizasyonuyla tamamlanan eksik verilerin ortalaması alınıp bu sonuçlar birleştirilmektedir (Kabir ve diğerleri, 2020: 368).

2.4.5. Naive Bayes Algoritması

Giriş bölümünde de ifade edildiği gibi Naive Bayes algoritması, çalışma kapsamında, eksik verilerin daha önce bahsedilmiş olan makine öğrenmesi yöntemleri ve temel yöntemler ile tamamlanmasından sonra elde edilen tamamlanmış veri setlerinde sınıflandırma operasyonu için uygulanacaktır.

Bu algoritma, metin madenciliği, sınıflandırma ve daha birçok amaçla sıklıkla kullanılan ve Bayes Teoremi'ne dayalı koşullu olasılık tabanlı bir yapıya sahiptir (Saritas ve Yasar, 2019: 90). Naive Bayes algoritmasında, veri setindeki değişkenlerin birbirinden koşullu bağımsız ya da oldukça düşük ve göz ardı edilebilecek düzeyde düşük bir korelasyona sahip olduğu varsayımı söz konusudur (Patil ve Sherekar, 2013: 257). Bu varsayımın, yöntemin zayıf tarafı olarak nitelendirildiği bir durum olarak yorumlanması da söz konusu olsa da, ilgili varsayım, yöntemin işleyişini ve analizi oldukça kolaylaştırmaktadır (Vembandasamy ve diğerleri, 2015: 442). Algoritmanın isminde yer alan "naive" ifadesi de bu sebeple verilmiştir.

Algoritmanın işleyişine gelindiğinde, daha önce hedef nitelik/değişken değerleri bilinen veri kayıtlarından yararlanarak yeni bir veri kaydının, tanımlanmış olan her bir hedef nitelik/değişken değerini alması olasılığı (koşullu) hesaplanmaktadır. İlgili olasılıklardan en büyük koşullu olasılık değerine sahip olan hedef nitelik/değişken değeri, o veri kaydının hedef nitelik/değişken değeri olarak atanmaktır (Yager, 2006: 579).

Bahsedilen hedef nitelik/değişken, kategorik bir değişken ise ilgili olasılıklar hesaplanırken hedef nitelik/değişken değerlerinin frekanslarından yararlanılmaktadır. Hedef niteliğin nümerik değerler alan bir değişken olması durumunda ise ilgili olasılıklar, ortalaması μ, varyansı σ^2 olan bir normal dağılış olasılık yoğunluk fonksiyonu yardımıyla hesaplanmaktadır (Han ve Kamber, 2006: 312). Buradaki ortalama ve varyans veri setindeki ilgili sayısal değişkenin aldığı değerlere ait ortalama ve varyans değerini ifade etmektedir.

X veri kümesi, (x_1, x_2, … , x_n) nitelik/değişken değerleri ve C hedef nitelik/değişken kümesi de (C_1, C_2, … , C_m) hedef nitelik/değişken değerlerini alan değişkenler olarak tanımlansın. Böylece, ilgili nitelik/değişken değerine göre koşullu olasılık (4) numaralı eşitlikteki gibi verilebilir.

$$P(Ci|X)= \frac{P(X|Ci).P(Ci)}{P(X)} \qquad (4)$$

Burada, $P(C_i|X)$ sonrasal olasılık, $P(X|C_i)$ koşullu olasılığı(Likelihood), $P(C_i)$ hedef niteliğe ait değerin önsel olasılığı ve P(X) ise tahminciye ait önsel olasılığı ifade etmektedir. Önce bahsedildiği gibi algoritmanın sahip olduğu bağımsızlık

varsayımdan dolayı, $P(X|C_i)$ olasılığı aşağıdaki gibi ifade edilebilmektedir. Paydadaki olasılık değeri ise, her bir hedef nitelik/değişken değeri için aynı olacaktır.

$$P(X/C_i) = \prod_{k=1}^{n} P(x_k/C_i) \quad (5)$$

Böylece, payda yazan ifade açısından en yüksek olasılık değerine sahip olan hedef nitelik/değişken değeri, yeni veri kaydının hedef nitelik/değişken değeri olarak atanacaktır. Bu yöntem, en büyük sonrasal sınıflandırma(MAP: Maximum A Posteriori Classification) olarak adlandırılmaktadır ve (6) numaralı eşitlikteki gibi gösterilmektedir.

$$\arg\max_{C_i} = \{P(X/C_i)P(C_i)\} \quad (6)$$

Çalışma kapsamında kullanılacak olan yöntemler, tez çalışmasının bu bölümüne kadar olan kısımda hem teorik hem de uygulamaya yönelik açıklamalarca ifade edilmiştir. Tablo 5'te uygulamada kullanılan tüm yöntemlere ilişkin özet bilgiler sunulmuştur.

Tablo 5: Uygulamada Kullanılmış Olan Tüm Yöntemler

	Yöntemler	**Açıklama**
Temel Yöntemler	Liste Boyunca Silme Yöntemi	Tamamlama yerine silme stratejisi ile hareket edilir. Veri setinde bulunan eksik değerlerin bulunduğu veri kayıtlarının tamamı veri setinden çıkarır.
	Son Gözlemi İleri Taşıma Yöntemi	Eksik değer, kendisinden önceki gözlenmiş olan değerle tamamlanır. Uzun süreleri kapsayan çalışmalar ve özellikle klinik çalışmalarında tercih edilir.
	Ortalama İle Tamamlama	Eksik değer, gözlenmiş değerleri aritmetik ortalaması ile tamamlanır. Eğer kategorik veriler söz konusu ise mod-medyan değeri kullanılabilmektedir. Veri kaybı yaşanmaz.
Makine Öğrenmesi Algoritmaları	En Yakın K-Komşu Algoritması	En yakın mesafe içerisinde bulunan k- adet komşu veriden öğrenme stratejisi kullanılır. Veri setine yeni dâhil edilen bir veri kaydının araştırılan değişken değeri, belli bir uzaklık ölçüsü limitlerindeki k adet komşu veriden benzerlik ataması yöntemidir.
	Rassal Orman Algoritması	Karar ağaçlarından oluşmuş bir karar ormanı yapısıdır. Kullanılan değişkenlerin rassal olarak seçilmesiyle dallanma yapılır. Rassal olarak oluşturulan karar ağacı modellerinin eğitimi gerçekleştirilir ve çoğunluk oylamasıyla, regresyon problemine yönelik birleştirmede ise ortalama kullanılır.
	Stokastik Regresyon	Bağımsız değişkenler ve normal dağılan rassal hata teriminden oluşan stokastik regresyon modeller kullanılır. Eksik değerler için tahmin edilmiş olan her değere rassal bir hata terimi eklenir.
	Amelia Algoritması	Bootstrap ve beklenti maksimizasyonunu içeren çoklu tamamlama mekanizmasıdır. Farklı örneklemler ile eksik verilere sahip alt veri setleri elde edilir. Birden fazla model kullanılıp her bir modelde, bootstrap ve beklenti maksimizasyonuyla tamamlanan eksik verilerin ortalaması alınıp bu sonuçlar birleştirilir.
	Naive Bayes Algoritması	Değişkenlerin birbirinden koşullu bağımsız olduğu varsayımıyla, hedef değişken değerleri bilinen veri kayıtlarından yararlanarak yeni bir kaydın, tanımlanmış olan her bir hedef değişken değerini alması olasılığı (koşullu) hesaplanmaktadır. En büyük koşullu olasılık değerine sahip olan hedef nitelik/değişken değeri, o veri kaydının hedef nitelik/değişken değeri olarak atanmaktır.

Kaynak: Yazar tarafından oluşturulmuştur.

ÜÇÜNCÜ BÖLÜM
PERFORMANS ÖLÇÜTLERİ VE PERFORMANSLARIN DEĞERLENDİRİLMESİ

3.1. PERFORMANS YAKLAŞIMI

Modellerin ve operatörlerin gösterdiği performansların anlaşılabilmesi ve başka operatörlere ait performanslarla karşılaştırılmasının yapılabilmesi için bazı metriklere ihtiyaç duyulmaktadır. Bu ihtiyaç hem kavramsal boyutta hem de değerlerle ifade edilebilecek bir boyutta olmalıdır. Aynı zamanda bahsedilen metriklerin sahip olduğu seviyelerin de ayrıca bir anlam taşıması, ilgili model veya operatörün standart olarak belirlenmiş bir eşik varsa, o eşiğe göre durumunu da ortaya koyulabilmesi dolayısıyla bu şekilde bir fonksiyonel bir özellik taşıması da oldukça önemlidir. Bu çerçevede, makine öğrenmesinde diğer disiplinlerde olduğu gibi birtakım istatistik ve değerlerden elde edilen, genel olarak performans değerlendirme ölçütleri şeklinde ifade edilen kriterler söz konusudur. Bu kriterler, makine öğrenmesi uygulamalarına yönelik çalışmalarda, algoritma ve modellerin başarılarına ilişkin sorgu çıktısı olarak kullanılmaktadır.

Uygulama bölümünde detaylı bir şekilde ele alınacağı üzere çalışma kapsamında kullanılacak olan veri seti manipüle edilerek farklı oranlarda eksiltilecektir. Eksiltilmiş veri setinin, önceki bölümlerde açıklanan yöntemlerle tamamlanmasının ardından tamamlanma operasyonları ile elde edilmiş tüm veri setleri, yine önceki bölümlerde açıklanmış olan Naive Bayes algoritmasıyla sınıflandırma işlemine tabi tutulacaktır. Bu noktada, algoritmaların eksik verileri tamamlama operasyonlarının, tamamlanmış olan tüm veri setleri için aynı şartlarda gerçekleştirilen sınıflandırma işlemine etkileri ortaya çıkacaktır. Bu amaçla, bir sonraki kısımda ele alınacak olan performans değerlendirme kriterleri ortaya koyulup açıklanacaktır. Böylece, algoritmaların bu uygulamadaki performansları değerlendirilip, sınıflandırma performansların değerlendirilmesi ve diğer sınıflandırma performansları ile karşılaştırılması işlemi gerçekleştirilecektir.

3.2. PERFORMANS DEĞERLENDİRME KRİTERLERİ VE ANLAMLARI

Performans ölçütleri ve bu ölçütlerin aldığı değerler "Confusion Matrix" adı verilen bir matris yapısında elde edilmektedir (Sokolova ve Lapalme, 2009: 429). Bu çalışmada da yürütüleceği gibi sınıflandırma başarısının tespiti, doğru ve yanlış sınıflandırılmış olan örnek ya da kayıt sayısına bağlı olarak tespit edilmektedir (Başer ve diğerleri, 2021: 115). Bahsedilen önek sayılarının verildiği "Confusion Matrix" Tablo 6'da ifade edilmiştir.

Tablo 6: Confusion Matrix

		Tahminlenen Sınıf	
		Sınıf A	Sınıf B
Gerçek	Sınıf A	a	b
	Sınıf B	c	d

Kaynak: Yazar tarafından oluşturulmuştur.

Verilen matris ışığında performans ölçütleri ve bu kriterlerin aldığı değerler olan sırasıyla a, b, c ve d değerleri ortaya çıkmaktadır. a: true positive (TP), b: false negative (FN), c: false positive (FP) ve d: true negative (TN) değerlerini ifade etmektedir (Almuniri ve Said, 2017: 59). Bu değerler, sınıflandırması yapılmış olan her kaydın, gerçek sonucu ile sınıflama sonucunda tahminlenmiş sonucunun kontrolünde oluşmaktadırlar. Örneğin: a=5 olması durumu, gerçekte Sınıf A içerisinde bulunup sınıflama sonrasında, 5 adet kaydın sınıf değerinin yine A olarak atanması sonucunu gösterir. Farklı bir örnek göstermek gerekirse, c=15 durumu ele alınabilir. Bu örnekte, gerçekte sınıf olarak B sınıfına üye olan fakat sınıflandırma işlemi neticesinde sınıf değerinin A olarak atandığı 15 adet kaydın olduğu sonucu çıkmaktadır.

Makine öğrenmesi ve veri madenciliği operasyonlarında, operasyonların değerlendirilmesinde kullanılan ölçütler, sınıflandırma için harcanan süre, doğru sınıflandırma oranı, kesinlik, duyarlılık, F-ölçümü ve ROC Alanı olarak ifade edilebilir.

Sınıflandırma için harcanan süre, kısaca sınıflandırma süresi şeklinde ifade edilebilir. Bu kriter sınıflandırmanın ne kadar zaman aldığını göstermektedir. İlk başta çok dikkat çekici gelmese de sınıflandırma süresi, özellikle çok büyük hacimli veri setleri ile çalışıldığı durumlarda oldukça öne çıkan bir kıstas haline gelmektedir.

Doğru sınıflandırma oranı veya doğruluk oranı, sınıf değeri doğru olarak atanmış kayıt sayısının, sınıflandırılan tüm kayıt sayısına oranını olarak tanımlanmaktadır.

$$\text{Doğruluk} = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

Kesinlik (Precision) oranı, (8) numaralı eşitlik ile elde edilmektedir. Performans değerlendirme kriterleri içerisinde bu ölçüt, aynı şekil ve yöntem ile elde edilmiş analiz çıktılarının birbirine olan yakınlığının bir ifadesidir.

$$\text{Kesinlik} = \frac{TP}{TP+FP} \tag{8}$$

Şekil 11: Doğruluk- Kesinlik İlişkisi

Kaynak: OACON, 20.11.2023

Bir diğer kriter olan duyarlılık (Recall) oranı, yapılan sınıflandırma operasyonuna ait sonuçların birbirine olan yakınlığının bir ifadesi olup (9) numaralı eşitlik ile hesaplanmaktadır (Miao ve Zhu, 2022: 1547)

$$\text{Duyarlılık} = \frac{TP}{TP+FN} \tag{9}$$

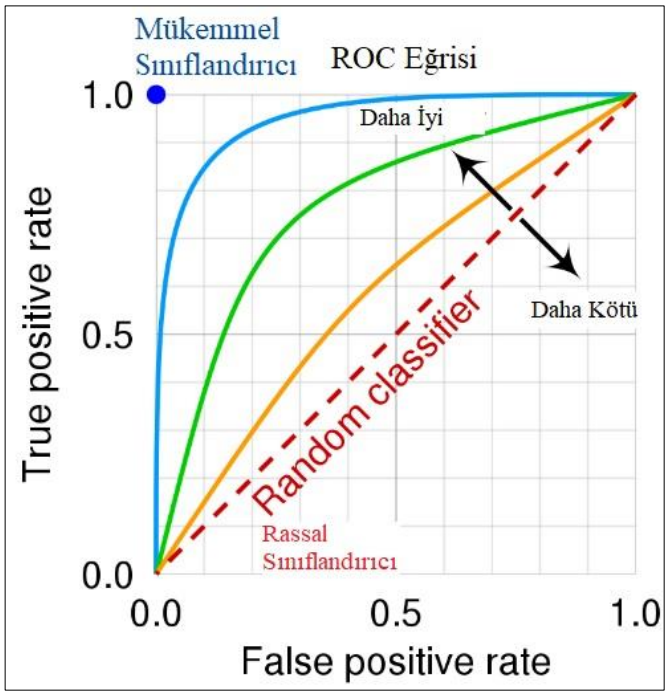
Her ne kadar duyarlılık ve kesinlik ölçütleri geçerli, kabul gören ve önemli performans değerlendirme ölçütleri olsalar da kıyaslayıcı bir unsur olarak tek başlarına yeterli kalitede olmayabilirler F-ölçümü (F-measure) veya F-ölçütü, açıklanmış olan

duyarlılık ve kesinlik ölçütlerinin bir bakış açısıyla kombinasyonundan oluşmaktadır. Bu iki ölçütün, her ikisi için eşit ağırlıklı olarak harmonik ortalaması F-ölçümü veya F-ölçütünü ortaya çıkarmaktadır (Bender ve diğerleri, 2022: 433). Duyarlılık ve kesinlik ölçütlerinin dengede olduğu durumlarda F-ölçütü yüksek değerlere sahip olmaktadır.

$$\text{F-ölçütü} = \frac{2 \text{ x Duyarlılık x Kesinlik}}{\text{Duyarlılık + Kesinlik}} \tag{10}$$

Performans değerlendirme ölçütlerinden ele alınacak son ölçüt ROC Alanı ölçütüdür. Bu alan, X eksenin FP ve Y ekseninde TP değerlerinin oluşturduğu bir alanı göstermektedir. Bu alan değeri, 1'e ne kadar yakınsa o kadar iyi sınıflandırıcı sonucunu ortaya çıkarmaktadır. Genel olarak 0,5 değerinin altındaki değerler literatürde, kötü bir sınıflandırıcı şeklinde yorumlanmaktadır (Bender ve diğerleri, 2022: 433).

Şekil 12: ROC Alanı



Kaynak: Wikipedia, 20.11.2023

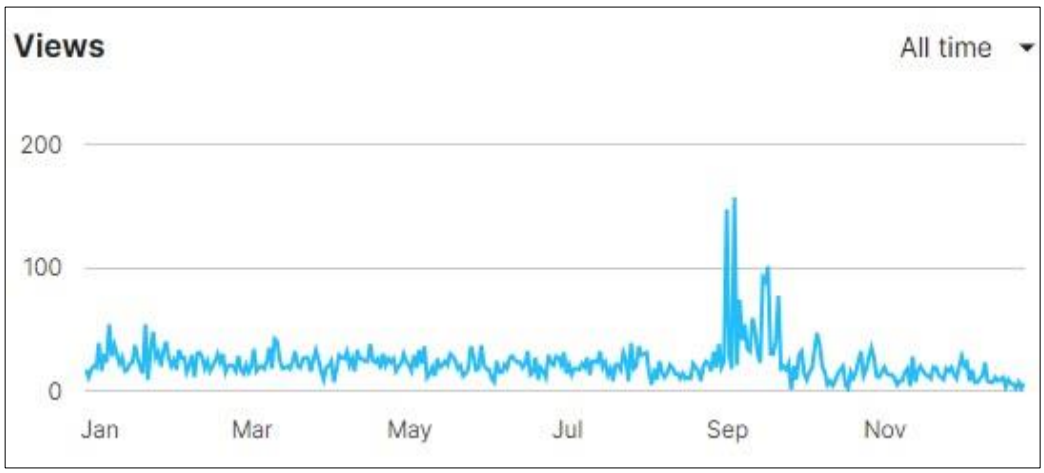
DÖRDÜNCÜ BÖLÜM
UYGULAMA

4.1. VERİ SETİ

Çalışmanın bu bölümünde, uygulama yapılacak veri seti tanıtılacaktır. Çalışma kapsamında, makine öğrenmesi, veri madenciliği ve çeşitli analizlerde oldukça sık kullanılan ve literatürde oldukça tanınan "Hitters" veri setinden faydalanılacaktır. Bu veri setine, makine öğrenmesi ve veri analizi konularında çalışan kişilerden oluşan online bir topluluk olan Kaggle platformundan ulaşılabilmektedir.

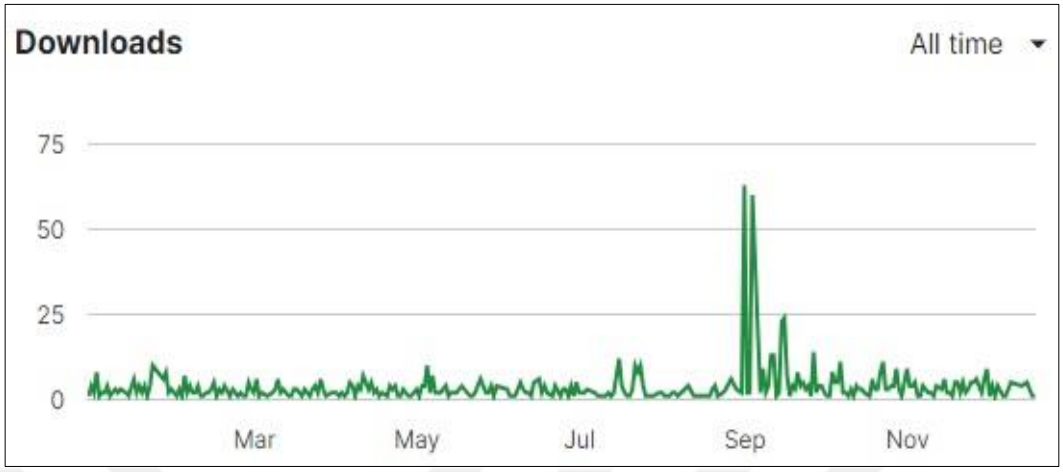
Hitters, 1986-1987 sezonu içerisinde Amerika'daki beyzbol ligi olan Major League'de oynayan beyzbol oyuncularına ait çeşitli istatistikleri sunan özellikle makine öğrenmesi çalışmaların oldukça büyük bir kullanım oranına sahip veri setidir. Hitters veriseti, ISLR R paketinin bir ögesidir aynı zamanda Ridge regresyon ve LASSO uygulamalarının R da nasıl gerçekleştirildiğini ifade eden James ve diğerlerinin (2013) kitap çalışmasında kullanılmıştır.

Şekil 13: Hitters Veri Setinin Aylara Göre Görüntülenme Durumu



Kaynak: Kaggle, 10.01.2023

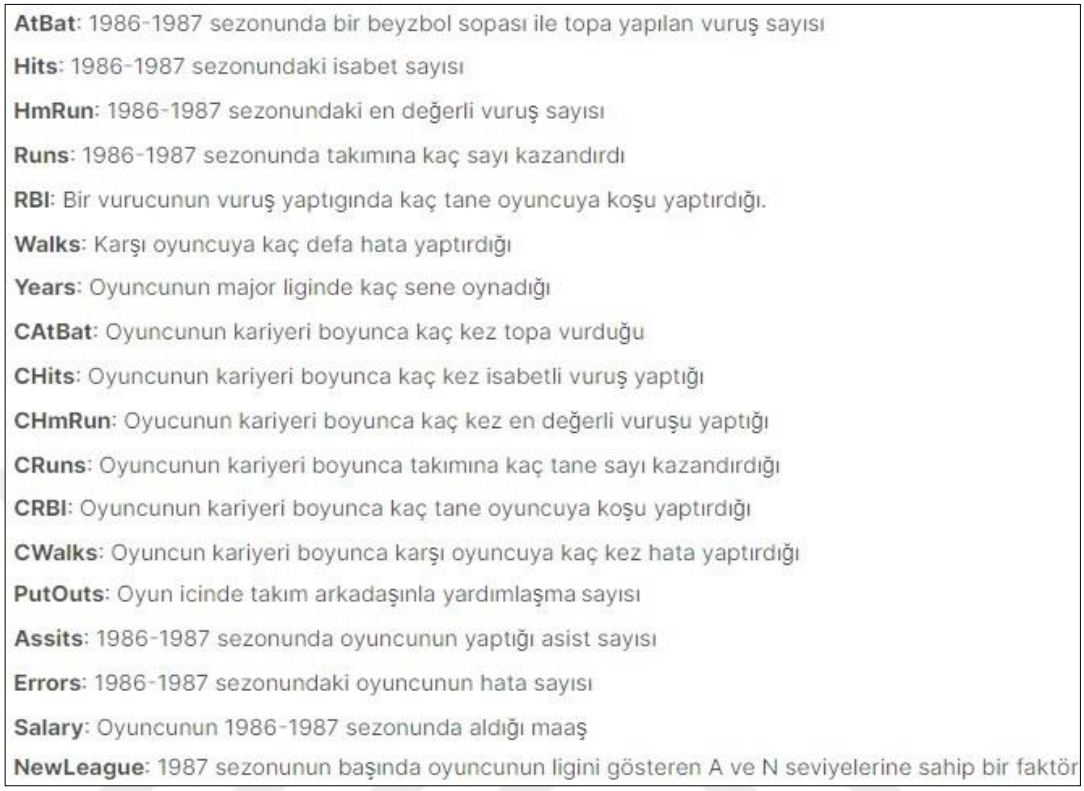
Şekil 14: Hitters Veri Setinin Aylara Göre İndirilme Durumu



Kaynak: Kaggle, 10.01.2023

Böylece, belirtilen sebeplerden dolayı veri setinin bilinirliği, makine öğrenmesi, veri analizi, veri madenciliği gibi çalışmalarda her zaman üst seviyede seyretmektedir. Şekil 13 ve Şekil 14'te gösterilen görüntülenme ve indirilme sayıları da, bahsedilen çalışmalarda sıklıkla tercih edilen bir veri seti olduğunun göstergesidir. Hitters veri setine ait daha detaylı bilgiler, https://www.kaggle.com/datasets web adresinden elde edilebilir. Bu veri setinde, 18 adet değişkenden 322 adet kayıtan yararlanılmaktadır. Şekil 15'te veri setindeki değişkenler ve açıklamaları yer almaktadır.

Şekil 15: Değişkenler



Kaynak: Kaggle, 10.01.2023

Şekil 15'te görüldüğü gibi New League değişkeni haricindeki nitelikler/değişkenler diğer bir ifadeyle değişkenler nümerik değişkenlerdir. New League değişkeni ise iki değer alan kategorik bir değişkendir. Veri setinde yer alan, "New League" değişkeni aynı zamanda çalışmanın sonunda yapılacak olan sınıflandırma işleminin sınıf niteliğidir.

4.2. R PROGRAMLAMA DİLİYLE İLGİLİ GENEL BİLGİLER VE KULLANILAN ARAÇLAR

Veri seti manipüle edilerek %5, %10 ve %15 gibi farklı oranlarda ve rassal olmak üzere R programından faydalanılarak eksiltilecektir. Bu kısımda eksik verilerin rassal olarak eksiltilmesinde ve ardından ortaya çıkan eksik değerlerin tamamlanmasında kullanılacak olan R yazılımından genel olarak bahsedilecektir.

R programlama dili, açık kaynaklı, Auckland Üniversitesi'nden Ross Ihaka ve Robert Gentleman tarafından istatistiksel hesaplamaların yapılabilmesi amacıyla 1993 yılında ortaya çıkmıştır. Böylece, yaratıcıların isimlerin baş harfi bu yazılımın adını da belirlemiştir. Fakat ilk sürümü 2000 yılında kullanıma sunulmuştur. R programlama dili ile S diline benzeyen ve S dilinin farklı bir uygulaması şeklinde ifade edilebilir.

Genel kamu lisansı projesi olan R dili, R Core Team ve "İstatistiksel Hesaplama için R Vakfı" tarafından desteklenmektedir. R, genel olarak istatistiksel hesaplama çerçevesinde ortaya çıksa da, birçok alanda analizler ve uygulamalar gerçekleştirmek için kullanılmaktadır. Doğrusal veya doğrusal olmayan modelleme, istatistik testleri, zaman serilerine yönelik analizler, sınıflandırma ve kümeleme gibi makine öğrenmesi yöntemleri, grafik ve görselleştirmeye dayalı uygulamalar R programlama dilinden faydalanılmaktadır.

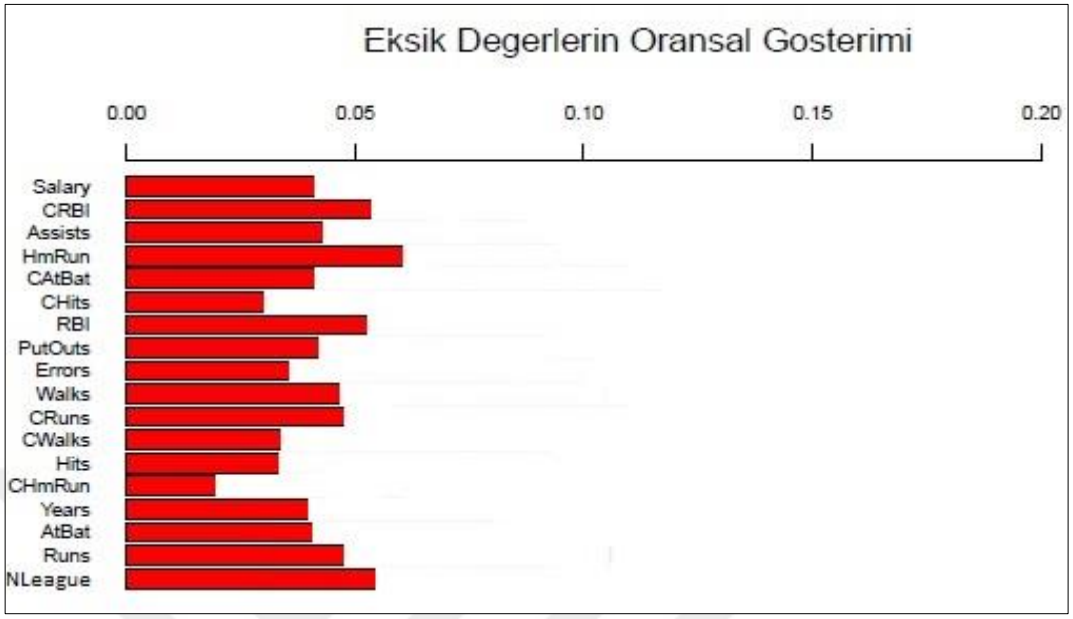
R programlama dilinde, yapılan uygulamalar ve daha birçok özel fonksiyon ve bazı alanlara ilişkin eklenmekte olan paketlerle gerçekleşmektedir. Oldukça büyük sayıda R paketleri, kısaca CRAN olarak ifade edilen "Comprehensive R Archive Network -Kapsamlı R Arşiv Ağı" hemen hemen tüm dosyaların olduğu bir depodadır.

Çalışma kapsamında bahsedilen verilerin rassal olarak eksiltilmesi, eksik değerlerin görselleştirilmesi ve eksiltilen verilerin ilgili yöntemlerle tamamlanması uygulamaları R programında gerçekleştirilmiştir. Bu uygulamalar, mice, missforest, amelia, VIM, lavaan, gglot2, kableExtra, knitr, dplyr, tidyverse, Performance Analytics ve ISLR2, xlsx, DMwR2 kütüphaneleri ve paketlerinden faydalanılarak gerçekleştirilmiştir.

4.3. VERİ SETİNİN RASTGELE EKSİLTİLMESİ VE EKSİK DEĞERLERİN TAMAMLANMASI

Bir önceki başlıkta ifade edildiği gibi veri seti, manipüle edilerek farklı oranlarda aynı zamanda rassal olarak eksiltilmiş ve eksiltilmiş olan değerler şekillerle görselleştirilmiştir. Hitters veri setinin, %5 oranında rassal olarak eksiltilmesi sonucunda, eksiltilmiş olan verilerin, değişkenler ve kayıtlar bazında gösterimi Şekil 16 ve Şekil 17'de gösterilmiş ve eksiltilen veri lokasyonları kırmızı renk ile ifade edilmiştir.

Şekil 16: %5 Eksiklik Oranında Değişkenlerdeki Eksik Değer Oranı



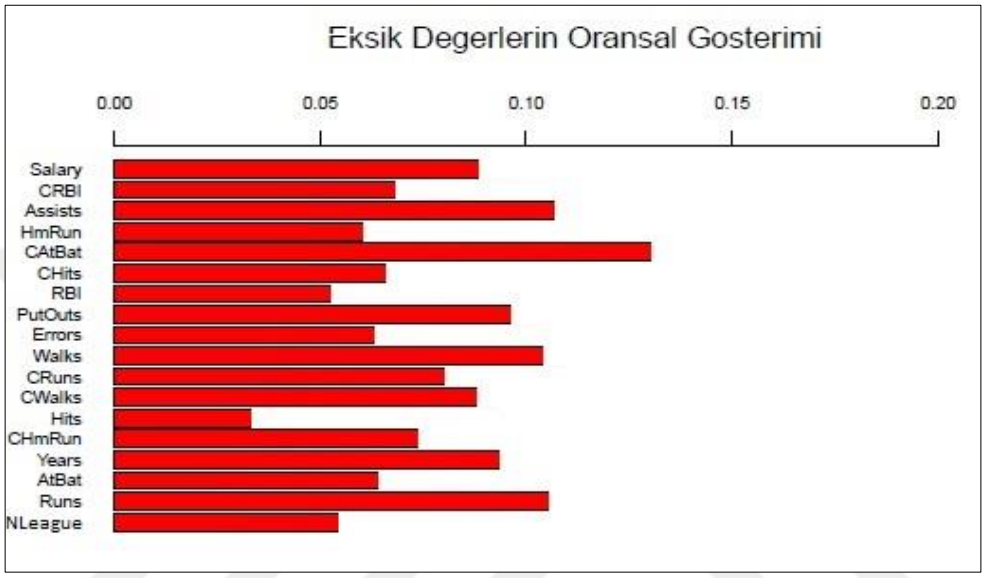
Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 17: %5 Eksiklik Oranında Rassal Eksik Kayıtlar

Kaynak: Yazar tarafından oluşturulmuştur.

Aynı şekilde veri setindeki eksik oranı %10 olmak üzere değerler rassal olarak eksiltilmiştir. Bu manipülasyonun veri seti üzerinde oluşturduğu yeni durum Şekil 18 ve Şekil 19'da nitelikler/değişkenler ve kayıtlar üzerinden ifade edilmiştir.

Şekil 18: %10 Eksiklik Oranında Değişkenlerdeki Eksik Değer Oranı



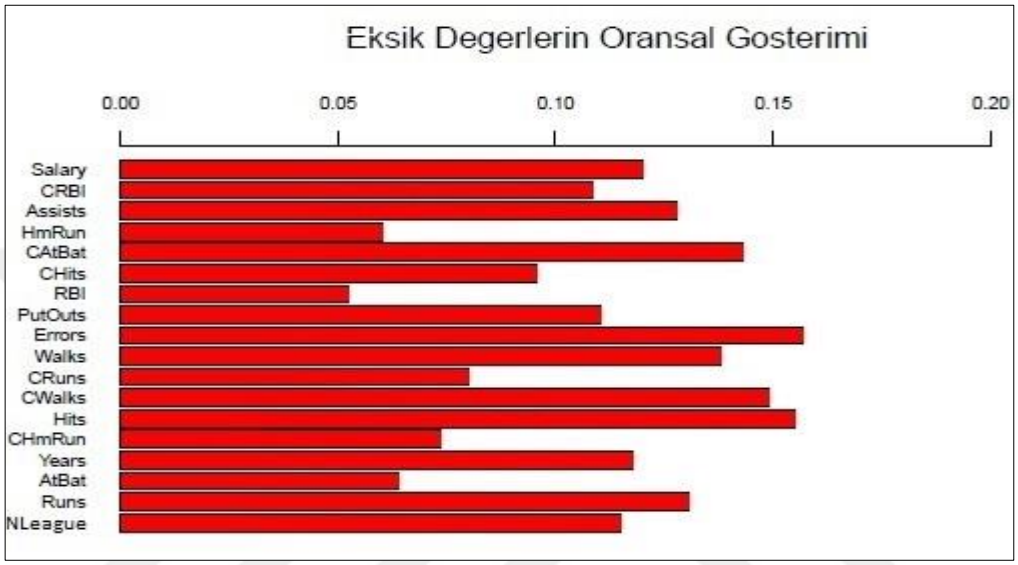
Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 19: %10 Eksiklik Oranında Rassal Eksik Kayıtlar

Kaynak: Yazar tarafından oluşturulmuştur.

Veri setinin %5 ve %10 eksiltilmesinin ardından veri setinin rastgele şekilde %15'i manipüle edilerek eksiltilmiştir. Eksiltme işleminin ardından ortaya çıkan durum Şekil 20 ve Şekil 21'de gösterilmiştir.

Şekil 20: %15 Eksiklik Oranında Değişkenlerdeki Eksik Değer Oranı



Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 21: %15 Eksiklik Oranında Rassal Eksik Kayıtlar

Kaynak: Yazar tarafından oluşturulmuştur.

Böylece, R programı kullanılarak yapılan manipülasyon sonucu, veri seti %5, %10 ve %15 oranlarında rassal olarak eksiltilmiştir. Veri setindeki eksiltilen değerlerin, nitelikler/değişkenler ve kayıtlar üzerinden görselleştirilmesi tamamlanmıştır.

Veri setinin farklı oranlarda eksiltilmesinin ardından, birinci ve ikinci bölümde açıklanan yöntemlerle veri setindeki eksik değerler, R programı ve daha önce bahsedilen kütüphanelerden yararlanarak tamamlanmıştır.

Böylece, eksiltilmiş olan tek bir veri setinden, ilgili veri setindeki eksik değerlerin farklı yöntemlerle tamamlanması ile her eksiklik oranı için bir anlamda yöntem sayısı kadar yeni veri setleri elde edilmiştir. Yeni veri setleri elde etme şeklinde ifade edilen bu operasyonlar, %5, %10 ve %15 seviyelerindeki her eksik değer oranı için yapılmıştır. Kıyaslama kalitesini yükseltmek adına, veri setinin orijinal yani eksiltilmemiş hali de farklı oranlarda eksiltilip ardından eksik değerleri tamamlanarak elde edilen tüm yeni veri setlerine arasına alınmıştır. Eksik verileri R programında, daha önce ifade edilen kütüphaneler ve fonksiyonlarla tamamlanan veri setlerinin yanı sıra veri setinin orijinal haline, "New League" değişkeni hedef nitelik/değişken olacak şekilde Naive Bayes algoritmasıyla sınıflandırma işlemi uygulanmıştır. Sınıflandırma uygulaması, WEKA programında gerçekleştirmiştir. Uygulama sonuçlarının ortaya koyulmasından ve yorumlanmasından önce WEKA programında kısaca bahsedilecektir.

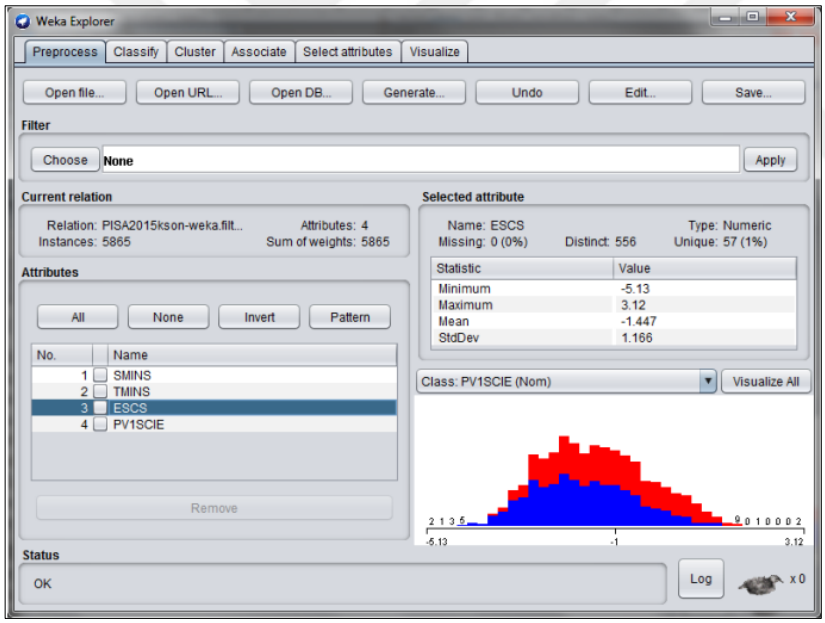
4.4. WEKA PROGRAMI VE SINIFLANDIRMA

WEKA (Waikato Environment For Knowledge Analysis) programı, Yeni Zelanda'da bulunan Waikato Üniversitesi tarafından geliştirilmiş, veri madenciliği ve makine öğrenmesi operasyonlarında oldukça sık bir kullanıma sahip GNU lisansına sahip bir yazılımdır. WEKA'nın Java dilinde geliştirilmiş olması, Java projelerine entegrasyonu kolay hale getirmektedir. Bu durum ve açık kaynak kodlu olması yazılımın yaygın bir kullanıma kavuşmasını destekleyen bir özellik olmuştur. Kolay bir kullanımı ve karmaşık olmayan bir ara yüze sahip olması, yazılımın modüler yapısını ortaya koyan unsurlardandır. WEKA ile ilgili tüm detaylara https://www.cs.waikato.ac.nz/ml/weka/index.html web sayfası üzerinden eşim

sağlanabilmektedir. Ayrıca, https://waikato.github.io/weka-wiki/downloading_weka/ web sayfasından WEKA programı indirilebilmektedir.

WEKA, kendisine özel olan ve "Arff (Attribute Relationship File Format)" adını taşıyan bir dosya türü ile çalışmaktadır. Temel veri madenciliği ve makine öğrenmesi algoritmalarını barındıran bu program, yayınlanan her güncelleme ile kullanılabilecek algoritma sayısının artırılması ve birtakım özelliklerin daha iyi hale getirilmesiyle her geçen gün daha çok kullanıcıya ulaşmakta ve fonksiyonel olma kalitesi yükselmektedir. Bunun dışında WEKA, veri ön işleme ve görselleştirme operasyonları, sınıflandırma, kümeleme, birliktelik kuralları, özellik seçme, eksik veri tamamlama ve daha birçok amaçla kullanılmaktadır.

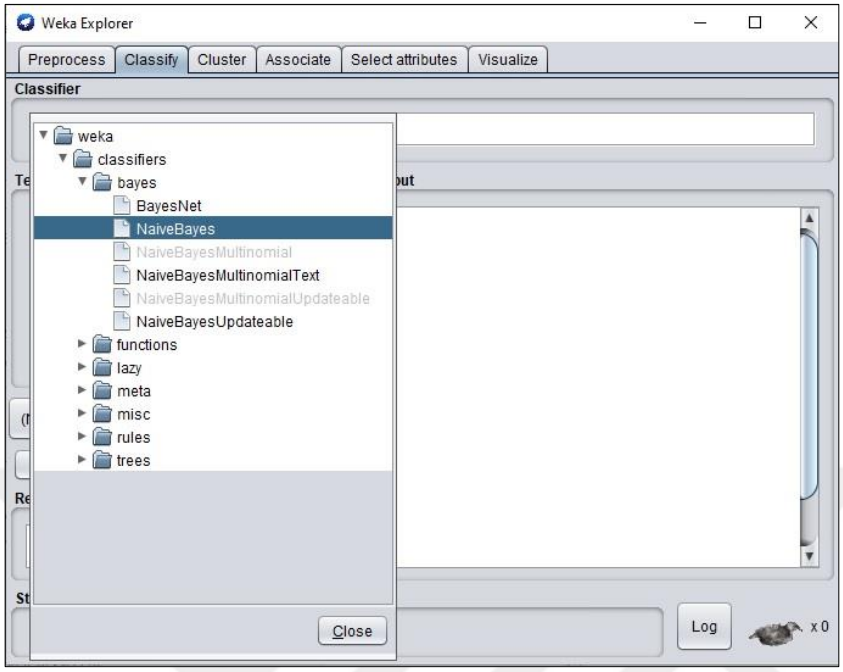
Şekil 22: WEKA- Veri Önişleme Menüsü



Kaynak: Aksu ve Doğan, 2019: 91.

Şekil 23'te çalışma kapsamında gerçekleştirileceği ifade edilen sınıflandırma işlemine ilişkin seçim ve özellik ekranı yer almaktadır. Naive Bayes algoritması ile yapılacak olan bir sınıflandırma işlemine ait girişler bu şekilde ifade edilmiştir.

Şekil 23: WEKA- Naive Bayes Algoritması Sınıflandırma Ekranı



Kaynak: Yazar tarafından oluşturulmuştur.

İfade edildiği gibi, farklı oranlarda eksiltilen veri seti, her eksiklik oranı için açıklanan yöntemlerle tamamlanmıştır. Ardından, bu veri setlerinde "New League" niteliği hedef nitelik/değişken olacak şekilde WEKA programı kullanılarak Naive Bayes algoritmasıyla sınıflandırma işlemi yapılmıştır. Sınıflandırma işlemi, test ayarlarında "Cross Validation" seçeneği ile "Fold" değeri 10 alınarak yapılmıştır. Bu işlem, sınıflandırma yapılacak veri setinin 10 adet parçaya bölüneceğini göstermektedir. Her bir parça sürekli bir şekilde değiştirilerek, elde edilen 10 parçanın 9 tanesinin eğitim verisi ve kalan 1 parçanın da test verisi olarak kullanılması anlamını taşımaktadır.

4.5. VERİ SETLERİNİN SINIFLANDIRILMASI VE PERFORMANS ÖLÇÜTLERİNİN ALDIĞI DEĞERLER

Önceki başlıkta ifade edildiği gibi farklı eksiklik oranına sahip veriler, yine çalışma kapsamında açıklanan yöntemlerle tamamlanmıştır. Hem orijinal hem de

tamamlanmış olan veri setleri “New League” niteliği hedef nitelik/değişken olacak şekilde Naive Bayes algoritması ile WEKA programı kullanılarak sınıflanmıştır. Sınıflandırma işlemlerine ait sonuç çıktıları şekiller üzerinden verilecektir.

Orijinal veri setinin yani veri setinin eksiltilmemiş halinin Naive Bayes algoritmasıyla sınıflandırılmasına ait sonuçlar ve performans ölçütlerinin aldığı değerler Şekil 24’de gösterilmiştir.

Şekil 24: Orijinal Veri Setinin Sınıflandırılması ve Performans Değerleri

```
Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         282               87.5776 %
Kappa statistic                          0.8081
Mean absolute error                      0.0978
Root mean squared error                  0.2735
Relative absolute error                 22.5981 %
Root relative squared error             58.8138 %
Total Number of Instances              322

=== Detailed Accuracy By Class ===
                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
Weighted Avg.    0,876    0,072    0,877      0,876    0,876      0,807    0,961
```

Kaynak: Yazar tarafından oluşturulmuştur.

Orijinal veri setinin sınıflandırılmasının ardından, %5 oranında rassal olarak eksiltilip bahsedilen yöntemlerle tamamlanmasıyla elde edilen veri setlerinin sınıflandırma sonuçları ve performans ölçütlerinin aldığı değerler, her bir tamamlama yöntemi için sıradaki şekiller ile ifade edilmiştir.

Şekil 25: Liste Boyunca Silme Yöntemiyle Tamamlanmış Veri Setinin Sınıflandırılması ve Performans Değerleri (%5 Eksiklik Oranı)

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         177               55.047  %
Kappa statistic                          0.4615
Mean absolute error                      0.2189
Root mean squared error                  0.3275
Relative absolute error                 68.8001 %
Root relative squared error             82.1168 %
Total Number of Instances              311

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   MCC      ROC Area
Weighted Avg.    0,578     0,117     0,497       0,578    0,527       0,439    0,888
```

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 26: Son Gözlemi İleri Taşıma Yöntemiyle Tamamlanmış Veri Setinin Sınıflandırılması ve Performans Değerleri (%5 Eksiklik Oranı)

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         220               68.323  %
Kappa statistic                          0.6024
Mean absolute error                      0.1282
Root mean squared error                  0.3227
Relative absolute error                 40.3034 %
Root relative squared error             80.9144 %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   MCC      ROC Area
Weighted Avg.    0,683     0,083     0,717       0,683    0,674       0,611    0,937
```

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 27: Ortalama İle Tamamlama Yöntemiyle Tamamlanmış Veri Setinin Sınıflandırılması ve Performans Değerleri (%5 Eksiklik Oranı)

```
Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         259               80.4348 %
Kappa statistic                          0.7468
Mean absolute error                      0.0835
Root mean squared error                  0.2644
Relative absolute error                 27.4489 %
Root relative squared error             67.815  %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
Weighted Avg.    0,804    0,046    0,819      0,804    0,806      0,754    0,963
```

Kaynak: Yazar tarafından oluşturulmuştur.

Birinci bölümde açıklanan temel eksik değer tamamlama yöntemleri kullanılarak oluşmuş veri setlerinin sınıflandırılmasının ardından, aynı eksiklik oranı için eksik değerlerin makine öğrenmesi yöntemleriyle tamamlanması sonucu oluşan veri setlerinin sınıflandırılması yapılmıştır. Sınıflandırma işlemlerine ait özet sonuç çıktısı ve performans değerleri şekilleri ile gösterilmiştir.

Şekil 28: En Yakın k-Komşu Algoritması ile Tamamlanmış Veri Setinin Sınıflandırılması ve Performans Değerleri (%5 Eksiklik Oranı)

```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         285               88.5093 %
Kappa statistic                          0.8554
Mean absolute error                      0.054
Root mean squared error                  0.195
Relative absolute error                 16.9924 %
Root relative squared error             48.9155 %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
Weighted Avg.    0,885    0,030    0,886      0,885    0,885      0,855    0,976
```

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 29: Rassal Orman Algoritmasıyla Tamamlanmış Veri Setinin Sınıflandırılması ve Performans Değerleri (%5 Eksiklik Oranı)

```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         278               86.3354 %
Kappa statistic                          0.8285
Mean absolute error                      0.0562
Root mean squared error                  0.2011
Relative absolute error                 17.6587 %
Root relative squared error             50.4365 %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
Weighted Avg.    0,863    0,035    0,869      0,863    0,864      0,829    0,981
```

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 30: Stokastik Regresyon İle Tamamlanmış Veri Setinin Sınıflandırılması ve Performans Değerleri (%5 Eksiklik Oranı)

```
Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         271               84.1615 %
Kappa statistic                          0.8013
Mean absolute error                      0.0727
Root mean squared error                  0.2259
Relative absolute error                 22.8044 %
Root relative squared error             56.5834 %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
Weighted Avg.    0,842    0,041    0,849      0,842    0,841      0,803    0,957
```

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 31: Amelia Algoritması İle Tamamlanmış Veri Setinin Sınıflandırılması ve Performans Değerleri (%5 Eksiklik Oranı)

```
Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         267               82.9193 %
Kappa statistic                          0.7853
Mean absolute error                      0.0781
Root mean squared error                  0.2366
Relative absolute error                 24.5441 %
Root relative squared error             59.3312 %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
Weighted Avg.    0,829    0,044    0,829      0,829    0,828      0,785    0,958
```

Kaynak: Yazar tarafından oluşturulmuştur.

Böylece, ilk önce %5 eksiklik oranı ile rastgele eksiltilmiş veri setinin, çalışma kapsamında açıklanan yöntemlerle tamamlanmasıyla elde edilmiş yeni veri setlerinin sınıflandırılmasına ilişkin özet sonuçlar ve performans değerleri şekiller itibari ile ifade edilmiştir. Aynı şekilde, %10 eksiklik oranı için yöntemler bazında tamamlanması ile elde edilen yeni veri setleri sınıflandırılmasına ait sonuçlar ve performans ölçütlerinin aldığı değerler sıradaki şekiller üzerinden açıklanacaktır.

Şekil 32: Liste Boyunca Silme Yöntemiyle Tamamlanmış Veri Setinin Sınıflandırılması ve Performans Değerleri (%10 Eksiklik Oranı)

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         147               49.4949 %
Kappa statistic                          0.3453
Mean absolute error                      0.2296
Root mean squared error                  0.4407
Relative absolute error                 65.5793 %
Root relative squared error            113.0132 %
Total Number of Instances              297

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
Weighted Avg.    0,506    0,171    0,507      0,506    0,505      0,339    0,668
```

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 33: Son Gözlemi İleri Taşıma Yöntemiyle Tamamlanmış Veri Setinin Sınıflandırılması ve Performans Değerleri (%10 Eksiklik Oranı)

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         199               61.8012 %
Kappa statistic                          0.5195
Mean absolute error                      0.1539
Root mean squared error                  0.3729
Relative absolute error                 48.4117 %
Root relative squared error             93.5409 %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
Weighted Avg.    0,618    0,095    0,628      0,618    0,617      0,526    0,884
```

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 34: Ortalama İle Tamamlama Yöntemiyle Tamamlanmış Veri Setinin Sınıflandırılması ve Performans Değerleri (%10 Eksiklik Oranı)

```
Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         229               71.118  %
Kappa statistic                          0.6098
Mean absolute error                      0.1108
Root mean squared error                  0.2959
Relative absolute error                 59.39   %
Root relative squared error             75.8817 %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
Weighted Avg.    0,711    0,120    0,703      0,711    0,693      0,601    0,901
```

Kaynak: Yazar tarafından oluşturulmuştur.

%10 eksiklik oranı ile eksiltilmiş olan Hitters veri seti, temel eksik veri tamamlama yöntemleri ile tamamlanarak yeni veri setleri elde edilmiş ve Naive Bayes algoritması ile sınıflandırma işlemi yapılmıştır. Ardından şekillerde belirtilen özet sonuçlar ve performans değerleri elde edilmiştir. Bu aşamada, aynı operasyonlar makine öğrenmesi yöntemleri ile yapılmış ve elde edilen sonuçlar sıradaki şekillerde ifade edilmiştir.

Şekil 35: En Yakın k-Komşu Algoritması ile Tamamlanmış Veri Setinin Sınıflandırılması ve Performans Değerleri (%10 Eksiklik Oranı)

```
Time taken to build model: 0.45 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         279               86.646  %
Kappa statistic                          0.8318
Mean absolute error                      0.0586
Root mean squared error                  0.2249
Relative absolute error                 16.8762 %
Root relative squared error             56.4196 %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
Weighted Avg.    0,866    0,034    0,869      0,866    0,867      0,834    0,982
```

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 36: Rassal Orman Algoritmasıyla Tamamlanmış Veri Setinin Sınıflandırılması ve Performans Değerleri (%10 Eksiklik Oranı)

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         269               83.5404 %
Kappa statistic                          0.793
Mean absolute error                      0.0658
Root mean squared error                  0.2566
Relative absolute error                 20.7167 %
Root relative squared error             64.37   %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
Weighted Avg.    0,835    0,041    0,842      0,835    0,836      0,797    0,897
```

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 37: Stokastik Regresyon İle Tamamlanmış Veri Setinin Sınıflandırılması ve Performans Değerleri (%10 Eksiklik Oranı)

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         254               78.882  %
Kappa statistic                          0.735
Mean absolute error                      0.087
Root mean squared error                  0.2704
Relative absolute error                 27.3899 %
Root relative squared error             67.826  %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
Weighted Avg.    0,789    0,052    0,796      0,789    0,787      0,738    0,957
```

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 38: Amelia Algoritması İle Tamamlanmış Veri Setinin Sınıflandırılması ve Performans Değerleri (%10 Eksiklik Oranı)

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         244               75.7764 %
Kappa statistic                          0.6964
Mean absolute error                      0.0969
Root mean squared error                  0.3113
Relative absolute error                 30.3993 %
Root relative squared error             77.9716 %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   MCC      ROC Area
Weighted Avg.    0,758     0,058     0,759       0,758    0,757       0,700    0,850
```

Kaynak: Yazar tarafından oluşturulmuştur.

Böylece, %10 eksiklik oranı ile rassal olarak eksiltilen Hitters veri seti, hem temel yöntemlerde hem de makine öğrenmesi ile tamamlanmış ve bu veri setlerinin sınıflandırılmasına ilişkin performans ölçütlerinin aldığı değerler belirtilmiştir. Aynı uygulama, Hitters veri setinin, %15 eksiklik oranı ile eksiltilmesi durumu için de gerçekleştirilmiştir. İlgili sonuçlar, sıradaki şekiller yoluyla gösterilmiştir.

Şekil 39: Liste Boyunca Silme Yöntemi İle Tamamlanmış Veri Setinin Sınıflandırılması ve Performans Değerleri (%15 Eksiklik Oranı)

```
Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         102               36.2989 %
Kappa statistic                          0
Mean absolute error                      0.3044
Root mean squared error                  0.3899
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              281

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
Weighted Avg.    0,373    0,373    0,139      0,373    0,202      0,000    0,485
```

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 40: Son Gözlemi İleri Taşıma Yöntemi İle Tamamlanmış Veri Setinin Sınıflandırılması ve Performans Değerleri (%15 Eksiklik Oranı)

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         164               50.9317 %
Kappa statistic                          0.0123
Mean absolute error                      0.503
Root mean squared error                  0.6774
Relative absolute error                101.4688 %
Root relative squared error            136.0495 %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
Weighted Avg.    0,509    0,497    0,510      0,509    0,510      0,012    0,476
```

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 41: Ortalama Atama Yöntemi İle Tamamlanmış Veri Setinin Sınıflandırılması ve Performans Değerleri (%15 Eksiklik Oranı)

```
Time taken to build model: 0.19 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         211               65.528  %
Kappa statistic                          0.5046
Mean absolute error                      0.2593
Root mean squared error                  0.3451
Relative absolute error                 85.1821 %
Root relative squared error             88.4891 %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
Weighted Avg.    0,655    0,189    0,622      0,655    0,599      0,516    0,839
```

Kaynak: Yazar tarafından oluşturulmuştur.

%15'lik eksik oranına sahip Hitter veri setinin temel eksik değer tamamlama yöntemleri ile tamamlanmasından sonra Naive Bayes algoritması ile sınıflandırılmasına ilişkin sonuçlar, bu şekilde gösterilmiştir. Bu sonuçların ardından, aynı eksiklik oranına sahip veri setinin, çalışma kapsamında kullanılan makine öğrenmesi algoritmalarıyla tamamlanmıştır. Tamamlanan veri setleri aynı şekilde, Naive Bayes algoritmasıyla sınıflandırılmıştır. Elde edilen sonuçlar ve performans değerleri sıradaki şekillerde olduğu gibi saptanmıştır.

Şekil 42: En Yakın k-Komşu Algoritması İle Tamamlanmış Veri Setinin Sınıflandırılması ve Performans Değerleri (%15 Eksiklik Oranı)

```
Time taken to build model: 0.07 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         226               70.1863 %
Kappa statistic                          0.6248
Mean absolute error                      0.1982
Root mean squared error                  0.3036
Relative absolute error                 62.1856 %
Root relative squared error             76.0566 %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
Weighted Avg.    0,702    0,077    0,691      0,702    0,694      0,621    0,886
```

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 43: Rassal Orman Algoritması İle Tamamlanmış Veri Setinin Sınıflandırılması ve Performans Değerleri (%15 Eksiklik Oranı)

```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         256               79.5031 %
Kappa statistic                          0.7421
Mean absolute error                      0.0848
Root mean squared error                  0.2673
Relative absolute error                 26.6653 %
Root relative squared error             67.0278 %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
Weighted Avg.    0,795    0,053    0,796      0,795    0,792      0,742    0,953
```

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 44: Stokastik Regresyon İle Tamamlanmış Veri Setinin Sınıflandırılması ve Performans Değerleri (%15 Eksiklik Oranı)

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         215               66.7702 %
Kappa statistic                          0.5871
Mean absolute error                      0.1399
Root mean squared error                  0.3504
Relative absolute error                 43.8798 %
Root relative squared error             87.771  %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
Weighted Avg.    0,668    0,079    0,739      0,668    0,655      0,610    0,935
```

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 45: Amelia Algoritması İle Tamamlanmış Veri Setinin Sınıflandırılması ve Performans Değerleri (%15 Eksiklik Oranı)

```
Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         224               69.5652 %
Kappa statistic                          0.6163
Mean absolute error                      0.1863
Root mean squared error                  0.2954
Relative absolute error                 58.5762 %
Root relative squared error             74.0823 %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
Weighted Avg.    0,696    0,075    0,692      0,696    0,693      0,620    0,903
```

Kaynak: Yazar tarafından oluşturulmuştur.

%5, %10 ve %15 eksiklik oranı ile rassal olarak eksiltilmiş olan Hitters veri seti, çalışma çerçevesinde açıklanan yöntemleriyle tamamlanmış ve ardından tamamlama operasyonları ile elde edilen yeni veri setleri Naive Bayes algoritması sınıflandırmıştır. Sınıflandırma işlemiyle birlikte elde edilen sonuçlar ve performans değerlendirme kriterlerinin aldığı değerler bu bölümde belirtilen şekiller ile ifade edilmiştir.

Farklı eksiklik oranlarına ilişkin yapılan eksik veri tamamlama işlemlerine yönelik ele edilen bulgular tablolarla aşağıdaki özetlenmiştir. Yanı sıra, temel eksik veri tamamlama yöntemleri ve makine öğrenmesi algoritmaları ile eksik veri tamamlama operasyonlarına ilişkin sonuçlar, verilen tablolar yoluyla her bir performans değerlendirme kriteri için özet şekilde sunulmuştur.

Hitters veri setinin eksiltilmeden Naive Bayes algoritması ile sınıflandırılması sonucu elde edilen performans değerleri Tablo 7'de verilmiştir.

Tablo 7: Hitters Veri Setinin Sınıflandırılmasına Ait Performans Değerleri

Sınıflandırma Süresi (saniye)	**Doğruluk (%)**	**Kesinlik**	**Duyarlılık**	**F-ölçütü**	**ROC Alanı**
0.03	87.5776	0.877	0.876	0.876	0.961

Kaynak: Yazar tarafından oluşturulmuştur.

Tablo 8: %5 Eksiklik Oranında Veri Setinin Tamamlanmasının Ardından Sınıflandırılmasına Ait Performans Değerleri

Yöntemler	Sınıflandırma Süresi (saniye)	Doğruluk (%)	Kesinlik	Duyarlılık	F-ölçütü	ROC Alanı
Liste Boyunca Silme	0.00	55.047	0.497	0.578	0.527	0.888
Son Gözlemi İleri Taşıma	0.00	68.323	0.717	0.683	0.674	0.937
Ortalama İle Tamamlama	0.04	80.435	0.819	0.804	0.806	0.963
KNN	**0.02**	**88.509**	**0.886**	**0.885**	**0.885**	**0.976**
Rassal Orman	0.02	86.335	0.869	0.863	0.864	0.981
Stokastik Regresyon	0.01	84.161	0.849	0.842	0.841	0.957
Amelia	0.03	82.919	0.829	0.829	0.828	0.958

Kaynak: Yazar tarafından oluşturulmuştur.

Tablo 9: %10 Eksiklik Oranında Veri Setinin Tamamlanmasının Ardından Sınıflandırılmasına Ait Performans Değerleri

Yöntemler	Sınıflandırma Süresi (saniye)	Doğruluk (%)	Kesinlik	Duyarlılık	F-ölçütü	ROC Alanı
Liste Boyunca Silme	0.00	49.495	0.507	0.506	0.505	0.668
Son Gözlemi İleri Taşıma	0.00	61.801	0.628	0.618	0.617	0.884
Ortalama İle Tamamlama	0.08	71.118	0.703	0.711	0.693	0.901
KNN	**0.45**	**86.646**	**0.869**	**0.866**	**0.867**	**0.982**
Rassal Orman	0.00	83.540	0.842	0.835	0.836	0.897
Stokastik Regresyon	0.00	78.882	0.796	0.789	0.787	0.957
Amelia	0.00	75.776	0.759	0.758	0.757	0.850

Kaynak: Yazar tarafından oluşturulmuştur.

Tablo 10: %15 Eksiklik Oranında Veri Setinin Tamamlanmasının Ardından Sınıflandırılmasına Ait Performans Değerleri

Yöntemler	Sınıflandırma Süresi (saniye)	Doğruluk (%)	Kesinlik	Duyarlılık	F-ölçütü	ROC Alanı
Liste Boyunca Silme	0.01	36.299	0.139	0.373	0.202	0.485
Son Gözlemi İleri Taşıma	0.00	50.932	0.510	0.509	0.510	0.476
Ortalama İle Tamamlama	0.19	65.528	0.622	0.655	0.599	0.839
KNN	0.07	70.186	0.691	0.702	0.694	0.886
Rassal Orman	**0.02**	**79.503**	**0.796**	**0.795**	**0.792**	**0.953**
Stokastik Regresyon	0.00	66.770	0.739	0.668	0.655	0.935
Amelia	0.06	69.565	0.692	0.696	0.693	0.903

Kaynak: Yazar tarafından oluşturulmuştur.

SONUÇ

Eksik değer veya başka bir ifade ile eksik veri sorunu, veriye ve veri analizine dayanan çalışmalarda oldukça sık karşılaşılan bir sorundur. Bu sorunun göz ardı edilmesi, yapılan analizlerin ve analiz sonuçlarının kalitesini olumsuz yönde etkilemektedir. Eksik değer problemi olmasına rağmen problemin yok sayılarak yani çözümlememiş olduğu durumda yapılmış olan analiz ve araştırılmalarda, sağlıklı ve etkin sonuçlar almak mümkün değildir. Özellikle, eksik değerlerin fazla sayılacak bir seviyede veya önemli bir pozisyonda olduğu veri gruplarında, istatistiksel gerçeklerin yansımış olduğu doğru sonuçları elde etmek ve dolayısıyla bu analizlerden doğru çıkarımlar üretmek neredeyse söz konusu değildir. Ancak göz ardı edilebilecek sayıda, oranda veya konumda verilerin eksik olduğu çalışmalarda, makul derecede hata payları ile doğru sonuçlara ve çıkarımlara erişmek mümkün olacaktır. Dolayısıyla belirtildiği gibi, pratik dünyada oldukça sık karşılaşan eksik değer sorunun mutlak suretle çözümlenmesi gerekliliği ortadadır.

Eksik değer sorununun çözümlenmesi için literatürde birçok yöntem mevcuttur. Bu yöntemler, farklı yaklaşımlarla, var olan eksik değer problemlerini, yaklaşımların amacı ve inşa edildiği durumlara göre ortadan kaldırmayı hedeflemektedir. Eksik değer probleminin çözümü için kullanılan yöntemler içerisinde eksik değer içeren veri unsurlarının veri topluluğundan çıkarılması yani silinmesi yönünde hareket eden yöntemler varsa da genel olarak, yöntemlerin neredeyse tamamı denilebilecek kadar büyük bir oranı eksik verilerin tamamlanması şeklinde bir çözüm üretmektedir.

Gerçekleştirilen tez çalışmasında, eksik değer tamamlama operasyonlarında makine öğrenmesinin kullanılmasının ortaya çıkardığı potansiyel etkilerin yanı sıra kullanılan temel yöntemlerin kullanılmasıyla elde edilen sonuçların kıyaslanması amaçlanmıştır. Bahsi geçen kıyaslama, temel yöntemler ve makine öğrenmesi yöntemleri ile tamamlanmış eksik değerleri bulunan veri seti üzerinden yapılmıştır. Veri setinin manipüle edilerek rastgele eksiltilmesiyle ortaya çıkan eksik değerlerin, birden fazla yöntemle tamamlanması neticesinde, doğal olarak, birden fazla yeni veri seti elde edilmiştir. Bir anlamda, eksik değerleri bulunan bir veri setinden eksik değerleri tamamlanarak birden fazla veri seti ortaya çıkmıştır.

Başlangıçta Hitters veri seti eksiltilmeden yani veri setinin orijinal hali Naive Bayes algoritmasıyla sınıflandırılmıştır. Bu sınıflandırma sonucu, yöntemlerin kendi aralarındaki eksik veri tamamlama operasyonlarının sınıflandırma performansına olan etkilerinin kıyaslanmasına olanak sağlarken aynı zamanda eksik veri tamamlama yöntemlerinin göstermiş olduğu performansların verinin orijinal haline ilişkin performans değerlerine ne derecede yakın olduğunun tespitini de sağlamaktadır.

Orijinal verinin sınıflandırılmasının ardından, Hitters veri seti manipüle edilerek %5 oranında eksiltilmiş ve eksik değerler çalışma kapsamında açıklanan tüm yöntemlerle tamamlanmıştı. Eksik değerlerin, yöntemlerce tamamlanıp elde edilen yeni veri setlerinin Naive Bayes algoritması ile sınıflandırılması sonucu elde edilen performans değerleri Tablo 7'de gösterilmiştir.

%5 rastgele eksiklik oranında, veri setinin eksik değerlerinin bahsedilen yöntemler ile tamamlanması sonucu ortaya çıkan veri setlerinin, Naive Bayes algoritmasıyla sınıflandırılmasına ait sonuçlar belirtilmektedir. Performans değerlendirme ölçütlerinin aldığı değerlerden önce bu sonuçların, eksik veri tamamlama performanslarının sınıflandırma performansına etkilerinin bir özeti olduğunu görmek mümkündür. Bunun nedeni, %5'nin rassal olarak azaltıldığı tek bir veri setinin ilgili yöntemlerce eksik değerlerinin tamamlanıp sınıflandırılmasıdır. Tüm operasyon süreçlerine bakıldığında, süreçlerin içerisinde tek farklı olan eksik değerlerin hangi yöntem ile tamamlandığıdır. Tamamlanma uygulaması için seçilen yöntem dışında, operasyon süreçlerindeki tüm unsurlar aynıdır. Dolayısıyla, performans ölçütlerinin aldığı değerlerin farklılaşmasını sağlayan tek unsur eksik veri tamamlama yöntemleridir. Böylece, bu şartlarda her bir yöntem için eksik verileri tamamlayarak sınıflandırma performansına etkilerinin görülmesi sağlanırken, yöntemleri kendi aralarındaki performans karşılaştırması yapılabilmektedir.

Tablo 8'e genel olarak bakıldığında, %5 rastgele eksiklik oranı için, makine öğrenmesi algoritmalarının temel yöntemlerin sınıflandırma performansına katkısına göre olumlu yönde gözle görülür bir fark ortaya koyduğu görülmektedir. Ortalama ile tamamlama yönteminin ROC alanı değerinde Stokastik Regresyon ve Amelia yöntemlerine göre çok küçük bir farkla daha iyi bir değer aldığı görülürken, genel anlamda nispeten daha iyi bir durumda olma söz konusu değildir. Görüldüğü gibi,

makine öğrenmesi algoritmalarının bariz bir şekilde sınıflandırma performansına olan katkıları temel eksik veri tamamlama yöntemlerinden üstündür.

Temel yöntemlere bakıldığında, öne çıkan özelliğin sınıflandırma süresi olduğu görülmektedir. Bu yöntemlerin, makine öğrenmesi yöntemlerine göre nispeten daha kısa sürede sınıflandırma işlemi yaptığı görülmektedir. Sınıflandırma süresi, özellikle çok büyük hacimli veri setleri ile çalışılırken öne çıkan bir özellik olsa da diğer performans ölçütlerinin aldığı değerlerinin daha düşük seviyelerde olması nedeniyle, bu çalışma içerisinde genel kıyaslama yapılırken avantaj yaratan bir özellik olmamıştır. Onun dışında kalan kesinlik, duyarlılık ve F-ölçütü kriterlerinin aldığı değerlerde üstünlük yine makine öğrenmesi yöntemlerinin tarafındadır. Temel eksik veri tamamlama yöntemleri arasından en çok katkının ortalama ile tamamlama yöntemine ait olduğu görülmektedir. Hem tüm yöntemlere hem de temel eksik veri tamamlama yöntemleri arasında nispeten en düşük veya diğer deyişle en kötü performansı gösteren yöntemin liste boyunca silme yöntemi olduğu görülmektedir.

Makine öğrenmesi yöntemlerine ait performanslarına gelindiğinde, %5 eksiklik oranı için, en iyi performansın En Yakın k- Komşu algoritmasına ait olduğu sonucu gözlenmektedir. Bu sonuca göre, En Yakın k- Komşu algoritması ile Rassal Orman algoritmalarının performans değerleri, orijinal verinin sınıflandırılmasına ait performans değerlerine oldukça yakındır. En Yakın k- Komşu algoritmasını sırasıyla Rassal Orman, Stokastik Regresyon ve Amelia algoritmalarının izlediği görülmektedir. Sadece sınıflandırma süresi ölçütünde, Stokastik Regresyon yönteminin diğer makine öğrenmesi algoritmalarına göre daha iyi bir performansa sahip olduğu tespit edilmiştir. En Yakın k-Komşu algoritması ile Rassal Orman algoritmasının sonuçlarının birbirine çok yakın olduğu gözlenirken diğer yöntemlere ait performans değerlerine doğru bu farkın nispeten büyüdüğü ortadadır.

%5 eksiklik oranı için Tablo 8'deki değerler incelendiğinde, çok önemli bir sonuç daha gözler önüne çıkmaktadır. Eksik değerlerin, %5 eksiklik oranında, En Yakın k-Komşu algoritması ile tamamlanmasıyla oluşan veri setinin sınıflandırılmasına ait performans değerlerinin, veri setinin eksiltilmeden yani orijinal haline ait olan sınıflandırma performansı değerlerinden daha iyi olmasıdır. Burada, En Yakın k-Komşu algoritmasının sınıflandırma performansını yukarıya çektiği gözlenmektedir. Elbette, eksiklik oranı yükseldikçe bu durumun ortaya çıkma

olasılığının düşeceği bir gerçektir ancak %5 gibi istatistiksel araştırmalarda da genel olarak hata payı şeklinde kabul edilen bir düzeyde oldukça çarpıcı bir sonuçtur. Bunun dışında, Rassal Orman, Stokastik Regresyon ve Amelia algoritmalarının veri setinin orijinal haline ilişkin sınıflandırma sonuçlarına oldukça yakın bir performans gösterdikleri ortadadır.

Böylece, çalışma kapsamındaki açıklanan şartlar ve %5 rastgele eksiklik oranı çerçevesinde, makine öğrenmesi algoritmalarının temel eksik veri tamamlama yöntemlerine göre daha iyi bir eksik veri tamamlama performansı gösterdiği sonucu, orijinal veri setine ait sınıflandırma sonuçları ile kıyaslandığında ortaya çıkmaktadır.

Hitters veri setinin %10'luk bir oranda rassal olarak eksiltilip makine öğrenmesi ve temel eksik veri tamamlama yöntemleriyle tamamlanmasının ardından Naive Bayes algoritması ile sınıflandırılmasına ilişkin sonuçlar Tablo 9'da gösterilmiştir. Tablo 9'da gözlenen sonuçlara bakıldığında, performans ölçütlerinin aldığı değerler ışığında, eksiltilen değerlerin makine öğrenmesi algoritmaları tarafından tamamlanmasıyla elde edilen veri setlerinin sınıflandırılması operasyonlarının daha başarılı olduğu gözlenmiştir. Eksik değerlerin temel eksik veri tamamlama yöntemleriyle tamamlanarak elde edilen veri setlerine ait sınıflandırma sonuçlarına ilişkin performans ölçütlerinin aldığı değerlerin anlamlı şekilde makine öğrenmesi algoritmalarına ait sonuçlardan daha düşük seviyede kaldığı gözlenmektedir.

Makine öğrenmesi algoritmaları arasında %5'lik eksiklik oranı ile ilgili sonuçlarda gözlendiği gibi %10 rassal eksik veri oranı için de En Yakın k-Komşu algoritmasının en iyi performansı gösterdiği görülmektedir. Makine öğrenmesi algoritmalarının kendi arasındaki karşılaştırılmada, en iyi performanstan en düşük seviyeli performans sıralamasının yine aynı şekilde gerçekleştiği görülmektedir. En iyi performansı gösteren En Yakın k-Komşu algoritmasıyla Rassal Orman algoritmasına ilişkin performans değerleri oldukça yakın gerçekleştirmiştir. İlk duruma göre eksiklik oranın 2 katına çıkmasına rağmen genel olarak makine öğrenmesi algoritmalarının eriştiği performans değeri seviyelerinin özellikle En Yakın k-Komşu algoritması ile Rassal Orman algoritmasına ilişkin performans değerlerinin iyi bir şekilde korunduğu sonucu da ortaya çıkmıştır. Aynı zamanda, önemli bir sonuç olarak yine bu iki

algoritmanın Hitters veri setinin eksiltilmemiş haline ilişkin sınıflandırma operasyonuna ait performans değerlerine oldukça yakın seyrettiği gözlenmiştir.

Diğer taraftan, %10'luk eksiklik oranı çerçevesinde, eksik değerlerin tamamlanması işlemlerinin, temel eksik değer tamamlama yöntemleriyle yapılmasıyla elde edilen veri setlerinin sınıflandırmasına ait performans değerleri incelendiğinde, bu değerlerin ciddi oranda düştüğü ilk bakışta farkedilmektedir. Temel eksik veri tamamlama yöntemlerinin kendi aralarındaki performans kıyaslamasında, iyiden kötüye doğru sıralamanın değişmediği ve en iyi performansın yine ortalama ile tamamlama yöntemine ait olduğu görülmektedir. Bu yöntemlere ilişkin performans değerleri yine incelendiğinde, eksiklik oranının 2 katına çıkmasıyla, performans değerlerinde gerçekleşen dramatik düşüşün, bu yöntemlerin eksiklik oranındaki artışlara daha duyarlı olduğu ve makine öğrenmesi algoritmalarına ilişkin sonuçlardaki durumun aksine, performans değerlerindeki seviyeleri koruyamadıkları gözlenmektedir.

Böylelikle, veri setindeki eksikliğin 2 katına çıkarılarak yani %10 seviyesinde rassal olarak eksiltilerek eksik değerlerin, hem temel eksik değer tamamlama yöntemleriyle hem de makine öğrenmesi yöntemleriyle tamamlanmasıyla elde edilen yeni veri setlerinin, Naive Bayes algoritmasıyla sınıflandırılmasına ilişkin sonuçların %5'lik eksiklik oranına ait sonuçlar ile sıralama ve üstünlük anlamında paralellik gösterdiği gözlenmiştir. Bu paralellik doğrultusunda, yine makine öğrenmesi algoritmalarının temel eksik değer tamamlama yöntemlerine göre daha iyi bir performans gösterdiği ortadadır. Bunun dışında, makine öğrenmesi algoritmalarına ilişkin sonuçların, veri setinin eksiltilmemiş haline ait sınıflandırılması sonuçlarına oldukça yakın olduğu ve eksiklik oranın 2 katına çıkmasına rağmen performans değerlerinin korunduğu da saptanmıştır. Diğer taraftan, temel eksik değer tamamlama yöntemlerine ait performans değerlerinin, makine öğrenmesi algoritmalarının ulaştığı performans değerlerinin gerisinde kaldığı gibi aynı zamanda %5'lik eksik değer oranın için gözlenen sonuçlar arasında olumsuz yönde, gözle görülür bir performans düşüşü saptanmıştır. Bunun sonucunda, %10'luk rassal eksiklik oranı çerçevesindeki değerlerin, veri setinin eksiltilmemiş haline ilişkin performans değerlerinden oldukça uzak değerler olduğu da görülmüştür.

Hitters veri setinin, rassal olarak %15 oranında eksiltilerek çalışma kapsamında kullanılan yöntemlerle tamamlanması sonucunda Naive Bayes algoritması ile sınıflandırılması uygulaması yapılmıştı. Bu uygulamanın sonuçlarının özetlendiği Tablo 10'a bakıldığında, tıpkı %5 ve %10 eksiklik oranına ilişkin sonuçlarda açık şekilde görüldüğü gibi %15 oranındaki rassal eksiklik durumunda da makine öğrenmesi yöntemlerinin eksik verileri tamamlamasıyla elde edileren yeni veri setlerine ait sınıflandırma sonuçlarında, temel eksik veri tamamlama yöntemlerine göre daha başarılı olduğu görülmüştür.

Makine öğrenmesi algoritmalarının %15'lik eksiklik oranına ilişkin sonuçlarında, performans ölçütlerinin aldığı değerler ışığında ortaya bir fark koyacak şekilde daha iyi performans göstermesinin yanı sıra %5 ve %10'luk eksiklik oranlarındaki sonuçlardan farklı olarak Rassal Orman algoritmasının en iyi performans gösterdiği gözlenmiştir. En Yakın k- Komşu algoritması, Stokastik Regresyon ve Amelia algoritmalarının ise yine temel eksik veri tamamlama yöntemlerinden daha iyi sonuçlar elde etmesiyle beraber birbirlerine oldukça yakın performanslar gösterdikleri de görülmektedir. Makine öğrenmesi yöntemlerine ilişkin Tablo 10'daki sonuçlardan elde edilebilecek bir diğer çıkarım da %5 ve %10 eksiklik oranına ilişkin sonuçlara göre %15'lik eksiklik oranına dair sonuçlarda performans ölçütlerinin nispeten daha düşük değerler almasıdır. Doğruluk oranı değerinin, diğer eksiklik oranlarına ait sonuçların aksine artık %80 ve üzerinde bir değerin elde edilmediği ve aynı şekilde diğer performans ölçütü değerlerinin diğer eksiklik oranına için olan değerlerden farkedilir şekilde daha düşük seviyelere kaydığı da gözlenmektedir.

Diğer taraftan, %15'lik rassal eksiklik oranına ve temel eksik veri tamamlama yöntemlerine ilişkin sonuçlara bakıldığında, en iyi sonucun yine eksik verilerin ortalama ile tamamlama yöntemi ile tamamlanarak elde edilen yeni veri setine sınıflandırma operasyonuna ait olduğu görülmektedir. Diğer eksiklik oranlarında ortaya çıkan sonuçlarda gerçekleşmiş performans sıralaması, yöntemler için aynı şekilde yer bulmuştur. Ortaya çıkan performans değerlerindeki önemli sonuçlardan belki de en çok göze çarpanı, temel eksik veri tamamlama yöntemlerinin dolaylı olarak oluşturduğu sonuçların dramatik seviyelere gerilediğidir. Bu yöntemlerde, doğruluk oranının %15 eksiklik oranı için %50 ve altındaki seviyelere düşen bir durumun ortaya çıktığı görülmektedir. Liste Boyunca Silme ve Son Gözlemi İleri Taşıma

yöntemlerinde yaşanan bu gerileme de, sınıflandırma işleminde Liste Boyunca Silme yöntemi için her 10 kaydın yaklaşık 6'sı, Son Gözlemi İleri Taşıma Yönteminde ise yaklaşık olarak her iki kayıttan birinin yanlış sınıflandırıldığı anlamına gelmektedir. Bundan dolayı, özellikle bu iki yöntem için, başta doğruluk oranı olmak üzere diğer performans ölçütü değerleri oldukça kötü bir performansın göstergesi niteliğindedir.

Makine öğrenmesi algoritmalarını ilişkin sonuçlarda, performans ölçütü değerlerinde diğer eksiklik oranlarına ait performans değerlerine göre farklılıklar göze çarpmaktadır. Bu ölçütlere bakıldığında, %15'lik eksiklik oranına dair sonuçlar için sınıflandırma sürelerinin arttığı gözlenmektedir. Bu durumun sebebi olarak ilgili eksiklik oranının ilk duruma göre artık 1.5 katına çıkması gösterilebilir çünkü eksiklik oranı arttıkça, veri setindeki eksikliğin tamamlanmasında faydalanılacak temel unsur olan veriler arasındaki muhtemel örüntü zayıflamaktadır. Bu nedenle, eksik veriler yerine tamamlanacak olan değerlerin nispeten daha zayıf bir örüntü ağından elde edilmesi sınıflandırma sürecini uzatmaktadır. Aynı şekilde, örüntünün nispeten diğer eksiklik oranıları için ortaya çıkan durumlara göre daha zayıf olması durumunun, bu eksik değerlerin tamamlanarak elde edilen veri setlerinin sınıflandırılma operasyonlarının başarısını değerlendiren diğer performans ölçütü değerlerinde de gözle görülür ve anlamlı bir şekilde düşüş yaşanmasına neden olduğu görülmektedir.

%15'lik eksik oranına ilişkin sonuçların elde edilmesiyle, ele alınan tüm eksiklik oranlarında, eksik veri tamamlama ve sınıflandırma operasyonları tamamlanmış ve performans ölçütü değerleri elde edilmiştir.

Yapılan çalışmada, Hitter veri seti manipüle edilerek sırasıyla %5, %10 ve %15'lik oranlarında rassal olarak eksiltilmişti. Tüm eksiltme işlemlerininin ardından, veri setindeki eksik değerler çalışma kapsamında açıklanmış olan temel eksik veri tamamlama yöntemleri ve makine öğrenmesi algoritmaları ile tamamlanarak bir anlamda yeni veri setleri elde edilmişti. Eksik değerlerin tamamlanmasının ardından yeni veri setleri Naive Bayes algoritması ile tamamlanmıştı. Uygulamanın bu şekilde yapılmasının nedeni, uygulama süreci içerisinde her eksiklik oranı için farklılık oluşturacak tek noktanın eksik veri tamamlama yöntemi olmasıdır. Böylece, Hitters veri setinin farklı oranlarda rassal olarak eksiltilmesinden eksik değerlerin tamamlanıp elde edilen veri setlerinin sınıflandırılmasına kadar olan süreçte eksik veri tamamlama yöntemi hariç tüm uygulama her durum için aynı olmuştur. Bu durumda, eksik

verilerin tamamlanmasının ardından yapılan veri setlerinin sınıflandırılmasına dayalı sonuçlar, eksik veri tamamlama yöntemlerinin, sınıflandırılma performanslarına olan etkilerinin ortaya koyulup kıyaslanmasına olanak sağlamıştır. Dolaylı olarak, veri setlerinin sınıflandırılma performansları üzerinden, eksik veri tamamlama yöntemlerinin başarısının kıyaslanabilmesi de mümkün hale gelmiştir.

Çalışmanın sonucunda elde edilmiş en genel sonuç, tüm eksik veri oranları için geçerli olmak üzere eksik veri tamamlama başarısı ve dolayısıyla sınıflandırma performansı bakımından, makine öğrenmesi algoritmalarının temel eksik veri tamamlama yöntemlerine göre anlamlı şekilde daha başarılı olduğudur. Tüm eksiklik oranlarında ve tüm performans ölçütlerinin aldığı değerlerde makine öğrenmesi algoritmalarının daha başarılı olduğu görülmüştür.

Performans ölçütleri incelendiğinde, söylenildiği gibi makine öğrenmesi algoritmalarının lehine değerlerin gerçekleştiği görülmüştür. Sınıflandırma süreleri göz önüne alındığında, genel olarak temel eksik veri tamamlama yöntemlerinin daha kısa sürede sınıflandırma işlemlerini tamamladığı gözlenmektedir. Sınıflandırma süresi, özellikle çok büyük hacimli veri setleri ile çalışıldığında önemli bir parametre olsa da tek başına değerlendirildiğinde farklı bilgiler içeren bir ölçek değeri taşımaktadır. Temel eksik veri tamamlama yöntemlerine ilişkin sınıflandırma süresinin daha kısa olmasının temel nedenlerinden birisi eksik verileri tamamlarken kullanılan yaklaşımın kompleks olmamasıdır. Örneğin, Liste Boyunca Silme yönteminde eksik değerleri bulunan kayıların direkt silinmesi veri setindeki kayıt sayısını azaltmaktadır bu da sınıflandırma süresini azaltıcı bir etki yaratmaktadır. Bunun dışında, Son Gözlemi İleri Taşıma yönteminde eksik değerden önce gözlenmiş olan son değerin eksik değer yerine atanması söz konusudur. Bu durum da, en azından ilgili nitelik/değişken için tekrar eden bir gözlem ifade etmektedir. Dolayısıyla, tekrar eden gözlemler için sınıflandırma kriterleri, sınıflandırma süreci içinde değerlendirilirken daha önce tekrar eden bu gözlemle işlem yapıldığından direkt olarak bir sonraki nitelik/değişken için sınıflandırma kriterleri sınanacaktır. Böylece, işlem hacmi azalacağından sınıflandırma süresi de kısalacaktır. Bu mantık, Ortalama İle Tamamlamada yöntemi için de düşünülebilir. İfade edilen bu sebeplerden kaynaklı, çalışma kapsamında kullanılan temel eksik veri tamamlama yöntemlerince oluşturulmuş veri setlerinin sınıflandırılması için harcanan süre daha kısa olmaktadır.

Makine öğrenmesi algoritmaları ile eksik verilerin tamamlanmasıyla oluşturulmuş veri setlerinde sınıflandırma süreleri incelendiğinde, genel olarak temel eksik veri tamamlama yöntemleriyle oluşturulmuş veri setlerinin sınıflandırılması için harcanan süreden daha uzun olduğu görülmektedir. Bu durum, daha önce ifade edildiği gibi eksik olan değerlerin tamamlanma aşamasında uygulanan yaklaşımın daha kompleks olması sonucu ortaya çıkan değerlerden kaynaklanmaktadır. Her eksik değer, tamamlanırken ele alınan makine öğrenmesi algoritmasının yaklaşımıyla farklı şekilde ele gözlendiğinden, ilgili veri setinin sınıflandırılması da nispeten daha uzun zaman almaktadır. Yine de, sınıflandırma sürelerine bakıldığında aradaki süre farkının oldukça az olduğu hatta %10 eksik değer oranı durumunda bir algoritmayaya ilişkin süre dışında gerçekleşen sürelerin aynı olduğu ve sınıflandırma için saniyenin %1'inden daha az süre harcandığı gözlenmektedir.

Diğer bir performans ölçütü olan doğruluk ölçütüne gelindiğinde, Makine öğrenmesi algoritmalarına ilişkin sonuçların, temel eksik veri tamamlama yöntemlerine ait sonuçlar ile karşılaştırılmasında net bir şekilde görülen doğruluk oranlarında makine öğrenmesi algoritmalarının açık ara ağırlığını koyduğu ortadadır. Özellikle En yakın k-Komşu algoritma ve Rassal Orman algoritmarının öne çıktığı özet tablolarda gözlenmektedir. %5'lik eksiklik oranı için ortaya çıkan çarpıcı bir sonuç başta doğruluk oranı olmak üzere diğer ölçütlerde de görülmüştür. Hitters veri setinin eksiltilmeden orijinal halinin Naive Bayes algoritması ile sınıflandırılması sonucu elde edilmiş performans ölçütü değerlerine bakıldığında, makine öğrenmesi algoritmalarından En Yakın k-Komşu algoritmasının %5'lik eksiklik oranına ilişkin performans değerlerinin, veri setinin orijinal haline ait değerlerden daha iyi olduğu görülmüştür. Böylece, veri setinin orijinal haline ait performansın üstüne çıkarak, ilgili algoritmanın, sınıflandırma performansını da yukarıya çektiği ve güçlendirdiği ortaya çıkmıştır. Diğer makine öğrenmesi algoritmalarında da %5'lik eksiklik oranına sahip veri setinin tamamlanmasına dayalı sınıflandırma performansına ait doğruluk oranlarının orijinal veri setine ait değerlere oldukça yakın olduğu gözlenmektedir. Bu durum, makine öğrenmesi algoritmalarının %5 eksiklik oranı için oldukça üstün bir performans gösterdiği sonucunu doğurmaktadır. Diğer taraftan, Ortalama ile Tamamlama yönteminin diğer temel eksik veri tamamlama yöntemlerininden nispeten az da olsa iyi anlamda bir fark ortaya koymuş olduğu görülse de makine öğrenmesi

algoritmalarına ait sonuçlar ile kıyaslandığında oldukça düşük bir performans seviyesi çizdiği ortaya çıkmıştır.

Makine öğrenmesi algoritmaları ile temel eksik veri tamamlama yöntemlerine ilişkin doğruluk oranı değerlerindeki, makine öğrenmesi algoritmaları lehine olan sonuçların, %10 ve %15 eksiklik oranlarına ilişkin performans değerlerinde aynı yönde devam ettiği gözlenmiştir. Bu noktada bir farklılık oluştuğu da gözlerden kaçmamalıdır. Bu farklılık, eksiklik oranı arttıkça algoritmaların dolaylı olarak oluşturduğu doğruluk oranlarının başlangıca göre nispeten daha düşük seviyede ilerlediğidir.

Doğruluk oranı kriteri üzerinden ifade edilen sonuçlara dair yorumlamaların, duyarlılık, kesinlik, duyarlılık ve kesinlik kriterlerinin harmonik ortalaması olan F-ölçütü ve ROC alanı kriterleri için de söylenebieceği ortadadır. Ortaya çıkan sonuçlar yine makine öğrenmesi algoritmaları lehinedir. Aynı şekilde, eksiklik oranı arttıkça ilgili performans değerlerinin azaldığı görülmektedir. Fakat temel eksik veri tamamlama yöntemlerine ilişkin sonuçların, eksik veri oranı arttıkça dramatik bir şekilde düşerek açık bir şekilde, kötü performans olarak nitelendirilebilecek duruma geldiği de gözlenmektedir. Temel eksik veri tamamlama yöntemleri arasından Liste Boyunca Silme Yönteminin en kötü performansa sahip olduğu ve eksiklik oranının artmasına paralel olarak ilgili yönteme ilişkin performans değerlerinde ciddi seviyede düşme gözlenmiştir. Bu noktada, makine öğrenmesi algoritmalarına ilişkin sonuç değerleri de aynı şekilde düşmüştür. Fakat makine öğrenmesi algoritmalarına ilişkin performans deperlerindeki düşüşün farkı, temel eksik veri tamamlama yöntemleri gibi dramatik bir şekilde gerçekleşmemiş olmasıdır.

Genel sonuç olarak direkt olarak çıkarılabilecek sonuçlardan biri de veri setinin eksik oranı arttıkça tüm yöntemlerinden kaynaklı elde edilen performans değerlerinin düşmesidir. Bahsedilen durum, oldukça doğal bir sonucun doğurduğu yeni bir sonuçtur. Eksiklik oranının artması veri setinin orijinal halinden daha da uzaklaşılması anlamına gelmektedir. Bu sebeple, eksiklik oranının her arttığında, veri setinin orijinal haline ait karakteristiğinin ve istatistiksel özelliklerin yansıtılması güçleşmektedir. Bu durum, tıpkı bir kumaş parçasında, kumaş bir bütün halinde sağlamken her aşamada farklı kısımlarından delik ve yırtıklar oluşturmak gibi bir duruma benzetilebilecek bir senaryo ifade etmektedir. Çünkü kumaş parçasının sağlam olması durumunda iken

kumaşın özellikleri ile ilgili bilgi ve çıkarım sağlamak nispeten kolaydır ve şaşırtıcı olmayan bir sonuçtur. Böyle bir sonuç, kumaştaki delik ve yırtıkların göz ardı edilebilecek bir seviyede olması durumunda da ortaya çıkabilmektedir. Fakat kumaştaki delik ve yırtıklar çoğaldıkça, kumaşa dair özellikler ile ilgili bilgi almak kolay olmamaktadır. Böyle bir durumda, kumaştaki problemlerin yani delik ve yırtıkların kumaşın özelliklerine göre giderilmesi de haliyle zorlaşmaktadır. Bu durumda da, eksikliklerin ve problemlerin giderilmesine dair ortaya koyulan performansta eksiklik ve problem oranını büyüklüğüne ve artmasına göre düşüşlerin görülmesi bu nedenlerle oldukça doğaldır. Bu çalışmada gerçekleştirilen uygulamada ortaya çıkan, eksiklik oranını arttıkça performans değerlerinin düşmesinin nedenleri bu şekilde ifade edilebilir.

Özetle, çalışma kapsamında yapılan Hitters veri setinin %5, %10 ve %15 eksiklik oranlarında rassal olarak eksiltilerek temel eksik veri tamamlama yöntemlerinden Liste Boyunca Silme, Son Gözlemi İleri Taşıma ve Ortalama ile Tamamlama ve makine öğrenmesi algortimalarından, En Yakın k- Komşu algoritması, Rassal Orman, Stokastik Regresyon ve Amelia algoritmalarıyla veri setinde eksiltilen veriler tamamlanmıştır. Hem veri setinin eksiltilmemiş orijinal hali hem de eksiltilip bahsedilen yöntemlerle tamamlanan böylece yeni bir veri seti olarak karşımıza çıkan veri setleri de Naive Bayes algoritmasıyla sınıflandırılmıştır. Bu sınıflandırma işlemine ait sonuçlar, performans değerlendirme ölçütlerinin aldığı değerler nezdinde hem yöntemlerin kendi içerisindeki kıyaslama hem de veri setinin orijinal haline ait sınıflandırma sonuçları ile kıyaslamada kullanılmıştır. Tez çalışması çerçevesinde gerçekleştirilen uygulama ve bahsedilen kıyaslamalar sonucunda, farklı rassal eksiklik oranlarında, veri setindeki eksik değerlerin, makine öğrenmesi algoritmaları ile tamamlanmasıyla ortaya çıkan veri setlerine ait sınıflandırma performansı değerlerinin, eksik değerlerin temel eksik veri tamamlama yöntemleri ile tamamlanarak elde edilmiş veri setlerine ait sınıflandırma performansı karşısında oldukça büyük bir farklı daha iyi olduğu neticesi elde edilmiştir. Uygulama süreci içerisindeki ait uygulamalar da dahil olmak üzere tek fark eksik değerlerin hangi yöntem veya algoritmayla tamamlandığıdır. Böylece, sınıflandırma performansı değerlerindeki farkı oluşturan nedenin eksik değer tamamlama yöntemi veya algoritması olması durumu ortaya çıkmaktadır. Bu detay da, çalışma kapsamında

kullanılan tüm yöntem ve algoritmaların hem kendi içerisinde bir kıyaslamaya hem de sınıflandırma operasyonu olan etkilerin ortaya çıkmasını sağlamıştır.

Sonuçların makine öğrenmesi algoritmaları lehine olmasının yanı sıra eksik oranı arttıkça tüm yöntem ve algoritmaların performans değerlerinin düştüğü neticesi de gözlenmiştir. Bu durumun, doğal bir sonuç olduğu ve bunun nedenleri de ayrıca belirtilmiştir. Doğal bir performans düşüşü tüm yöntem ve algoritmalarda gözlense de makine öğrenmesi algoritmalarının mevcut iyi performanslarından çok uzaklamadan her şeye rağmen iyi performans sayılabilecek seviyelerde seyrettiği de görülmüştür. Diğer taraftan, eksiklik oranı arttıkça temel eksik veri tamamlama yöntemleriyle tamamlanarak elde edilen veri setlerinin sınıflandırılmasına dayalı performans değerlerinde aynı şekilde düşüş gözlenmiştir. Fakat bu yöntemlere ilişkin perofrmans değerlerinde düşüklük oldukça sert ve dramatik bir seviyede gerçekleşmiştir. Eksiklik oranı arttıkça düşüş oranları da artarak azalan performans değerleri gözlenmiştir. Temel eksik veri tamamlama yöntemlerine ait bu sonuçlara bakıldığında %15'lik eksik veri oranı için yapılan sınfılandırma uygulamasına ait performans değerlerinin oldukça kötü bir performansı ortaya koyduğu saptanmıştır.

Eksik veri tamamlama konusu, pratik dünyada çalışılan veri setlerinde eksiklik bulunma durumunun genel olarak yaşanılabilecek seviyede olmasından dolayı literatürde geniş bir yer bulmaktadır. Fakat yapılan çalışmaların yeteri kadar zengin bir içeriğe sahip olmadığı ve bu alanda daha çok çalışmanın hayata geçirilmesi gerektiği, bu konu da literatüre büyük katkıda bulunan bilim insanları tarafından yapılan akademik çalışmalarda ifade edilmiştir. Nitelikli plan ve özellikle uzun bir periyodu gerektirse de yeterli miktarda zaman içerisinde hem litetatüre bir katkı hem de gündelik veya endüstriyel hayatta ortaya çıkan problemlerin çözümü için gerçekleştirilecek olan analizlere bilimsel bir destek olması adına aşağıda ifade edilmiş olan durumlarda eksik veri ve eksik veri tamamlama çerçevesinde çalışma önerileri sıralanmıştır. Bu öneriler, uzun vadede akademik bir çerçevede değerlendirilerek hayata geçirilecektir.

- Farklı veri türlerini içeren veri setlerinde bulundan eksik değerlerin tamamlanması
- Eksik veri barından barından veri setlerinde, yapay zeka destekli hibrit yaklaşımlar ortaya koyarak eksik verilerin tamamlanması

- Eksik verilerin tamamlanmasında, mesafe tabanlı makine öğrenmesi algoritmalarında eksik veri tamamlama ve verilerin analizine yönelik uygulama performanslarını geliştirecek şekilde uzaklık ölçütlerinin ele alınması
- Bulanık mantık yaklaşımının, niteliklerin veya kriterlerin ağırlıklandırma, olasılık tabanlı nitelik/değişken seçme ve mesafe tabalı yöntemlere entegre edilerek eksik verilerin tamamlanmasına yönelik entegre bir yaklaşımın ortaya koyulması
- Veri setinde yer alan verilerin homojen ve homojen olmayan veri yapıları şeklinde ele alınıp veri setindeki eksik değerlerin bu çerçevede değerlendirilerek tamamlanması
- Veri setindeki eksik değerlerin oranı dışında, veri setindeki lokasyonlarına göre ele alınması. Bu çerçevede, veri setinde volatilitenin yükseltiği kısımlarda eksik verilerin varlığı ve bu verilerin tamamlanması
- Eksik veri tamamlama konseptinin sistematik bir yaklaşımla ele alınması. Belli durumlarda ortaya çıkan eksik verilerin tamamlanması veya bazı durumlarda ortaya çıkan eksik değerlerin tamamlanmasının gerekli olup olmadığının değerlendirilmesine yönelik sistematik bir yaklaşımın geliştirilmesi
- Zaman serisi verilerinde ortaya çıkabilen mevsimsellik ve trend gibi durumlarda, bu durumların hakim olduğu veya etkilediği veri kısımlarında ortaya çıkmış olan eksik veriler için alternatif modeller üzerinden öngörümlemeye dayalı eksik veri tamamlama

KAYNAKÇA

Abidin, N. Z., Ismail, A. R., ve Emran, N. A. (2018). Performance Analysis of Machine Learning Algorithms for Missing Value Imputation. *International Journal of Advanced Computer Science and Applications*. 9(6): 442-447.

Acock, A. C. (2005). Working With Missing Values. *Journal of Marriage and Family*. 67(4): 1012-1028.

Aggarwal, C. C. (2015). *Data Mining: The Textbook.* New York: Springer.

Aittokallio, T. (2010). Dealing With Missing Values In Large-Scale Studies: Microarray Data Imputation And Beyond. *Briefings In Bioinformatics*. 11(2): 253-264.

Akar, Ö. ve Güngör, O. (2012). Classification Of Multispectral Images Using Random Forest algorithm. *Journal of Geodesy and Geoinformation*. 1(2): 105-112.

Aksu, G. & Dogan, N. (2019). Veri Madenciliğinde Kullanılan Güncel Bir Analiz Programı: WEKA. *Journal of Measurement and Evaluation in Education and Psychology*. 10(1): 80-95.

Almuniri, I. ve Said, A. M. (2017). School's Performance Evaluation Based On Data Mining. International Journal of Engineering and Information Systems. 1(9): 56-62.

Alpaydin, E. (2020). *Introduction to Machine Learning.* Cambridge: MIT press.

Alzubi, J., Nayyar, A., ve Kumar, A. (2018). Machine Learning From Theory To Algorithms: An Overview. *Journal of Physics: Conference Series*. 1142, 1-15

Aryanto, F., Fauzi, A., Masruriyah, A. F. N. ve Hananto, A. L. (2023). Sentiment Analysis Of Vaccination Using The K-Nearest Neighbor Algorithm. *Edutran Computer Science and Information Technology*. 1(1): 34-41.

Awan, S. E., Bennamoun, M., Sohel, F., Sanfilippo, F., ve Dwivedi, G. (2022). A Reinforcement Learning-Based Approach For İmputing Missing Data. *Neural Computing and Applications*. 34(12): 9701-9716.

Baraldi, A. N. ve Enders, C. K. (2010). An Introduction To Modern Missing Data Analyses. *Journal Of School Psychology*. 48(1): 5-37.

Baser, B. O., Yangin, M. ve Saridas, E. S. (2021). Makine Öğrenmesi Teknikleriyle Diyabet Hastalığının Sınıflandırılması. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*. 25(1): 112-120.

Bender, A., Schneider, N., Segler, M., Patrick Walters, W., Engkvist, O. ve Rodrigues, T. (2022). Evaluation Guidelines For Machine Learning Tools In The Chemical Sciences. Nature Reviews Chemistry. 6(6): 428-442.

Bertsimas, D., Pawlowski, C. ve Zhuo, Y. D. (2018). From Predictive Methods To Missing Data Imputation: An Optimization Approach. *Journal of Machine Learning Research*. 18(196): 1-39.

Bi, Q., Goodman, K. E., Kaminsky, J., ve Lessler, J. (2019). What Is Machine Learning? A Primer For The Epidemiologist. *American Journal Of Epidemiology*. 188(12): 2222-2239.

Biessmann, F., Rukat, T., Schmidt, P., Naidu, P., Schelter, S., Taptunov, A., ... & Salinas, D. (2019). Datawig: Missing Value Imputation For Tables. *Journal of Machine Learning Research*. 20(175): 1-6.

Bilgin, M. (2018). *Makine Öğrenmesi*. İstanbul: Papatya Yayınları.

Brynjolfsson, E., ve Mitchell, T. (2017). What Can Machine Learning Do? *Workforce Implications. Science*. 358(6370): 1530-1534.

Chen, M., Ebert, D., Hagen, H., Laramee, R. S., Van Liere, R., Ma, K. L., ... & Silver, D. (2008). Data, Information, And Knowledge In Visualization. *IEEE Computer Graphics And Applications*. 29(1): 12-19.

Chhabra, G., Vashisht, V. ve Ranjan, J. (2017). A Comparison Of Multiple Imputation Methods For Data With Missing Values. *Indian Journal of Science and Technology*. 10(19): 1-7.

Dangeti, P. (2017). *Statistics For Machine Learning*. Birmingham: Packt Publishing Ltd.

Datacamp. https://www.datacamp.com/blog/classification-machine-learning, (25.11.2023).

Datamites. https://datamites.com/blog/k-nearest-neighbor-knn-algorithm-in-machine-learning/, (25.11.2023)

Dayan, P., Sahani, M., ve Deback, G. (1999). *The MIT Encyclopedia of The Cognitive Sciences*. Cambridge: MIT Press.

Dayan, P., ve Niv, Y. (2008). Reinforcement Learning: The Good, The Bad and The Ugly. *Current Opinion In Neurobiology*. 18(2): 185-196.

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: Series B (Methodological)*. 39(1): 1-22.

Dogan, C. D. (2017). Applying Bootstrap Resampling To Compute Confidence Intervals For Various Statistics With R. *Eurasian Journal of Educational Research*. 17(68): 1-18.

Dolmans, D. H., De Grave, W., Wolfhagen, I. H. ve Van Der Vleuten, C. P. (2005). Problem-Based Learning: Future Challenges For Educational Practice and Research. *Medical Education*. 39(7): 732-741.

Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T. ve Moons, K. G. (2006). A Gentle Introduction To Imputation Of Missing Values. *Journal of Clinical Epidemiology*. 59(10): 1087-1091.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., ve Tabona, O. (2021). A Survey On Missing Data In Machine Learning. *Journal of Big Data*. 8(1): 1-37.

Enders, C. K. (2022). *Applied Missing Data Analysis*. New York: Guilford Publications.

Erken, Ş., ve Şenyay, L. Makine Öğrenmesi İle Eksik Veri Tamamlama Yöntemlerinin Sınıflandırma Performansına Etkileri. *Kayseri Üniversitesi Sosyal Bilimler Dergisi*. 5(1): 51-71.

Famili, A., Shen, W. M., Weber, R. ve Simoudis, E. (1997). Data Preprocessing And Intelligent Data Analysis. *Intelligent Data Analysis*. 1(1): 3-23.

Farhangfar, A., Kurgan, L. ve Dy, J. (2008). Impact of Imputation of Missing Values On Classification Error For Discrete Data. *Pattern Recognition*. 41(12): 3692-3705.

Friedman, J., Hastie, T., ve Tibshirani, R. (2001). *The Elements of Statistical Learning*. New York: Springer.

Gajawada, S. ve Toshniwal, D. (2012). Missing Value Imputation Method Based On Clustering And Nearest Neighbours. *International Journal of Future Computer and Communication*. 1(2): 206-208.

García, S., Luengo, J., Herrera, F., García, S., Luengo, J. ve Herrera, F. "Dealing With Missing Values", Data Preprocessing In Data Mining, Ed. Janusz Kacprzyk ve Lakhmi C. Jain, Springer International Publishing, New York, 2015, pp.59-105.

Gibson, B. R., Rogers, T. T. ve Zhu, X. (2013). Human Semi-Supervised Learning. *Topics In Cognitive Science*. 5(1): 132-172.

Gollapudi, S. (2016). *Practical Machine Learning*. Birmingham: Packt Publishing Ltd.

Gürlen, E. (2011). Probleme Dayalı Öğrenmenin Öğrenme Ürünlerine, Problem Çözme Becerisine, Öz-Yeterlik Algı Düzeyine Etkisi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*. 40(40): 221-232.

Gürsakal, N. (2019). *Veri Bilimi*. Bursa: Dora Yayıncılık.

Han, J. & Kamber, M. (2006). *Data Mining: Concepts and Techniques.* San Francisco: Morgan Kaufmann.

Honaker, J., King, G. ve Blackwell, M. (2011). Amelia II: A Program For Missing Data. *Journal of Statistical Software*. 45(7): 1–47.

Hu, L. Y., Huang, M. W., Ke, S. W. ve Tsai, C. F. (2016). The Distance Function Effect On K-Nearest Neighbor Classification For Medical Datasets. *SpringerPlus*. *5*(1): 1-9.

İnan, S. (2019). *Normal, Multinomial, Üssel (Exponential) ve Gamma Dağılım Gösteren Veri Yapıları Ve Eksik Veri Tiplerinde (MAR, MCAR, MNAR) Tamamlama Algoritmalarının Parametre Tahminleri Üzerine Etkileri*. (Yayınlanmamış Yüksek Lisans Tezi). Van: Yüzüncü Yıl Üniversitesi Fen Bilimleri Enstitüsü.

Jadhav, A., Pramod, D. ve Ramanathan, K. (2019). Comparison of Performance Of Data Imputation Methods For Numeric Dataset. *Applied Artificial Intelligence*. 33(10): 913-933.

James, G., Witten, D., Hastie, T., Tibshirani, R. ve Taylor, J. (2023). *An Introduction to Statistical Learning: with Applications in Python*. Cham: Springer International Publishing.

James, G., Witten, D., Hastie, T., ve Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York: Springer.

Javatpoint. https://www.javatpoint.com/supervised-machine-learning, (25.11.2023)

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M. ve Franco, L. (2010). Missing Data Imputation Using Statistical and Machine Learning Methods In A Real Breast Cancer Problem. *Artificial Intelligence In Medicine*. 50(2): 105-115.

Kabir, G., Tesfamariam, S., Hemsing, J. ve Sadiq, R. (2020). Handling Incomplete And Missing Data In Water Network Database Using Imputation Methods. *Sustainable and Resilient Infrastructure*. 5(6): 365-377.

Kaggle. https://www.kaggle.com/datasets/floser/hitters, (10.01.2023)

Kenward, M. G. ve Molenberghs, G. (2009). Last Observation Carried Forward: A Crystal Ball?. *Journal of Biopharmaceutical Statistics*. 19(5): 872-888.

Kenyhercz, M. W., ve Passalacqua, N. V. (2016). Missing Data Imputation Methods and Their Performance With Biodistance Analyses. *Biological Distance Analysis*. London: Academic Press.

Kulkarni, V. Y. ve Sinha, P. K. (2013). Random Forest Classifiers: A Survey And Future Research Directions. *Int. J. Adv. Comput*. 36(1): 1144-1153.

Kwak, S. K. ve Kim, J. H. (2017). Statistical Data Preparation: Management Of Missing Values And Outliers. *Korean Journal of Anesthesiology*. 70(4): 407-411.

Li, Y. F. ve Zhou, Z. H. (2014). Towards Making Unlabeled Data Never Hurt. *IEEE Transactions On Pattern Analysis and Machine Intelligence*. 37(1): 175-188.

Liaw, A., ve Wiener, M. (2002). Classification and Regression by RandomForest. *R News*. 2(3): 18-22.

Liew, A. (2007). Understanding Data, Information, Knowledge And Their Inter-Relationships. *Journal of Knowledge Management Practice*. 8(2): 1-16.

Little, R. J. ve Rubin, D. B. (1983). On Jointly Estimating Parameters And Missing Data By Maximizing The Complete-Data Likelihood. *American Statistician*, 37(3): 218-220.

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

Mahesh, B. (2020). Machine Learning Algorithms-A Review. *International Journal of Science and Research (IJSR)*. 9(1): 381-386.

Medium. https://towardsdatascience.com/bootstrapping-statistics-what-it-is-and-why-its-used-e2fa29577307, (25.11.2023)

Miao, J. ve Zhu, W. (2022). Precision–Recall Curve (PRC) Classification Trees. Evolutionary Intelligence. 15(3): 1545-1569.

Mohri, M., Rostamizadeh, A. ve Talwalkar, A. (2018). *Foundations of Machine Learning*. Cambridge: MIT Press.

Morimoto, J. ve Ponton, F. (2021). Virtual Reality In Biology: Could We Become Virtual Naturalists?. *Evolution: Education and Outreach*. 14(1): 7.

Mulak, P. ve Talhar, N. (2015). Analysis Of Distance Measures Using K-Nearest Neighbor Algorithm On KDD Dataset. *Int. J. Sci. Res*. 4(7): 2319-7064.

Myers, T. A. (2011). Goodbye, Listwise Deletion: Presenting Hot Deck Imputation As An Easy And Effective Tool For Handling Missing Data. *Communication Methods and Measures*. 5(4): 297-310.

Newman, D. A. (2014). Missing Data: Five Practical Guidelines. *Organizational Research Methods*. 17(4): 372-411.

OACON. https://www.oacon.com.tr/Dogruluk-Ve-Tekrarlanabilirlik, (20.11.2023)

Orchard, T. ve Woodbury, M. A. (1972). A Missing Information Principle: Theory And Applications. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics* (pp. 697-716): Düzenleyen: University of California. Berkeley. 21 Haziran- 19 Temmuz 1970.

Özkan, Y. (2008). *Veri Madenciliği Yöntemleri.* İstanbul: Papatya Yayıncılık Eğitim.

Öztemel, E. (2003). *Yapay Sinir Ağları.* İstanbul: Papatya Yayınları.

Patil, T. R., ve Sherekar, S. S. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. *International Journal of Computer Science and Applications*. 6(2): 256-261.

Patrician, P. A. (2002). Multiple Imputation For Missing Data. *Research In Nursing & Health*. 25(1): 76-84.

Peng, J., Jury, E. C., Dönnes, P. ve Ciurtin, C. (2021). Machine Learning Techniques For Personalised Medicine Approaches In Immune-Mediated Chronic Inflammatory Diseases: Applications And Challenges. *Frontiers In Pharmacology*. 12, 1-18.

PennState World Campus MIS204. https://courses.worldcampus.psu.edu/welcome/mis204/001/content/01_lesson/03_page.html, (25.11.2023)

Poulos, J. ve Valle, R. (2018). Missing Data Imputation For Supervised Learning. *Applied Artificial Intelligence*. 32(2): 186-196.

Pyle, D. (1999). *Data Preparation For Data Mining*. San Francisco: Morgan Kaufmann.

Raja, P. S. ve Thangavel, K. J. S. C. (2020). Missing Value Imputation Using Unsupervised Machine Learning Techniques. *Soft Computing*. 24(6): 4361-4392.

Rao, R. V. ve Selvamani, K. (2015). Data Security Challenges And Its Solutions In Cloud Computing. *Procedia Computer Science*. 48, 204-209.

Rässler, S., Rubin, D. B. ve Zell, E. R. (2013). Imputation. *Wiley Interdisciplinary Reviews: Computational Statistics*. 5(1): 20-29.

Ratnasari, D. (2023). Comparison of Performance of Four Distance Metric Algorithms in K-Nearest Neighbor Method on Diabetes Patient Data. *Indonesian Journal of Data and Science*. 4(2): 97-108.

Rawal, S., Gupta, S. C. ve Singh, S. (2017). Predicting Missing Values In A Dataset: Challenges and Approaches. *International Journal of Recent Research Aspects*. 4(3): 34-38.

Rebala, G., Ravi, A., ve Churiwala, S. (2019). *An Introduction To Machine Learning*. Berlin: Springer.

Rigatti, S. J. (2017). Random Forest. *Journal of Insurance Medicine*. 47(1): 31-39.

Rodríguez, R., Pastorini, M., Etcheverry, L., Chreties, C., Fossati, M., Castro, A. ve Gorgoglione, A. (2021). Water-Quality Data Imputation With A High Percentage Of Missing Values: A Machine Learning Approach. *Sustainability*. 13(11): 6318.

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

Samuel, A. L. (1959). Some Studies In Machine Learning Using The Game Of Checkers. *IBM Journal of Research and Development*. 3(3): 210-229.

Saritas, M. M. ve Yasar, A. (2019). Performance Analysis Of ANN And Naive Bayes Classification Algorithm For Data Classification. *International Journal Of Intelligent Systems And Applications In Engineering*. 7(2): 88-91.

Serokell. https://serokell.io/blog/random-forest-classification, (25.11.2023)

Silahtaroğlu, G. (2016). *Veri Madenciliği*. İstanbul: Papatya Yayınları.

Silva-Ramírez, E. L., Pino-Mejías, R., López-Coello, M. ve Cubiles-de-la-Vega, M. D. (2011). Missing Value Imputation On Missing Completely At Random Data Using Multilayer Perceptrons. *Neural Networks*. 24(1): 121-129.

Sokolova, M. ve Lapalme, G. (2009). A Systematic Analysis Of Performance Measures For Classification Tasks. Information Processing & Management. 45(4): 427-437.

Sumathi, S. ve Sivanandam, S.N. (2006). *Introduction to Data Mining and Its Applications*. Berlin: Springer.

Sutton, R. S., ve Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. Cambrige, MA: MIT Press.

Tang, F. ve Ishwaran, H. (2017). Random Forest Missing Data Algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*. 10(6): 363-377.

TechVidvan. https://techvidvan.com/tutorials/reinforcement-learning/, (25.11.2023)

Thomas, T., ve Rajabi, E. (2021). A Systematic Review Of Machine Learning-Based Missing Value Imputation Techniques. *Data Technologies and Applications*. 55(4): 558-585.

Tsai, C. F. ve Hu, Y. H. (2022). Empirical Comparison Of Supervised Learning Techniques For Missing Value Imputation. *Knowledge and Information Systems*. 64(4): 1047-1075.

UCI. https://archive.ics.uci.edu/, (15.03.2023)

Van Buuren, S. (2018). Flexible Imputation Of Missing Data. Florida: CRC Press.

Van der Heijden, G. J., Donders, A. R. T., Stijnen, T. ve Moons, K. G. (2006). Imputation Of Missing Values Is Superior To Complete Case Analysis And The Missing-İndicator Method In Multivariable Diagnostic Research: A Clinical Example. *Journal of Clinical Epidemiology*. 59(10): 1102-1109.

Van Engelen, J. E., ve Hoos, H. H. (2020). A Survey On Semi-Supervised Learning. *Machine learning*. 109(2): 373-440.

Vembandasamy, K., Sasipriya, R. ve Deepa, E. (2015). Heart Diseases Detection Using Naive Bayes Algorithm. *International Journal of Innovative Science, Engineering & Technology*. 2(9): 441-444.

Vercellis, C. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making.* Chichester: Wiley.

Wikipedia. https://en.wikipedia.org/wiki/Receiver_operating_characteristic, (20.11.2023)

Witten, I. H., Frank, E. ve Hall, M. A. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. USA: Morgan Kaufmann Publishers.

Yager, R. R. (2006). An Extension Of The Naive Bayesian Classifier. *Information Sciences*. 176(5): 577-588.

Yilmaz, M. (2009). Enformasyon Ve Bilgi Kavramları Bağlamında Enformasyon Yönetimi Ve Bilgi Yönetimi. *Ankara Üniversitesi Dil ve Tarih-Coğrafya Fakültesi Dergisi*. 49(1): 95-118.

Zebari, S.C.H (2023). *A Machine Learnıng Classification Approach for Diabetes and Biomedıcal Data*. (Yayınlanmamış Yüksek Lisans Tezi). Şanlıurfa: Harran Üniversitesi Fen Bilimleri Enstitüsü

Zhang, S. (2012). Nearest Neighbor Selection For Iteratively Knn Imputation. *Journal of Systems and Software*. 85(11): 2541-2552.

Zhang, Z. (2016). Multiple Imputation For Time Series Data With Amelia Package. *Annals of Translational Medicine*. 4(3): 1-10.

Zhu, C. ve Gao, D. (2016). Influence Of Data Preprocessing. *Journal of Computing Science and Engineering*. 10(2): 51-57.

Zins, C. (2007). Conceptual Approaches For Defining Data, İnformation, And Knowledge. *Journal Of The American Society For İnformation Science And Technology*. 58(4): 479-493.

Zoubir, A. M. ve Iskandler, D. R. (2007). Bootstrap Methods And Applications. *IEEE Signal Processing Magazine*. 24(4): 10-19.