İZMİR BAKIRÇAY ÜNİVERSİTESİ

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
AKILLI SİSTEMLER MÜHENDİSLİĞİ A.B.D.

KALP HASTALIĞI TANISINDA WEKA TABANLI MAKİNE ÖĞRENMESİ ALGORİTMALARININ PERFORMANS ANALİZİ

YÜKSEK LİSANS TEZİ

Bekir Can TELKENAROĞLU

Tez Danışmanı: Doç. Dr. Bahar DEMİRTÜRK

Temmuz 2024

KALP HASTALIĞI TANISINDA WEKA TABANLI MAKİNE ÖĞRENMESİ ALGORİTMALARININ PERFORMANS ANALİZİ

Yüksek Lisans Tezi

Bekir Can TELKENAROĞLU İzmir 2024

KALP HASTALIĞI TANISINDA WEKA TABANLI MAKİNE ÖĞRENMESİ ALGORİTMALARININ PERFORMANS ANALİZİ

Bekir Can TELKENAROĞLU

YÜKSEK LİSANS TEZİ

Akıllı Sistemler Mühendisliği

Danışman: Doç. Dr. Bahar DEMİRTÜRK

İzmir

İzmir Bakırçay Üniversitesi

Lisansüstü Eğitim Enstitüsü

Temmuz 2024

JÜRİ VE ENSTİTÜ ONAYI

İzmir Bakırçay Üniversitesi Lisansüstü Eğitim Enstitüsü Akıllı Sistemler Mühendisliği Anabilim dalında öğrenim görmekte olan Bekir Can Telkenaroğlu' nun "İzmir Bakırçay Üniversitesi'nin Kalp Hastalığı Tanısında Weka Tabanlı Makine Öğrenmesi Algoritmalarının Performans Analizi" başlıklı tezi 16.07.2024 tarihinde aşağıdaki jüri tarafından değerlendirilerek "İzmir Bakırçay Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği" nin ilgili maddeleri uyarınca, Akıllı Sistemler Mühendisliği Anabilim dalında Yüksek Lisans tezi olarak kabul edilmiştir.

Jüri Üyeleri	**Ünvanı Adı Soyadı**	**İmza**
Üye (Tez Danışmanı)	Doç. Dr. Bahar DEMİRTÜRK	
Üye	Doç. Dr. Volkan KILIÇ	
Üye	Dr. Öğr. Üyesi Bayram KÖSE	

Prof. Dr. Özge TÜZÜN ÖZMEN
Lisansüstü Eğitim Enstitüsü Müdürü

FINAL APPROVAL FOR THESIS

The thesis titled "Performance Analysis of Weka-Based Machine Learning Algorithms in Heart Disease Diagnosis of İzmir Bakırçay University" by Bekir Can Telkenaroğlu, who is studying in the Department of Intelligent Systems Engineering at İzmir Bakırçay University Graduate Education Institute, was evaluated by the following jury on 16.07.2024 and "İzmir Bakırçay University In accordance with the relevant articles of the "Postgraduate Education and Examination Regulations", it has been accepted as a Master's thesis in the Department of Intelligent Systems Engineering.

Committee Members	**Title, Name and Surname**	**Signature**
Member (Supervisor)	Assoc. Prof. Dr. Bahar DEMİRTÜRK	
Member	Assoc. Prof. Dr. Volkan KILIÇ	
Member	Assist. Prof. Dr. Bayram KÖSE	

Prof. Dr. Özge TÜZÜN ÖZMEN
Director of Graduate Education Institute

ÖZET

KALP HASTALIĞI TANISINDA WEKA TABANLI MAKİNE ÖĞRENMESİ ALGORİTMALARININ PERFORMANS ANALİZİ

Bekir Can TELKENAROĞLU

Akıllı Sistemler Mühendisliği Anabilim Dalı

İzmir Bakırçay Üniversitesi, Lisansüstü Eğitim Enstitüsü, Temmuz 2024

Danışman: Doç. Dr. Bahar DEMİRTÜRK

Kalp hastalığı, kalbin normal işlevlerini yerine getiremediği ve genellikle kardiyovasküler sistemdeki sorunlarla ilişkilendirilen bir durumdur. Erken teşhis, tedavi ve önlemler açısından hayati öneme sahiptir. Makine öğrenmesi algoritmaları, bilgisayar sistemlerinin verilerden öğrenme yeteneği kazanmasını sağlayan matematiksel modellerdir. Sınıflandırma, regresyon, kümeleme gibi farklı görevler için kullanılan bu algoritmalar, veri analizi ve örüntü tanıma gibi birçok alanda kullanılır. Bu çalışma Weka ile makine öğrenmesi algoritmalarının kalp hastalıklarını teşhis etme yeteneğini incelemek ve karşılaştırmak amacıyla yapılmıştır. Weka, açık kaynaklı bir veri madenciliği ve makine öğrenmesi algoritmalarının kullanıldığı bir platformdur. Weka, araştırmacılar tarafından geniş bir alanı kapsayan birçok projede tercih edilmektedir. Veri madenciliği yöntemleri ile veri seti analiz edilerek regresyon, sınıflandırma ve kümeleme algoritmaları kullanılarak bu çalışma gerçekleştirilmiştir. Kullanılan algoritmaların performansını değerlendirmek için kullanılan parametrelerle, sonuçlar kapsamlı bir şekilde analiz edilmiştir. Algoritmaların performansları incelendiğinde bulgular, Weka ile uygulanan çeşitli makine öğrenmesi algoritmalarının kalp hastalığı teşhisinde başarı sağladığını göstermektedir. Bu çalışma, sağlık profesyonelleri ve araştırmacıları için kalp hastalığı teşhisinde makine öğrenmesi uygulamalarını daha iyi anlamalarına ve potansiyel olarak hastalıkların teşhis süreçlerinin geliştirilmelerine yardımcı olacaktır.

Anahtar Sözcükler: Kalp hastalığı teşhisi; Weka; Makine Öğrenmesi Algoritmaları; Veri analizi; Performans Analizi

ABSTRACT

PERFORMANCE ANALYSIS OF WEKA-BASED MACHINE LEARNING ALGORITHMS IN HEART DISEASE DIAGNOSIS

Bekir Can TELKENAROGLU

Department of Intelligent Systems Engineering

İzmir Bakırçay University, Graduate Education Institute, July 2024

Supervisor: Assoc. Prof. Dr. Bahar DEMİRTÜRK

Heart disease is a condition in which the heart cannot perform its normal functions and is often associated with problems in the cardiovascular system. Early diagnosis is vital for treatment and precautions. Weka is a software language program that uses an open source data mining and machine learning algorithm. Weka is preferred by researchers in many projects covering a wide area. Machine learning algorithms are mathematical models that enable computer systems to gain the ability to learn from data. These algorithms, which are used for different tasks such as classification, regression and clustering, are used in many areas such as data analysis and pattern recognition. This study was conducted to examine and compare the ability of Weka and machine learning algorithms to diagnose heart diseases. This study was carried out by analyzing the data set with data mining methods and using regression, classification and clustering algorithms. The results have been extensively analyzed with the parameters used to evaluate the performance of the algorithms used. When the performances of the algorithms are examined, the findings show that various machine learning algorithms implemented with Weka are successful in diagnosing heart disease. This study will help healthcare professionals and researchers better understand the applications of machine learning in diagnosing heart disease and potentially improve patients' diagnostic processes.

Keywords: Heart disease diagnosis; Weka; Machine Learning Algorithms; Data Analysis; Performance Analysis

TEŞEKKÜR

Tez çalışması sürecinde yol gösterici olan, hiçbir konuda desteğini esirgemeyen, zorlandığım noktalarda bana verdiği motivasyon ile tekrar çalışmaya odaklanmamı sağlayan ve kıymetli bilgisiyle desteklerini benden esirgemeyen sayın danışmanım Doç. Dr. Bahar DEMİRTÜRK' e teşekkürlerimi sunarım.

Son zamanların en zorlu yıllarında bana olan inancı ve sonsuz desteği ile çalışma boyunca yanımda olan sevgili eşim Kidonya Eylül TELKENAROĞLU' na, eğitim hayatım boyunca desteğini esirgemeyen, ilkokul yıllarımdan itibaren daha iyi eğitim alabilmem için her türlü fırsatı sunan annem ve babama, hayatımın her alanında elinden geldiğince yanımda olmak için çaba sarf eden biricik ablama teşekkürlerimi sunarım.

16/07/2024

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Bu tezin bana ait, özgün bir çalışma olduğunu; çalışmamın hazırlık, veri toplama, analiz ve bilgilerin sunumu olmak üzere tüm aşamalarında bilimsel etik ilke ve kurallara uygun davrandığımı; bu çalışma kapsamında elde edilen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi; bu çalışmanın İzmir Bakırçay Üniversitesi tarafından kullanılan Turnitin bilimsel intihal tespit programıyla tarandığını ve hiçbir şekilde "intihal içermediğini" beyan ederim. Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçları kabul ettiğimi bildiririm.

Bekir Can TELKENAROĞLU

16/07/2024

STATEMENT OF COMPLIANCE WITH ETHICAL PRINCIPLES&RULES

I hereby truthfully declare that thesis is an original work prepared by me; that I have behaved in accordance with the scientific ethical principles and rules throughout the stages of preparation, data collection, analysis and presentation of my work; that I have cited the sources of all the data and information that could be obtained within the scope of this study, and included these sources in the references section; and that this study has been scanned for plagiarism with Turnitin scientific plagiarism detection program used by İzmir Bakırçay University, and that "it does not have any plagiarism" whatsoever. I also declare that, if a case contrary to my declaration is detected in my work at any time, I hereby Express my consent to all the ethical and legal consequences that are involved.

Bekir Can TELKENAROĞLU

İÇİNDEKİLER

Sayfa

ÇİZELGELER DİZİNİ

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

SİMGELER VE KISALTMALAR DİZİNİ

WHO : Dünya Sağlık Örgütü

UCI : University Of California, Irvine

DT : Decision Tree

NB : Naive Bayes

CHD : Coronary Heart Disease

PSO : Parçacık Sürü Optimizasyonu

ICA : Independent Component Analysis

KDD : Knowledge Discovery in Databases

IJCAI : International Joint Conferences on Artificial Intelligence

WEKA : Waikato Environment for Knowledge Analysis

IBK : Instance-Based k Nearest Neighbors

SMO : Sequantial Minimal Optimization

LibSVM : Library for Support Vector Machines

EM : Expectation-Maximization

KI : Kappa İstatistiği

KK : Korelasyon Katsayısı

OMH : Ortalama Mutlak Hata

HKOK : Hataların Karelerinin Ortalamasının Karekökü

GMH : Göreceli mutlak hata

GMHK : Göreceli Mutlak Hata Karekökü

DP : Doğru Pozitif

YP : Yanlış Pozitif

DN : Doğru Negatif

YN : Yanlış Negatif

F-Ö : F-Ölçüsü

AİK : Alıcı İşlem Karakteristikleri

KHE : Kesinlik Hassasiyet Eğrisi

1. GİRİŞ

Teknolojinin gelişimiyle birlikte işletmelerin iş yapış şekillerinde, tüm sektörlerde köklü bir değişim yaşanmıştır. Bu değişim, sağlık alanında da etkisini göstermiş ve özellikle hastalıkların erken teşhisi konusu büyük önem kazanmıştır. Erken teşhis, bireylerin potansiyel tehlikelerle karşılaşmasını önlerken, yaşamlarını kurtarmanın yanı sıra sağlık harcamalarını da önemli ölçüde azaltabilir. Dünya Sağlık Örgütü(WHO)'nün verilerine göre, dünya genelinde ölüm nedenlerinin başında kalp hastalıkları gelmektedir. Tahminlere göre, 2019 yılında 18.6 milyondan fazla insan kalp hastalıkları nedeniyle hayatını kaybetmiştir. Bu rakam, dünya genelindeki ölümlerin yaklaşık %31'ini oluşturmakta ve kalp hastalıklarının teşhis yöntemlerinin ne kadar kritik olduğunu göstermektedir (http-1, s.69). Bu durum dikkate alındığında kalp hastalığı riskinin azaltılması için önlemler alınması gerektiği ortadadır. Ancak hastalık teşhisi tıp alanında zorlu bir süreçtir. Teşhis genellikle hastanın bulguları, semptomları ve fizik muayenesine dayanır. Çoğu zaman, doktorlar bilgi ve deneyimlerine dayanarak kalp hastalığını öngörebilirler. Ancak, doğru tanı koymak için tek bir insan zekası yeterli olmayabilir. Hastaların verdiği bilgiler birbiriyle ilişkili semptomları içerebilir ve özellikle hastaların şikayetleri çeşitli hastalıkları işaret edebilir. Bu gibi durumlar, doktorların doğru teşhisi koymasını zorlaştırabilir (Topol, 2019, s.69).

Sağlık alanında yapılan bilimsel çalışmalarda, hastalardan toplanan veri kümeleri uzun süredir istatistiksel yöntemler kullanılarak incelenmektedir. Bu süreçte, tıp alanında çalışanların istatistiksel yöntemleri yakından takip ettikleri bilinmektedir. Ancak, veri madenciliği yöntemlerinin sağlık verisi üzerindeki uygulamaları genellikle enformatik ve mühendislik gibi alanlarda uzmanlaşmış araştırmacılar tarafından gerçekleştirilmektedir. Bu durum, sağlık çalışanları ve araştırmacıları ile veri madenciliği arasında çok sıkı bir bağlantı olmadığını göstermektedir. Halbuki veri madenciliği analizlerinin sağlık alanında da büyük potansiyeli bulunmaktadır. Sağlık alanı, hem çok geniş bir veri yelpazesine sahiptir hem de bu veri miktarı sürekli olarak artmaktadır (Holzinger, 2014, s.66).

Veri madenciliği yöntemleri, hastalıkların tespiti, hastalık tahmini ve erken teşhis koyma gibi alanlarda önemli katkılar sağlayabilir. Ayrıca, hastalıkların gruplandırılması, doktorlara destek sağlayacak karar destek sistemlerinin geliştirilmesi ve ender görülen hastalıkların ya da anormalliklerin tespiti gibi alanlarda da kullanılabilir. Bu şekilde, veri

madenciliği yöntemleri, sağlık alanında daha etkili ve verimli bir şekilde kullanılarak hastalık yönetimi ve tedavi süreçlerinin iyileştirilmesine yardımcı olabilir (Chawla ve Davis, 2013, s.65).

Günümüzde, birçok teşhis yöntemi mevcut olmasına rağmen, yapay zeka alanındaki gelişmeler, özellikle makine öğrenmesi algoritmaları sayesinde büyük ilerlemeler kaydetmiştir. Makine öğrenmesi algoritmaları, mevcut veri setleriyle eğitilerek tahminlerde bulunabilme yeteneğine sahiptir. Yüksek doğruluk oranlarına sahip makine öğrenmesi algoritmaları, yeni hastalıkların teşhisinde önemli bir rol oynayabilir ve hastaların tedavi süreçlerine değerli katkılarda bulunabilir. Bu nedenle, makine öğrenmesi algoritmalarının kullanımı, kalp hastalığı gibi ciddi sağlık sorunlarının erken teşhisinde önemli bir araç olarak kabul edilmektedir (Cheng ve Montagnon, 2018, s.65).

Kalp hastalığı, dünya genelinde sağlık sorunları arasından önde gelen ölüm nedenlerinden biridir. Erken teşhis, hastaların yaşam kalitesini artırabilir ve tedaviye başlama sürecini hızlandırabilir. Bu nedenle, kalp hastalığı teşhisi için etkili yöntemlerin geliştirilmesi önemlidir. (Kandemir ve Turhan, 2015, s.67).

Kalp hastalığının teşhisi, hastaların sağlığını iyileştirmek ve sağlık hizmetlerini optimize etmek için kritik bir adımdır. Makine öğrenmesi, sağlık sektöründe veri analizi ve hastalık teşhisi için giderek daha fazla kullanılmaktadır. Weka, ücretsiz ve açık kaynaklı bir veri madenciliği yazılım aracıdır. İçeriğinde pek çok makine öğrenmesi algoritması bulundurmaktadır. Bu çalışma Weka kullanılarak makine öğrenmesi algoritmalarının kalp hastalıklarını teşhis etme yeteneğini incelemeyi, karşılaştırmayı, performans analizini ve sağlık profesyonellerine yeni araçlar sunmayı amaçlamaktadır.

Kalp hastalığı, birçok farklı alt türü içerir, bunlar arasında koroner arter hastalığı, kalp yetmezliği, hipertansiyon ve aritmiler gibi yaygın olanlar bulunmaktadır. Erken teşhis, bu hastalıkların etkilerini hafifletmek ve hastaların yaşam süresini uzatmak açısından kritik öneme sahiptir (Nattel ve Dobrev, 2017, s.68).

Makine öğrenmesi, tıp ve sağlık sektöründe kalp hastalığı teşhisi gibi karmaşık problemleri çözmek için kullanılan bir araç haline gelmiştir. Bu teknikler, büyük veri setlerini analiz etme ve problemleri çözmede etkindir (Esteva, 2017, s.66)

Bu bağlamda, Weka gibi veri madenciliği araçları, araştırmacılara ve sağlık profesyonellerine veri analizi ve makine öğrenmesi algoritmalarını uygulamak için kullanışlı bir platform sunar. Weka'nın açık kaynaklı ve kullanıcı dostu yapısı,

araştırmacıların farklı algoritmaları karşılaştırmasını kolaylaştırır.

1.1. Kalp Hastalıkları

Teknolojinin günlük yaşantımıza entegre olmasıyla birlikte insanların yaşam tarzlarında büyük değişiklikler meydana gelmiştir. Teknolojik gelişmelerin sağladığı avantajlarla birlikte, hareketsizlik stres ve beslenme alışkanlıklarındaki değişim gibi dezavantajlar da sağlık durumumuzu olumsuz etkilemektedir (Aydın ve Uzunboylu, 2019, s.65).

Kalp hastalığı, kalbi ve kan damarlarını etkileyen çeşitli rahatsızlık ve durumları kapsayan genel bir terimdir. Kalbin verimli çalışması sağlıklı bir yaşam için gereklidir. Kalbin düzgün çalışmasında herhangi bir aksaklık, beyin, böbrek veya vücudun diğer bölgeleri üzerinde doğrudan olumsuz etkilere yol açabilir. (Özdemir ve Arslan, 2018, s. 68).

Tahminlere göre 2019 yılında, yaklaşık 18.6 milyon insan kardiyovasküler hastalıklar nedeniyle hayatını kaybetmişken, bu sayı dünya genelindeki tüm ölümlerin %31' ine denk gelmektedir. Özellikle kalp krizi ve felç gibi hastalıkları tetikleyen faktörler; sağlıksız beslenme, sigara içme, fiziksel aktivitenin azalması ve alkol kullanımıdır. Bu faktörler sırasıyla yüksek tansiyon, yüksek kan şekeri, aşırı kilo ve obeziteye yol açarak kalp sağlığını tehdit etmektedir (http-2, s.69).

1.1.1. Kalp Hastalığı Tehlikesini Artıran Etkenler

Kalp hastalığının meydana gelmesinde etken olan muhtelif nedenler vardır. Bunlar içinde yaş, yüksek tansiyon, kolesterol seviyesi, sigara kullanma bağımlılığı, şeker hastalığı (diyabet), obezite, hareketsiz bir ömür ve genetik yatkınlık mühim rol oynar. (Erem, 2018, s.66).

Yaşlanma, kalp kasında incelme ya da kalınlaşma ile arterlerde daralma gibi durumların ortaya çıkmasına neden olabilir, bu da kalp hastalığı riskini artırır. Yaşlanmanın etkisiyle, kalp kasının esnekliğinin azalması ve arterlerde plak birikiminin artması kalp hastalığı riskini artırır. Bu nedenle, yaşlanma sürecinde düzenli tıbbi kontrollerin yapılması ve sağlıklı yaşam şeklinin benimsenip, bu alışkanlıkların sürdürülmesi önemlidir. (Eroğlu, 2019, s.66).

Yüksek tansiyon, kan basıncının normalden yüksek olması durumudur. Yüksek tansiyon, kalp ve damarlar üzerindeki baskıyı artırarak kalp hastalığı tehlikesini artırabilir. Yüksek kolesterol seviyeleri, arterlerde plak birikimine ve arterlerin daralmasına yol açarak kalp hastalığı riskini artırır (Bilgin ve Eren, 2019, s.66). Sigara içmek, damar duvarlarının hasar görmesine ve daralmasına neden olabilir, bu da kalp hastalığı tehlikesini artırır ve kan pıhtılarının oluşumunu teşvik edebilir (Çelik ve Erdoğan, 2020, s.66).

Şeker hastalığı veya diyabet, kan şekerinin yükselmesine neden olabilir. Bu durum, arterlerdeki hasar riskini artırarak kalp hastası olma potansiyelini doğurur. Obezite, vücutta inflamasyonu artırabilir, kan basıncını ve kolesterol seviyelerini etkileyebilir. Böylece kalp hastalığı riski yükselmiş olur (Gregg, Sattar ve Ali, 2016, s. 66).

Hareketsiz bir yaşam tarzı da kalp hastalığı tehlikesini yükselten etmenlerden birisidir. Düzenli egzersiz yapmamanın kilo kontrolünü zorlaştırabileceği, kan basıncını ve kolesterol seviyelerini artırabileceği ve böylece kalp hastalığı riskini artırabileceği bilinmektedir. Ayrıca, ailede kalp hastalığı öyküsü bulunması da bireyin kalp hastalığı tehlikesini artırabilir ve genetik faktörlerin kalp hastalığına yakalanmada rol oynayabileceğini gösterebilir (Taşçılar, 2021, s. 69).

Bu risk faktörleri, kalp hastalığının gelişiminde önemli bir etkiye sahiptir ve sağlıklı yaşam tarzı seçimleri ve düzenli tıbbi kontroller gibi önlemlerle azaltılabilir veya yönetilebilir.

1.1.2. Kalp Hastalığı Çeşitleri

Kalp hastalığı çeşitleri arasında şunlar bulunmaktadır.

Koroner Arter Hastalığı: Bu durum, kalp kasına kan taşıyan damarların daralması veya tıkanması sonucu oluşur. Kalp kası, yeterli oksijen alamadığında anjinaya (göğüs ağrısı), kalp krizine veya ani ölüme neden olabilir.

Kalp Yetmezliği: Kalbin yetersiz kan pompalaması durumudur. Genellikle sol ventrikül etkilenir ve sol kalp yetmezliği olarak adlandırılır. Ayrıca sağ kalp yetmezliği de meydana gelebilir.

Aritmi: Kalp atışlarının düzensiz veya anormal olduğu durumdur. Bu durumlar kalp

kasındaki elektrik sinyallerinin düzensizliği sonucu oluşabilir ve kalp ritmini etkileyebilir.

Kalp Kapak Hastalıkları: Kalbin dört kapakçığından birinin daralması, kanı sızdırması veya zayıflaması sonucu oluşan durumlardır. Bu durum, kanın kalbin odacıkları arasında düzgün bir şekilde akmasını engelleyebilir.

Miyokardit: Kalp kasının iltihaplanması sonucu meydana gelir. Viral enfeksiyonlar, bakteriyel enfeksiyonlar veya diğer nedenlerden kaynaklanabilir.

Perikardit: Kalbi saran dış zarın iltihaplanması sonucu oluşan bir durumdur. Bu durumda, perikart yumuşak dokusu iltihaplanır ve kalp atışlarını zorlaştırabilir (Beltrame ve Crea, 2018, s. 65).

Bu çeşitli kalp hastalıkları, bireyin yaşam tarzına, genetik yatkınlığa ve diğer sağlık faktörlerine bağlı olarak farklı semptomlar ve tedavi yöntemleri gerektirebilir.

2. LİTERATÜR ARAŞTIRMALARI

Literatürde, birçok tıbbi araştırmada makine öğrenmesi algoritmalarının kullanıldığı görülmektedir.

Dr.A.Govrdhan, K.Srinivas ve B.Kavihta Ravihta Rani tarafından 2010 yılında gerçekleştirilen bir araştırmada, kalp krizlerini tahmin etmeye yardımcı olacak bir sistem tasarımı üzerinde çalışılmıştır. Araştırmada, UCI Machine Learning Repository' den alınan tıbbi veri seti üzerine bağımlılık artırılmış saf Bayes sınıflandırıcısı (ODA NB) ve Bayesian ağı (BN) kullanılmıştır. Bu tekniklerin WEKA platformunda uygulanması sonucunda en yüksek doğruluk oranının %84 olduğu belirlenmiştir (Govrdhan ve diğerleri, 2010, s.68).

H.Jindal ve diğerleri, 2021 yılında yayımladıkları "Kalp Hastalığı Tahmini" başlıklı çalışmalarında UCI' den elde edilen kalp hastalığı veri setini kullanarak k-en yakın komşu algoritması, lojistik regresyon ve rastgele orman algoritmalarını değerlendirmişlerdir. Çalışmanın bulgularına göre, k-en yakın komşu algoritması ve lojistik regresyonun %88.5 doğruluk oranıyla en yüksek performansı göstermiştir (Jindal ve diğerleri, 2021, s.67).

B. Bahrami ve M.Hosseini Shirvani, klinik veri setinde KNN ve J48 gibi farklı sınıflandırma algoritmalarını kullandıkları çalışmalarında, 209 bireyden elde edilen kayıtların olduğu veri seti üzerinde, J48 sınıflandırma algoritmasının WEKA platformunda uygulanmasıyla %83.7' lik bir doğruluk oranı elde etmişlerdir (Bahrami ve Shirvani, 2015, s.65).

B. Venkatalakshmi ve M.V.Shivsankar, kalp hastalığının teşhisine yönelik bir öngörüsel veri madenciliği geliştirerek UCI' den alınan tıbbi veri setini kullandıkları çalışmalarında Karar Ağacı (DT) ve Naive Bayes (NB) algoritmalarını WEKA platformundanda uygulamışlardır. Yaptıkları çalışmada Naive Bayes sınıflandırıcısı %85' lik doğruluk oranıyla en iyi sonucu veren algoritma olmuştur (Venkatalakshmi ve Shivsankar, 2014, s.69).

E.Maini ve diğerleri, kalp hastalığı tahmini için Güney Hindistan'daki bir hastaneden 1670 bireyin kaydını içeren veri setinden veri setinin %70' ini eğitim, %30' unu test verisi olarak ayırıp Python programlama dilinde bir çalışma gerçekleştirmişlerdir. Yaptıkları çalışmadan Naive Bayes, adaboost, rastgele orman ve k-en yakın komşu algoritması gibi çeşitli makine öğrenmesi algoritmalarını kullanmışlardır.

Bu algoritmalar arasında, en yüksek doğruluk oranını %93.8 ile rastgele orman algoritmasının sağladığı gözlemlenmiştir (Maini ve diğerleri, 2021, s.69).

T.Puyalnithi ve M.Vankadara UCI' den elde ettikleri bir veri setini ve Orange aracını kullanarak farklı türde algoritmalar ile bir çalışma gerçekleştirmişlerdir. Çalışmalarında Naive Bayes algoritması %93' lük doğruluk oranıyla en iyi sonucu sağladığını gözlemlemişlerdir (Puyalnithi ve Vandakara, 2017, s.68).

Z. Mahmoodabadi ve M.S. Abadeh Koroner Arter Hastalığı (CHD) teşhisi için yaptıkları çalışmalarında Cleveland ve Hungarian klinik veri setlerini kullanarak MATLAB yazılım dilinde Parçacık Sürü Optimizasyonu (PSO) ve Bağımsız Bileşen Algoritması (ICA) kullanmışlardır. ICA algoritması yaptıkları çalışmada %94.92' lik doğruluk oranıyla en iyi sonucu göstermiştir (Mahmoodabadi ve Abadeh, 2014, s.67).

R.R. Sanni ve diğerleri, kalp hastalığını tahminlemek üzerine yaptıkları çalışmada, kalp hastalığıyla ilgili elde ettikleri veri setini, Python programlama dilini ve denetimli öğrenme algoritmalarından karar ağacı, lojistik regresyon, k-en yakın komşu ve rastgele orman algoritmalarını kullanmışlardır. Araştırmanın sonuçlarına göre, en yüksek doğruluk oranını %85 ile karar ağacı algoritması sağlamıştır (Sanni ve Guruprasad, 2021, s.68).

R.Katarya ve S. K. Meena, UCI Repository' den elde ettikleri veri setini kullanarak kalp hastalığı tahmini gerçekleştirmişlerdir. Çalışmalarında, veri setindeki 76 öznitelikten 14' ünü kullanmışlardır. Araştırmada lojistik regresyon, Naive Bayes, destek vektör makineleri, k-en yakın komşu algoritması, karar ağacı, rastgele orman algoritması, yapay sinir ağları, derin sinir ağı ve çok katmanlı algılayıcı gibi çeşitli algoritmalar değerlendirilmiştir. Bu algoritmalar karşılaştırmalı analizlere tabi tutulmuş ve sonuçlar, lojistik regresyonun %93.40, destek vektör makineleri ve yapay sinir ağlarının %92.30 doğruluk oranına sahip olduğunu, rastgele orman algoritmasının %95.60 doğruluk oranıyla en yüksek performansı sergilediğini göstermiştir. Katarya ve Meena' nın yaptığı bu çalışma, farklı makine öğrenmesi algoritmalarının kalp hastalığı tahminindeki performanslarını karşılaştırarak, en etkili yöntemi belirleme amacını taşımaktadır (Katarya ve Meena, 2021, s.68).

3. YAPAY ZEKA

Yapay zeka terimi, ilk kez 1955 yılında düzenlenen Dartmouth Konferansı' nda ortaya atılmış ve o tarihten bu yana tanımı üzerinde çeşitli çalışmalar yapılmıştır. Konferansta yapay zeka kavramını ilk kez kullanan Prof. John McCarthy, yapay zekayı insan benzeri düşünebilen, kararlar alabilen, insanların yaptığı işleri gerçekleştirebilen ve problemleri çözebilen makineler olarak tanımlamıştır. Ancak günümüzde, yapay zeka tanımı bu geniş kapsamlı tanımla sınırlı kalmayıp daha çeşitli ve kapsamlı bir biçimde ele alınmaktadır (Boden, 2016, s.65).

Yapay zeka tarihinin kökenleri, İngiliz matematikçi Alan Mathison Turing' in 1950 yılında Mind dergisinde yayımlanan "Computing Machinery and Intelligence" adlı makalesine dayanır. Turing' in bu makalesinde ortaya attığı "Makineler düşünebilir mi?" sorusu, yapay zeka çalışmalarının başlangıcı olarak kabul edilir. John McCarthy, "What is Artificial Intelligence?" adlı makalesinde, yapay zeka alanındaki ilk çalışmaların Alan Turing tarafından yapıldığını belirtmiş ve yapay zeka terimini "Akıllı makinelerin, bilhassa zeki bilgisayar programları yapma bilimi ve mühendisliği" olarak tanımlamıştır. (Copeland, 2004, s.65).

Alan Turing "Computing Machinery and Intelligence" adlı makalesinde "Taklit Oyunu" olarak adlandırdığı bir oyunu tanımlar. Bu oyunda erkek, kadın ve sorgulayıcı olarak belirlenen üç oyuncu vardır. Sorgulayıcı, diğer iki oyuncuyu göremeyeceği bir odada bulunur ve onlara yazılı olarak sorular sorar. Cevaplar da yazılı olarak verilir, böylece sorgulayıcı ses tonlarına bakmadan hangi cevabın hangi oyuncuya ait olduğunu tahmin etmeye çalışır. A' nın amacı sorgulayıcıyı yanıltmak iken, B' nin amacı sorgulayıcıya doğru cevapları vererek ona yardımcı olmaktır. Alan Turing' in bu oyunda ortaya koyduğu temel soru, eğer A kişisinin yerini bir makine alırsa sorgulayıcının hangi oyuncunun makine olduğunu belirleyip belirleyemeyeceğidir. Bu bağlamda, Turing makine düşünmesinin doğası hakkında derin bir düşünce yürütür ve "Makineler düşünebilir mi?" sorusuyla yapay zekanın temelini atmış olur. Bu düşünce deneyi, yapay zeka çalışmalarının başlangıcı olarak kabul edilir, zira makinenin insan gibi davranıp davranamayacağı, insan zekasının ve düşünce sürecinin temelini oluşturan önemli bir soruyu ortaya koyar (French, 2000, s.66).

Yapay zekanın temel amacı, parlak zeka seviyesindeki varlıkların davranışlarını yansılayıp, insanlara destek olmak ve muhtelif problemlere çözümler üretmektir. Bu

doğrultuda, yapay zeka teknolojileri, hızla gelişen değişen teknolojinin ilerlemesiyle beraber birçok alanda insan dünyasına kolaylık sağlamaktadır. Günlük iş hayatında sanal asistanlar vasıtasıyla meydana getirilen yardımlar, sürücüsüz araçlarla gerçekleştirilen seyahatler, kişiselleştirilmiş önerilerin sunulmasındaki uygulamalar, yapay zekanın günlük hayatta etkili bir halde kullanılmasına emsal teşkil etmektedir. Ayrıca, kronik rahatsızlıkların tedavisinden iklim değişikliği ile mücadeleye, siber güvenlik tehditlerinin tahmininden ticaret ve finans alanlarına kadar geniş bir yelpazede bu uygulamalar yer almaktadır (Kaplan ve Haenlein, 2019, s.67).

Yapay zeka kavramı, insanların genellikle "zeka" olarak adlandırdığı faaliyetlerin bir makine tarafından da gerçekleştirilebilmesi olarak tanımlanır. Bu, bir bilgisayarın veya bilgisayar kontrolündeki bir robotun, zeki canlılara benzer şekilde çeşitli görevleri yerine getirme yeteneğidir. Yapay zeka sistemleri dışarıdan gelen verileri toplar, bu verileri doğru bir şekilde yorumlar, bu verilere dayanarak öğrenme yeteneği geliştirir ve ardından bu öğrenmeleri insan tarafından belirlenen hedeflere ve görevlere uygun şekilde uygular (Öztürk ve Genç, 2019, s.68).

Yapay zeka kavramı, literatürde değişik tanımlarla karşımıza çıkmaktadır. Kesin bir tanımı olmamakla birlikte, yapay zeka çoğu zaman canlılara has özellikleri benimseyip, geçmiş bilgilerden öğrenen bu detayları yorumlayabilen ve sonuçlar üretebilen bilgisayar sistemleri olarak tanımlanmaktadır. Bu alanın öncülerinden Marvin Lee Minsky' ye göre yapay zeka, çeşitli durumlarda etkin ve güvenilir bir şekilde hareket edebilen zeki ve akıllı makineler oluşturmakla ilgilidir. Bu tanımlar, yapay zeka terimini geniş bir perspektiften ele alarak, bu alandaki çalışmaların kapsamını ve hedeflerini açıklamaktadır (McCarthy, 2022, s.68).

Yapay zeka, temel olarak insan beyninin işlevlerini taklit etmeye yönelik yazılımların özelliğidir. İnsan zekası bağlamında zeka sahibi olarak tanımlanan yapay zeka sistemleri, günümüzde hayret verici yeteneklere sahiptir. Ancak, gelecekte yapabilecekleri hakkındaki öngörüler, yapay zekayı insan olmayan ancak insan gibi düşünebilen bir varlık olarak betimlemektedir. Dolayısıyla yapay zeka, insan gibi, neredeyse insan gibi veya hatta insan olmayı uman bir varlık olarak da değerlendirilebilir (Erkan, 2017, s.66).

Bazı görüşlere göre, bir sistemin yapay zeka olarak tanımlanabilmesi için dört temel özelliği içermesi gerekmektedir. Bu özellikler insana benzer biçimde hareket edebilme, insana benzer biçimde düşünebilme, akılcı hareket edebilme ve akılcı

düşünebilmedir. Bu özelliklerin bir arada bulunması, bir sistem üstünde yapay zeka bulunduğunu belirlemek için temel kriterler olarak kabul edilir. Yapay zeka, bu özellikleri taşıyan sistemler vasıtasıyla insan zekasının sınırlarını keşfetmeye ve veri işleme kapasitesini artırmaya yönelik bir alan olarak kabul edilir (Yılmaz ve Karahan, 2020, s.69).

Yapay zeka terimine ilişik farklı tanımların kullanılmasının sebebi, bu kavramın mühendislik, psikoloji, sosyoloji, tıbbi bilimlere benzer biçimde birçok değişik alanda uygulanabilir olmasıdır. Dolayısıyla, yapay zeka ancak bilgisayar bilimleri ile sınırlı olarak kalmaz. Bu disiplinler, görüntü işleme, yapay sinir ağları, makine öğrenmesi, doğal dil işleme, genetik algoritmalar, robotbilim, uzman sistemler ve bulanık mantığı içinde bulunduran yapay zekanın alt dallarında yaygın olarak kullanılmaktadır. Bu şekilde, yapay zeka, değişik bilim dalları içinde disiplinler arası bir etkileşimi teşvik ederek benzer ve farklı alanlarda yenilikçi çözümler sunarak topluluğu etkilemeye ve geliştirmeye devam etmektedir (Kaya ve Ertürk, 2019, s.67).

3.1. Makine Öğrenmesi

Günümüzde, hızla gelişen değişen teknolojiyle beraber sağlık, eğitim, finans, güvenlik ve ulaşım alanlarıyla birlikte birçok farklı alanlarda teknolojinin yoğun olarak kullanıldığı görülmektedir. Bu durum, reel ve tüzel kişiler tarafınca üretilen veri miktarının artmasına neden olmuştur. Teknolojinin ilerlemesiyle beraber verilerin depolanması ve erişimi daha basit hale gelmiştir. Bu da verinin değerini artırmıştır. Özellikle, finans sektöründe etkinlik yayınlayan yeni nesil banka ve aracı kuruluşlar, işlemleri komisyon tutarı almadan gerçekleştirebilmektedir. Bu durumun temel nedeni, daha çok veriye erişim sağlayarak çalışılan veri tabanını genişletmek ve stratejik kararlar almak için daha çok veri kullanmaktır. Ancak, bu tahminlerin yapılabilmesi için büyük oranda veriye ve uzmanlığa ihtiyaç vardır. Bu noktada, makine öğrenmesi, veri analizi ve tahminlerde insanları destekleyen bir rol oynamaktadır (Russell ve Norvig, 2016, s.68).

Sanayi devriminin ilk aşamalarından günümüze kadar, makine terimi büyük bir değişiklik geçirmiştir. Bilgisayarların yaşamımıza girmesiyle beraber, makine terimi artık yazılımların içinde bulunduğu birçok donanımı kapsamaktadır. Bu kavramlardan en mühim olanlarından birisi de makine öğrenmesidir. Makine öğrenmesi, donanımlara öğrenme kabiliyetini bilgisayar yazılımları ile kazandıran bir süreçtir. Geleneksel

programlamada, bir görevin yerine getirilmesi için her adım kodlanırken, makine öğrenmesinde algoritmalar veri üstünde öğrenim alarak rolü çözmeyi öğrenir. Blum` a göre, makine öğrenmesi; kural öğrenme kabiliyeti olan, kendini güncelleyebilen ve performansını deneyimlerle geliştiren yazılımları tasarlamayı ihtiva eder. Bu bağlamda, makine öğrenmesinde veriler kullanılarak yazılımlar ile kendi kendine öğrenme kabiliyeti geliştirilir. Makine öğrenmesi alanında özetle şu hususlar vurgulanabilir:

- Makine öğrenmesi bilgisayarlı sistemlerin öğrenme kabiliyetini içerir ve bu öğrenme çoğu zaman bilgisayar yazılımları ile gerçekleşir.
- Deneyimle öğrenme ve performansın artması, öğrenmenin temel unsurları arasındadır ve deneyimlerle beraber performansın artması öğrenme olarak kabul edilir.
- Öğrenme sürecinde veri setleri temel alınır ve bu veri setlerinden yararlı bilgiler çıkarılır, yazılımın ise değişikliklere uyum sağlaması beklenir.
- Öğrenme düzeyi, belli kriterler kullanılarak kontrol edilir ve modellerde parametre ayarlaması yapılabilir.

Bu noktalar, makine öğrenmesinin temel prensiplerini oluşturur ve bu alanda meydana getirilen çalışmaların odak noktalarını belirler (Domingos, 2012, s.66).

Makine öğrenmesi, belirli bir rolü yerine getirmek için kalıplara ve çıkarımlara dayalı algoritmalar ve istatistiksel modeller kullanır. Makine öğrenmesi algoritmaları, eğitim verileri olarak adlandırılan istatistiksel örnek veri modeli oluşturarak öngörülerde ya da kararlarda bulunur. Makine öğrenmesi, yapay zekanın bir alt konusu olarak kabul edilerek uygun komutların geliştirilemediği birçok uygulamada kullanılır. Hesaplamalı istatistik bilimi olarak da geçen makine öğrenmesi, öğrenilen verilerden elde edilmiş tahminlere odaklanarak malum özelliklere dayalı olarak çalışır.

Makine öğrenmesi, kompleks algoritmaları kullanarak veri setlerinden öğrenme sağlar ve bu öğrenme sürecinde istatistiksel modellerden yararlanır. Öğrenme süreci, belirlenen bir rolü gerçekleştirmek için lüzumlu olan kalıpları tarif etmek ve çıkarımlar yapmak üstüne odaklanır. Bu süreçte, algoritmalar eğitim verilerini kullanarak bir model oluşturur ve çıkarımlar yapar.

Bu bağlamda, makine öğrenmesi ve derin öğrenme, yapay zeka alanının alt kolları olarak düşünülebilir. Yapay zeka, geniş bir terimi kapsarken, makine öğrenmesi ve derin öğrenme, yapay zekanın altındaki belirli metot ve tekniklerin uygulanmasını belirtir. Derin öğrenme ise, makine öğrenmesinin daha spesifik bir alt dalı olarak kabul edilir, bilhassa büyük ve karmaşık veri kümeleri üstünde çalışırken etkilidir. Şekil 3.1.'de yapay

zeka, makine öğrenmesi ve derin öğrenme arasındaki ilişki, yapay zeka alanının geniş kapsamını temsil eden bir hiyerarşik yapı içinde gösterilmektedir.

Şekil 3.1. Yapay zeka makine öğrenmesi ve derin öğrenme ilişkisi (Dalkıran ve Ozan, 2022, s.66)

Makine öğrenmesi ile oluşturulan projeler, yapay zeka projelerini destekleyebilecek durumlarda kullanılabilir. Ancak yapay zeka ve makine öğrenmesi birbirleriyle iç içe olduğundan birçok alanda beraber kullanılabilmektedir. Yapılan çalışmalar, tanımlama işleminin makine öğrenmesi örnekleri kullanarak gerçekleştirildiğini göstermektedir. Bu nedenle öğretilen bilgilerin bütün muhtemel senaryolara uyumluluğunun önemi büyüktür. Makine öğrenmesi, veri girişi yeterli olmadığında tahmin hatalarının artmasına sebep olabilir. Dolayısıyla makine öğrenmesinin iyi bir halde eğitilmesi gerekmektedir..

Makine öğrenmesi süreci, üç temel aksiyondan oluşur. İlk olarak, verilerin gözlenmesi ve hafızaya alınarak hemen sonra yapılacak değerlendirme için hatırlamanın sağlanmasıdır. Sonrasında çıkarım aşaması gelmektedir ve verilerin kullanıma imkân verecek halde dönüştürülmesini içerir. Sonuncusu ise genelleştirme aşamasıdır ve aksiyonun alınmasına temel oluşturan verilerin kullanılmasını içerir. Daha detaylı olarak, makine öğrenmesi uygulaması için verilerin toplanmasıyla süregelen bir süreç vardır. Bu süreçte, çözümlenecek veri elektronik ortama aktarılmalıdır (Can ve Kaya, 2019, s.66).

Makine öğrenmesi algoritmasının kalitesi, elde edilmiş verinin kalitesine bağlıdır.

Verinin incelenmesi ve hazırlanması aşaması büyük seviyede insan müdahalesini ve tesirini içerir. Analistin veriyi anlaması ve lüzumlu düzeltmelerin yapılması önemli bir konudur. Modelin eğitilmesi sürecinde, ilk önce veriden öğrenilmek istenen şey açıkça tanımlanmalıdır. Modelin performansının değerlendirilmesi sürecinde, kontrol verileri kullanılarak modelin aldığı kararlar kontrol edilir.

Son olarak, test verileri ile sınanmakta olan modelin istenilen performansı göstermediği durumda, modelin değiştirilmesi ya da veri eklemelerinin yapılması şeklinde alternatifler değerlendirilir. Ham verinin bilgiye dönüşüm dönemi, ham veriler içinde gizli saklı kalmış ilişkilerin ortaya çıkarılması olarak tanımlanır. Bu süreçte, verinin elde edilmesinin yanı sıra tarz uyumsuzlukları, veri tutarsızlıkları, noksan ve yanlış girilmiş veriler ele alınır. Bu sorunlar veri ön işleme olarak adlandırılan süreçten sonra makine öğrenmesi algoritmaları ile analizler yapılarak ham veriden bilgiye dönüşüm dönemi tamamlanır (Tüfekçi ve Doğan 2020, s.69).

Denetimli öğrenme, denetimsiz öğrenme ve takviyeli(pekiştirmeli) öğrenme olmak üzere üç kategoriye ayrılan makine öğrenmesi türleri Şekil 3.2' de gösterilmektedir.

Şekil 3.2. Makine öğrenmesi türleri (Şeyranlıoğlu, 2022, s.69).

Denetimli öğrenme, hem giriş hem de çıkış parametrelerinden oluşan örneklerden yararlanarak model geliştirilmesi ve yeni girdilerin çıktılarla eşleştirilmesi işlemidir. Bu süreçte, verilerdeki öznitelikler çözümlenerek belirli bir sınıfa dair çıktı değişkeni

arasındaki ilişki incelenir. Daha sonra, model bu ilişkiyi temsil eden bir fonksiyon üretir ve bu fonksiyon, kontrol verileri kullanılarak doğrulanır.

Denetimsiz öğrenmede ise model, etiketlenmemiş verilerle çalışır ve temel gaye veri kalıplarını anlamaktır. Bu metot çoğu zaman verilerin kümelenmesine benzer uygulamalarda kullanılır.

Takviyeli öğrenmede ise yazılımların, belirli bir rolü yerine getirmek suretiyle direkt programlanması yerine, çevreleriyle etkileşime girerek öğrenmesi sağlanır. Bu süreç, karşılaşılan duruma bağlı olarak alınan aksiyonlara dayalı ödüllerin ya da cezaların verilmesiyle gerçekleşir. Temel olarak, pekiştirmeli öğrenme deneme-yanılma yöntemine dayanır (Hastie ve diğerleri, 2009, s.66).

3.2. Veri Madenciliği

Hızla değişen teknolojinin gelişmesiyle birlikte, farklı sektörlerde büyük hacimli veri tabanlarının ve geniş çaplı veri oluşumunun artması kaçınılmaz olmuştur. Bu durum, veri tabanlarının ve bilgi teknolojisinin araştırılmasını, bu kıymetli verilerin depolanması ve işlenmesi için daha efektif yöntemlerin geliştirilmesini teşvik etmiştir. Bu bağlamda, büyük veri setlerinden kıymetli bilgilerin ve örüntülerin çıkmasına imkan veren veri keşfi mühim bir role sahiptir. Bunun yanında veri keşfi bilgi edinme periyodu olarak adlandırılır ve bu süreç veri madenciliği olarak bilinir.

Veri madenciliği, büyük veri miktarlarını inceleyerek faydalı desenleri ve detayları ortaya çıkarma sürecidir. Temel amacı, büyük veri havuzlarından anlamlı bilgiler çıkarmaktır. Elde edilen bu bilgiler, işletmelerde mühim kararların alınmasında kullanılabilir. Dolayısıyla, veri madenciliği işletmelerin rekabet avantajları elde etmelerine ve daha doğru kararlar almalarına destek olabilir.

Bilginin her geçen gün daha çok değerlenmesiyle birlikte, bilgiye erişim ve elde edilmiş verilerin çözümlenme zorunluluğu bulunmaktadır. Geçmişte bilgiye ulaşmak zor ve vakit alıcı bir süreçken, sınırlı karar mekanizmaları kullanıldığı için beklenen veriye ulaşmak kolay olmuyordu. Günümüzde ise bilgiye erişim kolaylaşmış ve verilerin çözümlenmesi için birçok metot ve karar mekanizması geliştirilmiştir.

Veriden en yüksek faydayı elde etmek için uygulanan yöntemlerin karmaşıklığı ve zorluğu, bilgisayarların kullanımını artırmış ve farklı algoritmalar geliştirilerek "Veri Madenciliği" terimi ortaya çıkmıştır. Bu kavram, büyük veri kümelerinden anlamlı

bilgilerin ve desenlerin keşfedilmesini/çıkarılmasını sağlayarak, işletmelerin daha bilgili kararlar almasına imkan tanır (Konuralp, 2018, s.67).

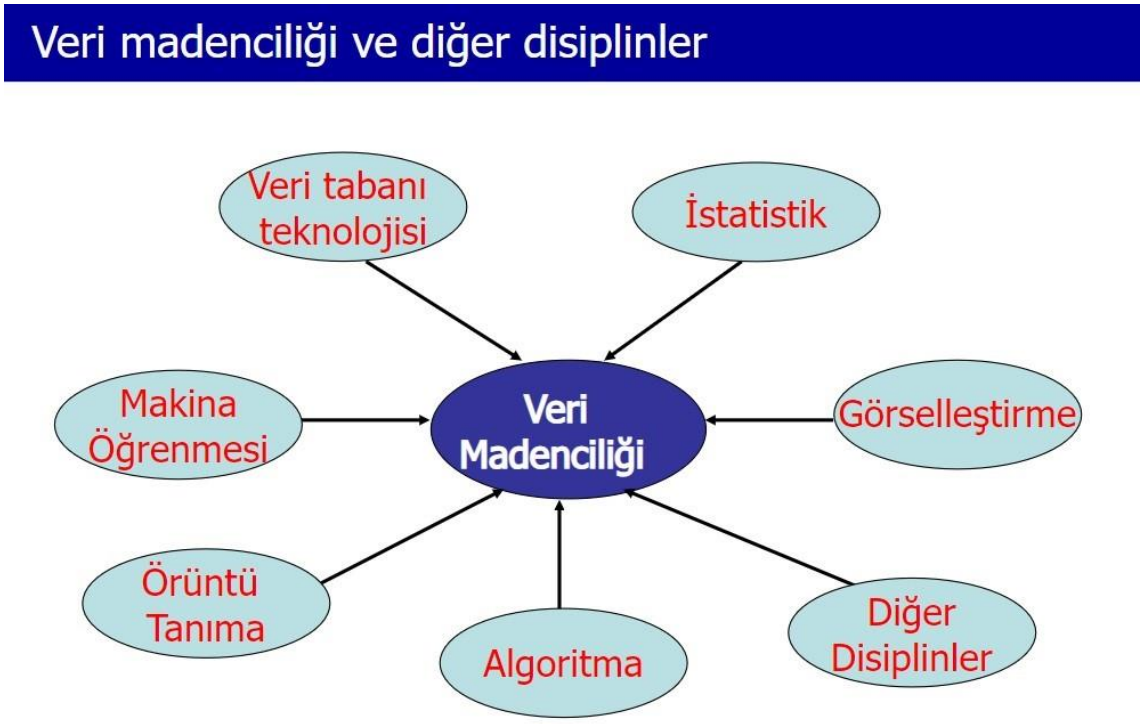
Veri madenciliği, geçmiş yıllardan günümüze kadar devamlı olarak gelişen ve hayatımızı kolaylaştıran bir alan olmuştur. 1950` lerde süregelen yapay zeka ve makine öğrenmesi çalışmaları, veri madenciliği tekniklerinin gelişimine imkan tanımıştır. Bilgisayar teknolojilerindeki ilerlemeler, 1970` lerden 1990` lara kadar olan süreçte yeni programlama dilleri ve algoritmaların ortaya çıkmasını ve yeni tekniklerin geliştirilmesini sağlamıştır.

Veri madenciliği alanındaki mühim bir adım, 1989` da KDD (IJCAI)-89 Veri Tabanlarında Bilgi Keşfi Çalışma Grubu toplantısı ve 1991` de yayımlanan "Knowledge Discovery in Real Databases: A Report on the IJCA89 Workshop" makalesi ile atılmıştır. Bu süreç, 1991` de veri madenciliği için ilk yazılımın geliştirilmesiyle zirve noktasına ulaşmıştır (Piatetsky-Shapiro ve Frawley, 1991, s.68).

2000` li yıllarda, veri madenciliği devamlı olarak gelişmiş ve birçok alanda uygulanmaya başlanmıştır. Veri madenciliği uygulamalarındaki faydaların görülmesiyle birlikte, alaka artmış ve veri madenciliği giderek daha mühim hale gelmiştir. 1990` lardan itibaren, veri madenciliği terimi, veri tabanlarındaki bilgilerin keşfi çalışmalarında kullanılmış ve zaman içinde veri tabanında bulunmayan verileri de kapsayacak halde genişlemiştir. Bu geçmişteki çalışmalar, veri madenciliğinin geleceği hakkındaki mühim ipuçları sunmaktadır.

Veri madenciliği, bilgisayar bilimi alanında fazlaca önemli bir yer tutan ve veri kümeleri içerisindeki ilişkileri keşfetme ve gelecekteki tahminler için çözümler bulma amacı güden bir tekniktir. Bu yöntem, hususi teknikler ve metotlar kullanarak veri ilişkilerini tespit eder ve veriyi anlamlı bir bilgiye dönüştürür. Veri madenciliği, veri tabanı yönetimi, yapay zeka, istatistik ve makine öğrenmesi benzer biçimde değişik alanlardan gelen teknikleri bir araya getirerek bilgi keşfi yapmayı olası kılar (Akay, 2015, s.65).

Veri madenciliği, modern bilgi çağında büyük öneme sahip bir alandır. Bu alandaki teknikler, farklı disiplinlerden beslenerek gelişmiş ve birçok değişik uygulama sahasına yayılmıştır. Şekil 3.3' te veri madenciliğinin disiplinlerle ilişkisi belirtilmiştir. Bu disiplinlerin bir araya gelmesiyle, veri madenciliği bugün endüstriyel uygulamalardan bilimsel araştırmalara kadar geniş bir yelpazede kullanılmaktadır.



Şekil 3.3. Veri madenciliği uygulama alanları (Kumdereli, 2012, s.67).

Bilgisayar bilimi alanında büyük bir öneme sahip olan veri madenciliği, günümüzde giderek artan oranda verinin üretilmesi ve birikmesiyle beraber büyük bir önem kazanmaktadır. "Madencilik" terimi, yeryüzündeki gizli saklı ve kıymetli kaynakların keşfedilmesini konu edinirken, "veri madenciliği" ise veri kümeleri içerisindeki kıymetli bilgilerin keşfedilmesi ve çıkarılmasını çağrıştırmaktadır.

Veri madenciliği, bir bilim disiplini olarak tanımlandığında, değişik araştırmacılar ve uzmanlar tarafınca değişik şekillerde ele alınabilir. Literatürde yer edinen tanımlar incelendiğinde, veri madenciliğinin makine öğrenmesi, örüntü tanıma, istatistik, veri tabanı ve görselleştirme tekniklerini birleştirerek geniş veri tabanlarından bilgi çıkarma amacı güden bir disiplin olduğu ortaya çıkar (Fayyad ve diğerleri, 1996, s.66).

Bu disiplini tarif etmek için kullanılan tanımlardan biri, gözlemsel veri setlerinin analiz edilerek veri sahibi için anlaşılabilir ve yararlı bilgilerin çıkarılması periyodunu vurgular. Diğer bir tarif ise verinin otomatik ya da yarı-otomatik çözümlemeyle gizli saklı örüntülerin bulunmasını öne çıkarır. Büyük veri depolarında yararlı bilgilerin otomatik olarak keşfedilmesi ve veri ambarlarında depolanan büyük miktardaki verinin istatistiksel ve matematiksel tekniklerle beraber incelenerek yeni ilişkiler, örüntüler ve eğilimlerin bulunması da veri madenciliği süreçlerinin mühim bileşenleridir (Witten ve diğerleri,

2011, s.69).

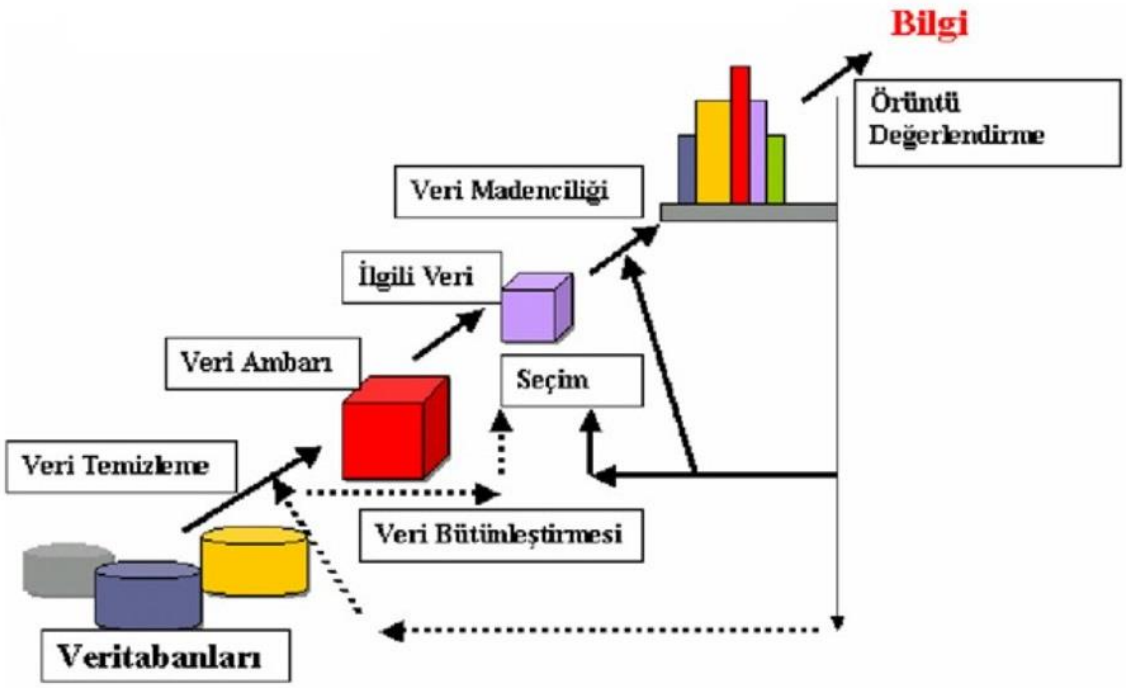
Veri madenciliği, büyük veri setlerinde ya da veri tabanlarındaki veriler arasında bilinmeyen, klasik yöntemlerle tespit edilmeyen, basit olmayan ilişkileri, örüntüleri, belirli yapıları veya eğilimleri ortaya çıkarmak için istatistik, matematik, makine öğrenmesi ve bilgisayar uygulamalarının birleşimini kullanarak verileri analiz etme ve bu analiz sonuçlarını anlamlı bir şekilde özetleyip görselleştirme sürecidir.

Bu doğrultuda, veri madenciliği disiplini, modern bilgisayar bilimi ve istatistiksel yöntemleri sentezleyerek veri analizi alanında önemli bir rol oynamaktadır. Bu disiplin, bilgiye dayalı karar verme süreçlerini destekleyerek, iş dünyasında, bilimde ve diğer birçok alanda uygulama alanları bulmaktadır. Veri madenciliği, veri odaklı dünyamızda bilgiye ulaşma, anlama ve değerlendirme sürecini güçlendiren ve zenginleştiren önemli bir araç olarak kabul edilmektedir (Şen, 2018, s.69).

Veri madenciliği periyodu belirli aşamalardan doğar ve bu aşamaların dikkatle izlenmesi gerekir. Veri ambarları, bir organizasyonun gereksinimlerine uygun olarak büyük miktarda verinin kolay erişilebilir bir yapıda saklanmasını sağlayan bilgisayar tabanlı depolama sistemleridir. 1990` lı yıllarda ortaya çıkan veri ambarları, veriyi kullanılabilir ilişkiler ve profillerde sınıflandırmayan ancak potansiyel bilgi içeren veri tabanlarıdır. Gerçek değerli bilginin keşfedilmesini sağlayan ise veri madenciliği gibi tekniklerdir.

Veri ambarından veriyi çekebilmek için hangi verinin gerektiğini ve bu verinin nerede olduğunu belirlemek önemlidir. Gerekli verinin çoğu zaman değişik sistemlerde ve değişik formatlarda bulunması nedeniyle, ilk aşamada veri temizleme ve düzenleme işlemi yapılmalıdır. Veri ambarının kurucusu olan W.H. Inmon' a göre, veri ambarı, verinin temizlendiği, birleştirildiği ve tekrar düzenlenmiş olduğu merkez ve bütünleşmiş bir depodur Özkan, 2017, s.68).

Veri madenciliği, veri tabanları, veri ambarı, veri bütünleştirmesi, ilgili veri ve seçim gibi çeşitli kavramları içeren kapsamlı bir süreçtir. Şekil 3.4’ te veri madenciliği süreci görselleştirilmiştir. Bu kavramlar, büyük veri kümelerinden anlamlı bilgi çıkarmak için kullanılan temel bileşenleri temsil eder. Veri madenciliği projelerinde, bu kavramlar doğru şekilde anlaşılmalı ve etkili bir şekilde uygulanmalıdır.



Şekil 3.4. Veri madenciliği süreci (Altıntaş, 2006)

Veri madenciliği alanında kullanılan birçok program mevcuttur. Bunlar arasında, açık kaynak kodlu olanlar özellikle önemlidir. Örnek olarak, R, Orange, OpenNN, Carrot2, Gate, Elki, Torch ve Weka gibi programlar verilebilir. Bu programlar, veri madenciliği çalışmalarında kullanılabilecek çeşitli araçlar ve özellikler sunar. Bununla birlikte, açık kaynaklı programlar dışında, pazarda birçok ücretli program da bulunmaktadır. Örneğin, RapidMiner, IBM SPSS Modeler, Microsoft Analysis Services, Oracle Data Mining, Clarabridge gibi programlar, geniş bir kullanıcı kitlesi tarafından tercih edilmektedir. Bu programlar, genellikle daha fazla özellik ve destek sunarak kullanıcıların veri madenciliği projelerini daha etkili bir şekilde yönetmelerine olanak tanır. Ancak, hangi programın en uygun olduğu, projenin gereksinimlerine ve kullanıcıların tercihlerine bağlı olarak değişebilir. Bu nedenle, kullanıcılar projelerine en uygun programı seçerken dikkatli olmalı ve gereksinimlerini göz önünde bulundurmalıdır Demir, 2019, s.66).

4. YÖNTEM

4.1. Weka

Weka, günümüzde oldukça popüler ve sıkça araştırılan bir veri madenciliği uygulamasıdır. Yeni Zelanda`nın Waikato Üniversitesi tarafından ücretsiz GNU lisansı altında geliştirilmiş olan bu modüler uygulama, birçok metot, algoritma, hazır fonksiyon ve kütüphane içermektedir. Kullanıcıların isteklerine göre Weka platformundan indirilebilen ve programa entegre edilebilen birçok özelliği barındıran Weka, modüler yapısı sayesinde geniş bir kullanım imkanı sunmaktadır.

Weka platformu, veri madenciliği süreçlerinde sınıflandırma, kümeleme, ilişkilendirme, veri ön işleme ve görselleştirme vb. işlemleri kolayca gerçekleştirebilme imkanı sağlar. Ancak, bu işlemleri gerçekleştirebilmek için kullanılacak verilerin arff formatında olması gerekmektedir. Ayrıca, Weka değişik uzantılardaki verilerin dönüştürülmesini kolaylaştıran özelliklere de sahiptir.

Waikato Üniversitesi tarafınca geliştirilmiş ve açık kaynak kodlu bir makine öğrenmesi aracı olan Weka` nın ismi, "Waikato Environment for Knowledge Analysis" ifadesinin baş harflerinden gelmektedir. Bunun yanısıra, Weka ismi Yeni Zelanda' da nesli tükenmekte olan bir kuşun ismini da taşımaktadır. Weka, veri madenciliği ve makine öğrenmesi alanında sıkça kullanılan bir platform olup, içerisinde bulunan farklı algoritmalar yardımıyla birçok veri setinde analizler yapılmasını sağlar (Çelik ve Öztürk, 2019, s.66).

Weka, veri setlerinin düzenlenmesine ve birçok veri kümesinde farklı yöntemlerin kullanılarak hızlı sonuçlar elde edilmesine imkan sağlayacak biçimde tasarlanmıştır. Ayrıca, Weka kullanıcıları sınıflandırma, kümeleme ve birliktelik vb. temel veri madenciliği çalışmalarını kullanabilmektedir.

Biyoinformatik veri kümelerinde gerçekleştirilen analizlerde de Weka platformu başarıyla kullanılmaktadır. Örneğin, Kretschmann ve arkadaşları Weka kullanarak proteinlerin açıklanması üzerine başarı göstermiş çalışmalar gerçekleştirmiştir. Kretschmann ve diğerleri, 2015, s.67). Benzer şekilde, Tobler ve arkadaşları Weka`yı gen dizilimi analizi için kullanmış (Tobler ve diğerleri, 2015, s.69), Li ve arkadaşları ise kanser teşhisi konulu çalışmalarda Weka platformunu tercih etmişlerdir (Li ve diğerleri, 2019, s.67).

Weka, kullanıcılara işlemlerin niteliğine göre beş farklı arayüz sunmaktadır: Explorer, Experimenter, KnowledgeFlow, WorkBench ve Simple CLI' dir. Bu arayüzler, kullanıcıların detaylı analizler yapmalarına olanak tanır ve çeşitli işlemleri kolayca gerçekleştirmelerine yardımcı olur. Bu tez çalışması Explorer ve Experimenter arayüzlerindeki işlemlerden faydalanılarak gerçekleştirilmiştir. Şekil 4.1'de Weka ana ekranında bulunan ara yüzler gösterilmiştir.

Şekil 4.1. Weka ana ekranı

Bu tez çalışması gerçekleştirilirken Weka` nın 3.8.6 versiyonu kullanılmış olup programın ana ekranın sağ kısmında 5 değişik ara yüzü gözükmektedir. "Explorer" Weka` nın kullanımında en çok tercih edilen arayüzdür. "Experimenter", seçeneği birden fazla deneyin sistemli bir biçimde yürütülmesine imkan tanır. Farklı algoritmalar, parametre ayarı ya da veri seti kombinasyonunu kontrol etmek istediğinizde kullanışlıdır. Sonuçları karşılaştırmak için istatistiksel analizler de sağlar. "KnowledgeFlow", bir görsel programlama aracıdır ve veri akışlarındaki işlemleri ve analizleri yapmaya imkan tanır. Veri ön işleme, model eğitimi, sınıflandırma ve değerlendirme şeklinde işlemlerin görsel olarak tasarlanmasına izin verir. "Workbench", özel makine öğrenmesi iş akışlarını yapmaya yarayan bir araçtır. Özelleştirilmiş modüller eklenebilir ve mevcut araçların işlevselliği artırılabilir. "Simple CLI", komut satırı ara yüzüdür. Bu, otomatikleştirilmiş işlemler için kullanışlıdır ya da Weka` nın komut satırından entegrasyonunu sağlamak isteyen kullanıcılar için idealdir. "Explorer", veri setlerinde keşifsel veri analizi

yapılmasını sağlar. Veri setleri bu kısımdan Weka` ya yüklenir. Yüklenilen veri setlerinin görselleştirilmesi, ön işlemesi, model seçilmesi, eğitilmesi, sonuçların değerlendirilmesi ve performans analizi mevzusunda kullanıcıya destek olur. Şekil 4.2' de Weka Explorer ara yüzü ve verinin ön işleme kısmı gösterilmektedir. Bu çalışmada Explorer ve Experimenter arayüzleri kullanılarak gerçekleştirilmiştir.

Weka ana ekranının üst bölümünde dört değişik seçenek sunmaktadır. "Program" seçeneği Weka` nın temel işlevselliğini içerir ve çeşitli makine öğrenmesi algoritması görevleri için araçlar bulunur. "Visualization", verilerinizi ve sonuçlarınızı görselleştirmenizi sağlar. Örneğin, veri dağılımını görselleştirmek için farklı grafikler oluşturabilir ya da model performansını değerlendirmek için çeşitli metriklerin grafiklerle görüntülenmesini sağlar. "Tools", Weka` nın destek araçlarını içerir. Veri ön işleme araçları, veri dönüştürme araçları, özellik tarzı araçları ve öteki yardımcı programlar burada bulunabilir. Bu araçlar veri setleri hazırlanırken ve analiz için kullanışlıdır. Weka` nın default kısmında bizlere sunmuş olduğu algoritmalara, harici veri tabanında bulunan algoritmalar bu kısımdan eklenebilir. "Help", Weka` nın kullanımıyla alakalı yardım belgelerine erişimi sağlar. Bu belgeler Weka` nın temel kavramları, işlevselliği ve nasıl kullanılacağına dair kılavuzlar içerir.

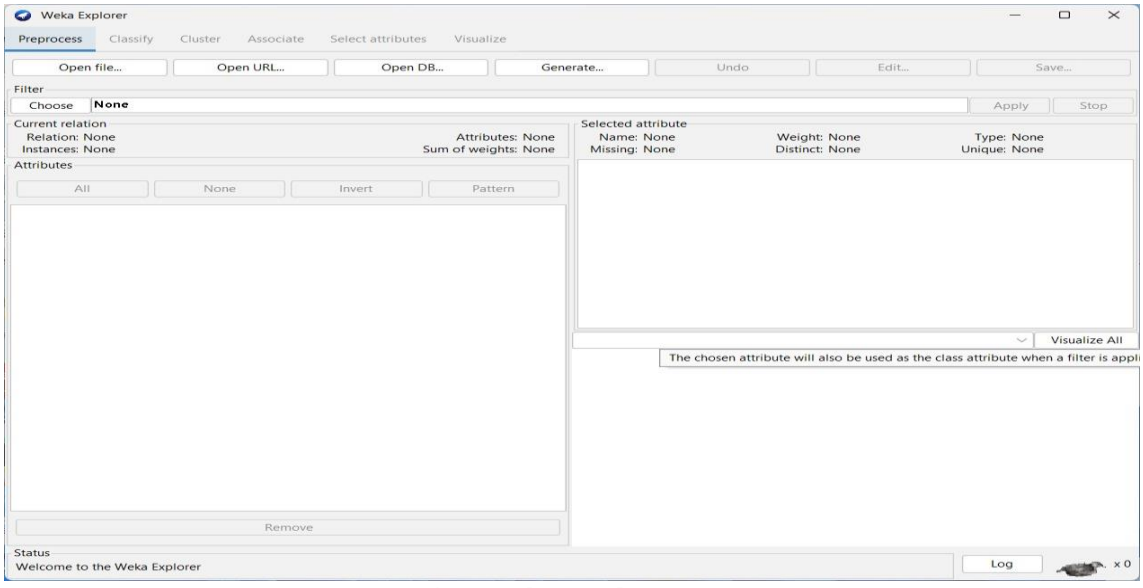
"Preprocess", veri seti farklı ön işleme işlemlerini gerçekleştirmeyi sağlar. Örneğin, eksik değerleri doldurma, veri normalleştirme ya da standartlaştırma, veri dönüşümleri yapma vb. işlemleri içerir. Veri setini makine öğrenmesi algoritmalarına hazırlamak için bu kısım önemlidir. "Classify", veri seti üstünde sınıflandırma ve regresyon işlemlerinin gerçekleştirilmesini sağlar. Sınıflandırma, verileri belirli bir sınıfa belirleme işlemidir. Bu seçenek, değişik algoritmalar kullanarak modeller yapmaya ve kontrol etmeye imkan tanır. "Cluster", veri seti üstünde kümeleme işlemleri gerçekleştirmeyi sağlar. Kümeleme, veri noktalarını benzerliklerine nazaran gruplandırma işlemidir. Bu gruplar, veri setindeki doğal yapıyı keşfetmek için kullanılabilir. "Associate", veri setinde ilişki analizi ya da kural çıkarma işlemlerini gerçekleştirmeyi sağlar. Bir market sepeti analizi yaparak müşterilerin hangi ürünleri beraber satın aldığını belirlemek buna örnek olabilir. "Select Attributes", modeli eğitmek için kullanılan öznitelikleri seçmeyi sağlar. Gereksiz ya da nötr öznitelikleri çıkarmak, modelin performansını artırabilir ve hesaplama maliyetini azaltabilir. "Visualize", veri setinin ve açılan modelin görsel olarak incelenmesini sağlar. Veri setinin dağılımını, sınıflar arasındaki ilişkileri ya da modelin sınıflandırma doğruluğunu görsel olarak

görebilmeye destek olur (Özlen, 2022, s.68).

Weka Explorer menüsündeki "Open file", "Open URL", "Open DB", "Generate", "Undo", "Edit", "Save" gibi seçenekler, veri işleme ve çalışma akışını yönetmek için kullanılan temel araçlardır. İşlevleri şu şekildedir:

“Open file”, bilgisayarınızda bulunan dosyayı Weka` ya yüklemenizi sağlar. “Open URL”, bir URL' yi kullanarak çevrimiçi bir veri setini Weka' ya yüklemenizi sağlar. Veri setinin web üstünde bulunmuş olduğu durumlarda kullanışlıdır. “Open DB”, bir veri tabanından veri seti almanızı sağlar. Weka, veri tabanı linkleri vesilesiyle veri alabilir ve analiz edebilir. “Generate”, örnek veri setleri ya da yapay veri setleri kurmayı sağlar. Örneğin, Weka' nın entegre etmiş olduğu veri oluşturma araçları yardımıyla bir dağılıma ait veri setleri oluşturabilir. “Undo”, son meydana getirilen değişikliği geri almanıza imkan tanır. Özellikle veri işleme ya da modelleme esnasında yanlış bir adım atıldığında kullanışlıdır. “Edit”, yüklü veri setlerini düzenlemeye imkan tanır. Veri setindeki bulunmayan değerleri düzenler. “Save”, çalışmanın kaydedilmesini sağlar (Özlen, 2022, s.68).

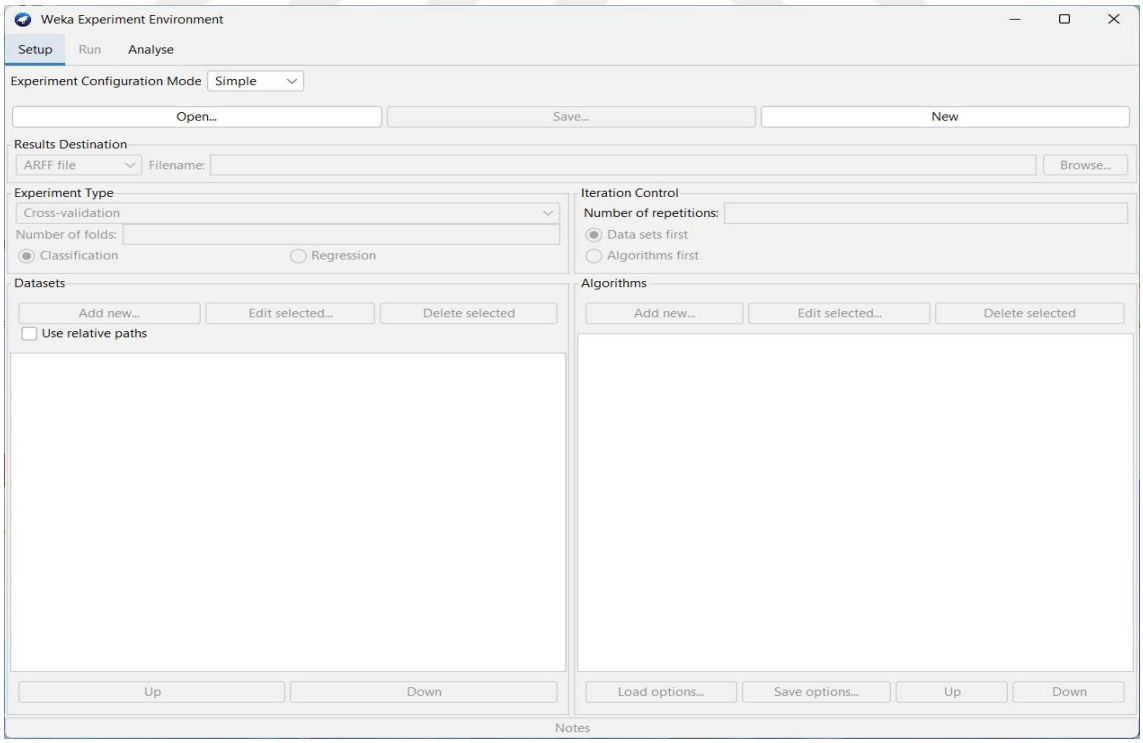
Weka Explorer kısmı genel olarak, Weka' nın keşifsel veri analizi, model oluşturma ve model değerlendirme gibi çeşitli veri madenciliği ve makine öğrenmesi görevlerini desteklemesini sağlar. Weka' nın veri yükleme, veri oluşturma, veri düzenleme ve çalışmayı kaydetme gibi temel işlevlerini içerir. Bu araçlar, kullanıcıların Weka içinde veri işleme ve analiz süreçlerini yönetmelerini kolaylaştırır. Şekil 4.2’ de Explorer kısmı ara yüzü gösterilmiştir.



Şekil 4.2. Weka Explorer ara yüzü

Weka Experimenter menüsündeki seçeneklerin görevleri şu şekildedir. “Setup”, yeni bir deneyin parametrelerini ve ayarlarını yapılandırmaya olanak tanır. Hangi algoritmaların kullanılacağı, hangi parametrelerin değiştirileceği, çapraz doğrulama yapıp yapılmayacağı gibi deneyin yapılandırması burada gerçekleştirilir. “Run”, belirlenen deneyi çalıştırmaya olanak tanır. Weka, seçilen parametrelerle belirlenilen algoritmayı veya algoritmaları veri seti üzerinde çalıştırır ve sonuçları toplar. “Analyse”, çalıştırılan deneyin sonuçlarını analiz etmeyi sağlar. Deneyin sonuçlarını görsel olarak incelemek, performans metriklerini karşılaştırmak veya istatistiksel analizler yapmak için kullanışlıdır. “Open”, daha önceden kaydedilmiş bir deneyin sonuçlarını açmanıza olanak tanır. Daha önce yapılan deneylerin sonuçlarını incelemek veya yeniden analiz etmek için kullanılır. “Save”, yapılandırılan veya sonuçlarını alınan bir deneyi kaydetmeyi sağlar. “New”, yeni bir deney oluşturulmasını sağlar.

Weka Experimenter arayüzü temel olarak, farklı algoritmaların performansını karşılaştırmak, parametre ayarlamak ve makine öğrenmesi modellerini iyileştirmek için kullanışlıdır. Şekil 4.3’ te Experimenter ara yüzü gösterilmiştir.



Şekil 4.3. Weka Experimenter ara yüzü

4.2. Veri Seti

Bu çalışmanın veri seti, Kaggle platformundaki "Heart Failure Prediction" başlığı altında bulunan bir veri setinden elde edilmiştir. Bu veri seti, bağımsız olarak var olan ancak önceden birleştirilmemiş farklı kalp veri setlerinin bir araya getirilmesiyle oluşturulmuştur. Toplamda beş farklı kalp veri seti (Cleveland, İsviçre, Macaristan, Long Beach VA, Stalog) incelenmiş ve bu veri setlerinde bulunan 11 ortak özellik birleştirilerek toplam 1190 gözlem yapılmıştır. Ancak, yapılan gözlemler sonucunda 272'si elenerek toplamda 918 veriden oluşan bir veri seti hazırlanmıştır (http-3, s. 69).

Kardiyovasküler hastalıklar, dünya genelinde yıllık tahmini 18.6 milyon can kaybına neden olarak, dünya genelindeki ölümlerin yaklaşık %31' ini oluşturan en önemli ölüm nedenlerinden biridir. Bu hastalıklardan kaynaklanan ölümlerin dörtte üçü kalp krizi ve felç gibi olaylardan meydana gelmektedir. Bu ölümlerin yaklaşık üçte biri ise 70 yaşın altındaki bireylerde erken yaşta meydana gelmektedir. Kalp yetmezliği, kardiyovasküler hastalıkların yaygın bir sonucu olup, bu veri seti, olası bir kalp hastalığını tahmin etmek için kullanılabilecek 11 farklı özelliği içermektedir (http-2, s. 69).

Kardiyovasküler hastalıklara sahip veya yüksek kardiyovasküler risk altında olan bireyler, hipertansiyon, diyabet, hiperlipidemi gibi bir veya daha fazla risk faktörünün varlığı nedeniyle erken teşhis ve müdahaleye ihtiyaç duymaktadırlar. Bu nedenle, bu tür hastalıkların belirlenmesi ve önlenmesi için etkili yöntemlerin geliştirilmesi büyük önem taşımaktadır.

Bu veri seti, toplamda 11 farklı özniteliği içermektedir. Bu öznitelikler; yaş, cinsiyet, göğüs ağrısı tipi, dinlenme kan basıncı, kolesterol seviyesi, açlık kan şekeri, istirahat elektrokardiyogram sonuçları, maksimum kalp atış hızı, egzersize bağlı anjina, depresyonda ölçülen değer ve ST eğiminin yönüdür. Bu özniteliklere bağlı değişken ise kalp hastalığıdır (http-3, s. 69).

Veri setinde yer alan dinlenme kan basıncı ve kolesterol özniteliğinde eksik veriler bulunmaktadır. Bu eksik veriler Excel ve Weka' nın preprocess kısmı ile düzenlenip tamamlanarak iki farklı veri üzerinde çalışılmıştır. Weka üzerinde kullanılacak olan algoritmaya göre bazı verilerin nümerik-kategorik, bazılarının ise nominal olması gerekmektedir. Bu nedenle, veriler nümerik ve nominal olarak düzenlenmiştir. Bu düzenleme, veri setinin analiz ve işlenme süreçlerinde kullanılabilirliğini artırmış ve doğru sonuçların elde edilmesine olanak sağlamıştır. Bu veriler iki farklı şekilde düzenlenerek sonuçlar incelenmiş ve performanslar kıyaslanmıştır. Veri seti

düzenlenirken kullanılan ilk yöntem Excel üzerinden tıbbi literatürde belirlenmiş standartlara göre veri setini düzenlemek olmuştur. Diğer yöntem ise Weka' nın veriyi önişleme kısmında bize sunduğu parametreler aracılığıyla veri setini düzenlemek olmuştur.

Orijinal veri setinde nümerik olarak gösterilen "RestingBP", "Cholesterol", "FastingBS", "MaxHR", "Oldpeak" gibi öznitelikler, tıbbi literatürde belirlenmiş standartlar baz alınarak değerlendirilmiş ve Excel ile Çizelge 4.1' de görüldüğü gibi kategorize edilmiştir. Weka' da nominal verilerin istendiği algoritmaların çalışması için veri seti bu şekilde oluşturulmuştur.

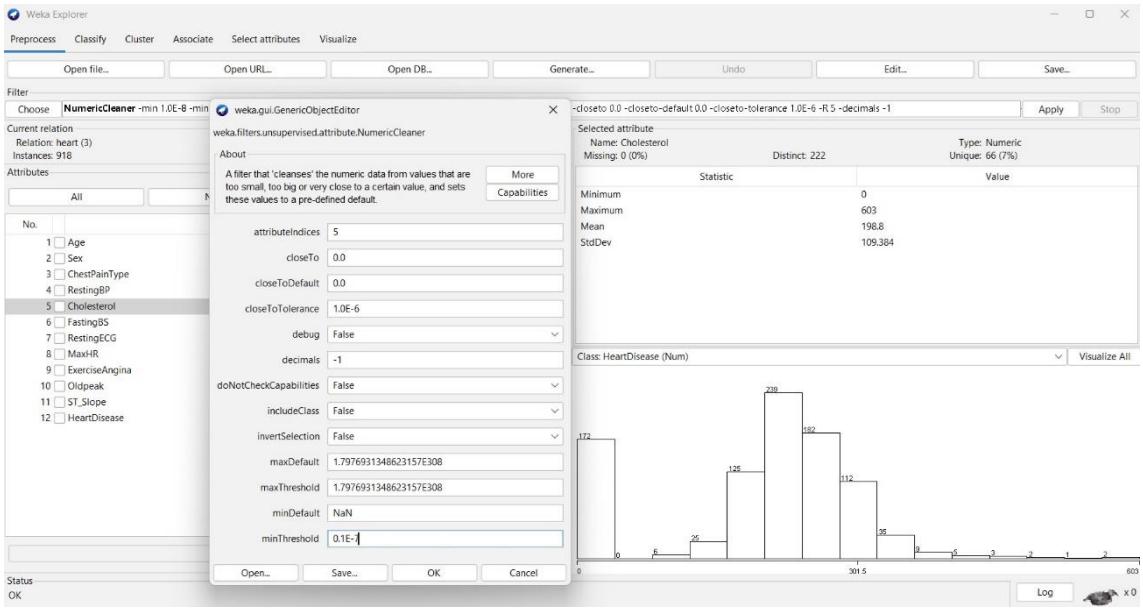
Çizelge 4.1. Veri seti özellikleri

Öznitelik	Tanımlama	Nümerik-Nominal- Kategorik
Age	**Yaş**	25 ≤ genç ≤ 49 50 ≤ orta yaş ≤ 64 65 ≤ yaşlı
Sex	**Cinsiyet**	1=erkek, 0=kadın
ChestPainType	**Göğüs Ağrısı Türü**	TA-1, ATA-2, NAP-3, ASY-4
RestingBP	**Dinlenme Kan Basıncı**	0 ≤ düşük ≤ 89, 90 ≤ normal ≤ 120, 121 ≤ yüksek
Cholesterol	**Serum Kolesterolü**	130 ≤ normal ≤ 230 231 ≤ yüksek ve yüksek≤129
FastingBS	**Açlık Kan Şekeri**	0 ≤ düşük ≤ 69 70 ≤ normal ≤ 100 101 ≤ yüksek
RestingECG	**İstirahat EKG Sonuçları**	Normal-1, ST-2, LVH-3
MaxHR	**Maksimum Kalp Atış Hızı**	0 ≤ düşük ≤ 100, 101 ≤ normal ≤ 200, 201 ≤ yüksek
ExerciseAngina	**Egzersize Bağlı Anjina**	Y = 1 N = 0
Oldpeak	**Depresyonda Ölçülen Değer**	0 ≤ düşük ≤ 0.5, 0.6 ≤ normal ≤ 1.5, 1.6 ≤ yüksek ve yüksek<0
ST_Slope	**ST Eğiminin Yönü**	Up(yukarı)-1, Flat(yatay)-2, Down(aşağı)-3
HeartDisease	**Hastalık Sınıfı**	1 = kalp hastası 0 = sağlıklı

Veri analizi ve düzenleme sürecinde, yalnızca Excel' e değil, Weka' nın ön işleme araçlarından faydalanılarak adımlar atılmıştır. İlk olarak, eksik verilerin belirlenmesi ve doldurulması için filtreler kullanılarak öznitelikler incelenmiştir. Ardından, veriler,

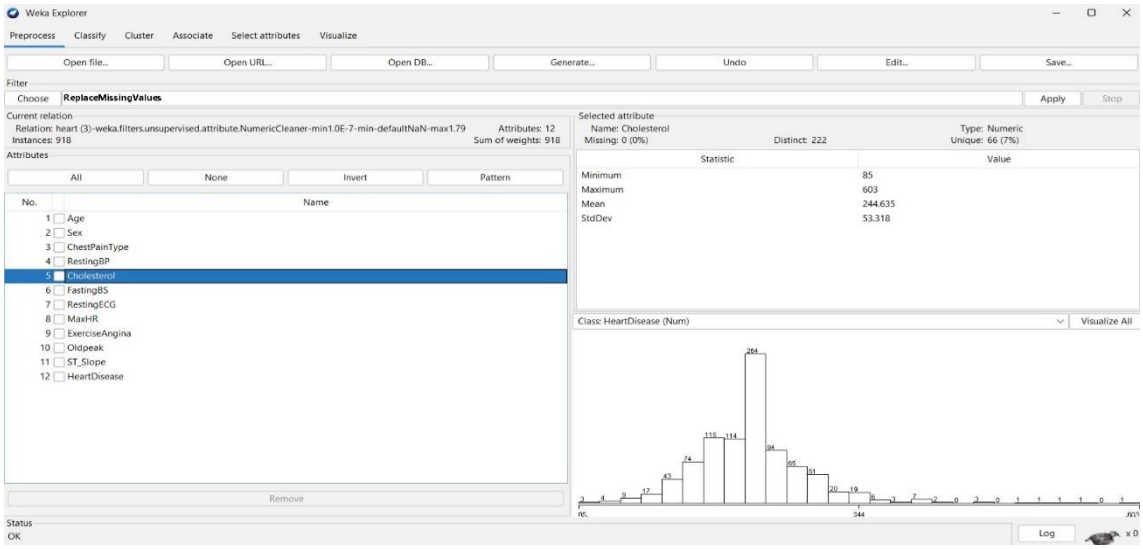
kullanılacak makine öğrenmesi algoritmalarına uygun şekilde ön işleme adımlarından geçirilerek nümerik ve nominal değerlere göre sınıflandırılmıştır. Bu yöntem, veri setinin daha kapsamlı bir şekilde değerlendirilmesini ve analiz edilmesini sağlamıştır.

RestingBP ve Cholesterol özniteliklerinde minimum değerler "0" olarak verildiği için bu değerler eksik değer olarak kabul edilir. Weka' nın ön işleme kısmı kullanılarak bu "0" değerlerinin veri setinden çıkarılması amaçlanır. Sıfır değerlerini kaldırmak için, NumericCleaner fonksiyonunun parametrelerini düzenleyip onaylamak gerekir. Bu durum, Şekil 4.4' te gösterilmiştir.



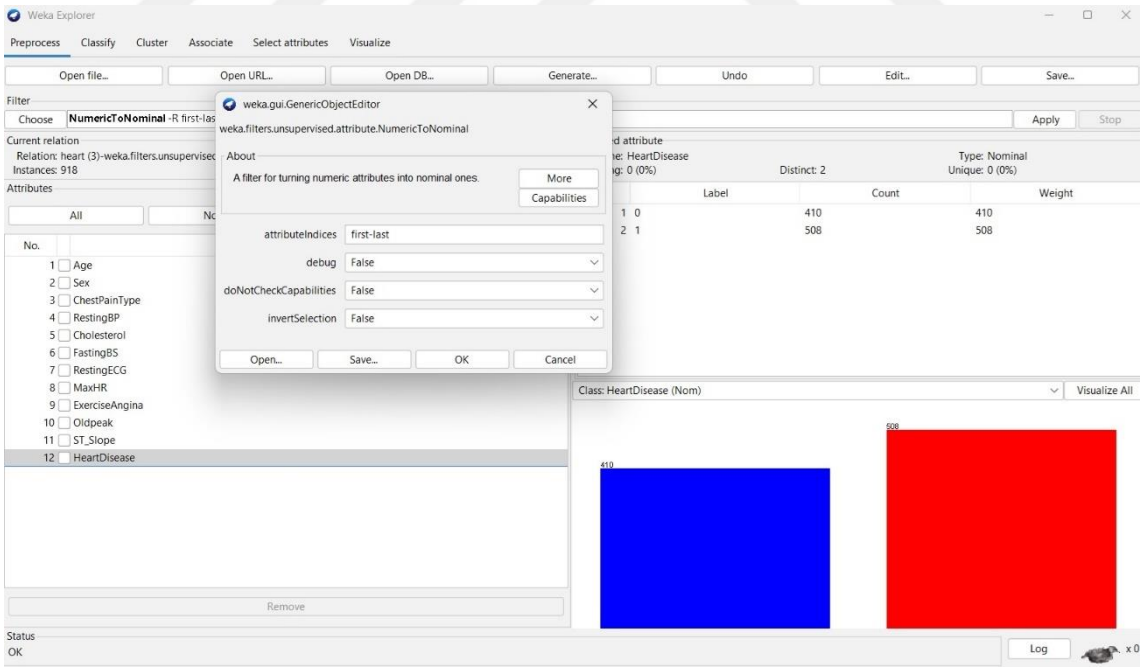
Şekil 4.4. Weka' da veri setinde eksik olan verilerin tespiti

Eksik olan veriler, ReplaceMissingValues fonksiyonu kullanılarak doldurulmuştur. Bu işlem, Şekil 4.5 'te gösterilmiştir.



Şekil 4.5. Weka' da veri setinde eksik olan verilerin doldurulması

Veri setinde çalıştırılacak algoritmaya göre veriler nümerik ya da nominal olarak sınıflandırılmış olup nominal olarak düzenlenen örnek veri seti Şekil 4.6' da gösterilmiştir.



Şekil 4.6. Weka' da nümerik değerlerin nominal değerlere dönüştürülmesi

4.3. Weka' da Çalışılan Makine Öğrenmesi Algoritmaları

Bu bölümde, kalp hastalığı teşhisinde Weka yazılımının sunduğu makine

öğrenmesi algoritmalarının etkili bir analizini sunulmaktadır. Kullanılan veri seti, beş farklı kaynaktan toplanarak bir araya getirilmiş ve toplamda 918 örneği içermektedir. Bu veri setinde bulunan 11 farklı öznitelik, bireylerin kalp hastalığı riskini değerlendirmek için önemli bilgiler sağlamaktadır.

Çalışma kapsamında, test ve analiz işlemleri için Weka platformundan yararlanılmıştır. Weka' nın 3.8.6 sürümü kullanılarak, Lineer Regresyon, M5P,Random Forest(regresyon-sınıflandırma), ZeroR, OneR, NaiveBayes, J48, IBK, SOM, LibSVM, K-Means, X-Means, Hiyerarşik Kümeleme, Expected Maximization, Self Organizing Maps gibi toplamda 16 farklı makine öğrenmesi yöntemi kullanılmıştır. Çalışmada, eğitim ve test işlemleri için Weka' nın "Explorer" modülü tercih edilmiştir. Sınıflandırma çalışmaları, sınıflandırma ve regresyon algoritmaları için "Clasify", kümeleme algoritmaları için "Cluster" sekmesi altında gerçekleştirilmiştir. Weka sınıflandırma ekranında sol üst köşeden kullanılacak olan algoritma seçimi mümkündür. Eğitim ve test işlemlerinde kullanılacak doğrulama oranı, Test Options alanından optimum performansa göre belirlenmiştir.

Öznitelikler, kullanılan algoritma türüne göre nümerik veya nominal olarak düzenlenmiştir. Bu düzenleme, algoritmaların doğru bir şekilde çalışabilmesi için kritik öneme sahiptir. Ayrıca, algoritmaların performansının değerlendirilmesi için doğru ölçütler belirlenek karşılaştırılmıştır.

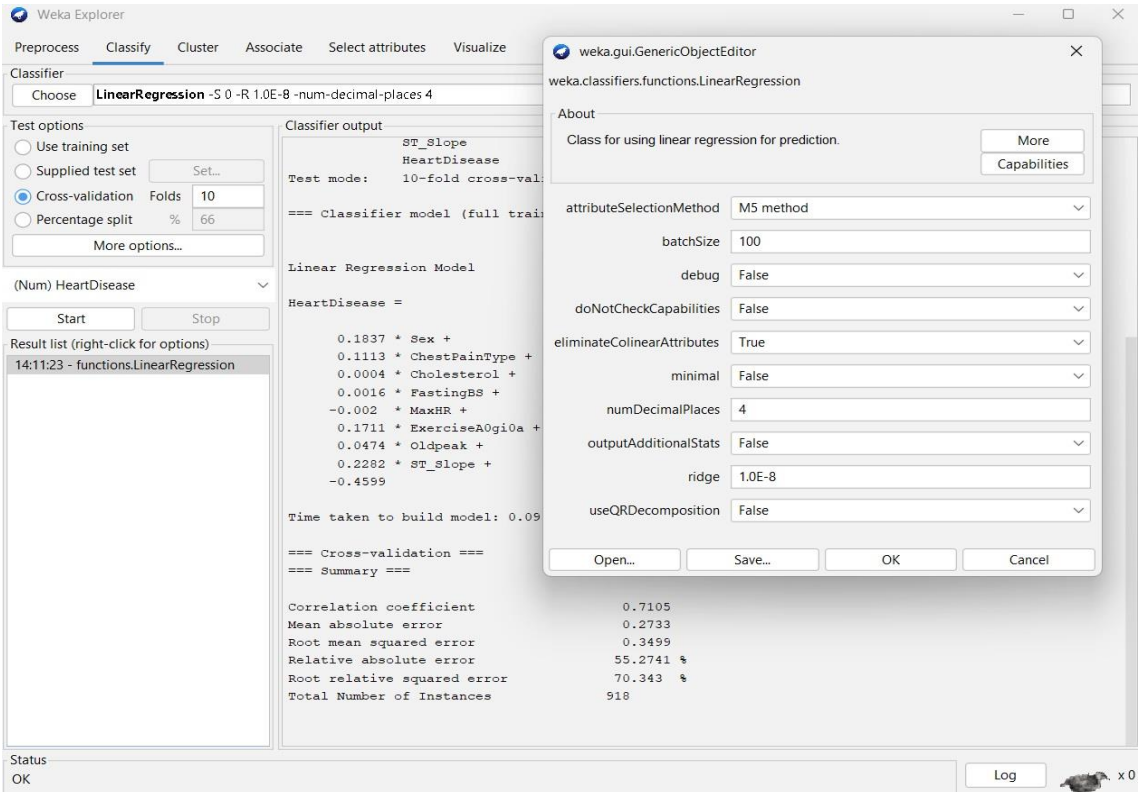
Weka platformunun kullanımıyla, algoritmaların seçimi ve performans değerlendirmesi gibi adımların, kalp hastalığı riskinin değerlendirilmesi ve hastaların erken teşhisine yönelik güçlü bir analitik temel oluşturduğu gözlenmektedir. Bu incelemeler, çalışmanın temelini oluşturmakta ve kalp hastalığı teşhisinde kullanılan makine öğrenmesi algoritmalarının etkili bir şekilde değerlendirilmesini sağlamaktadır.

4.3.1. Regresyon algoritmaları

4.3.1.1. Lineer regresyon algoritması

Weka` da Lineer Regresyon algoritması, bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiyi doğrusal bir model olarak değerlendirerek, veri setindeki bağımlı değişkenin tahmin edilmesinde ihtiyaç duyulan değeri hesaplamak için en uygun doğrusal denklemi bulmaya çalışan istatistiksel bir modelleme tekniğidir. Bu denklem, bağımsız değişkenlerin katsayılarını içerir. Katsayılar, her bir bağımsız değişkenin

bağımlı değişken üstündeki etkisini temsil eder. Lineer Regresyon algoritması, veri setindeki gözlemler arasındaki varyansı minimize etmek için en uygun doğrusal ilişkiyi bulur. Lineer Regresyon algoritması, çoğu zaman regresyon analizlerinde, tahmin ve modelleme problemlerinde kullanılır. Özellikle, bağımlı değişkenin sürekli olduğu durumlarda etkilidir. Ancak, bağımlı değişken kategorik olduğunda ise, sınıflandırma problemleri için kullanılan öteki algoritmalar tercih edilir (Taşpınar, 2013, s.69). Şekil 4.7' de çalışmada kullanılan lineer regresyon algoritması ve Weka' daki parametreleri görülmektedir.

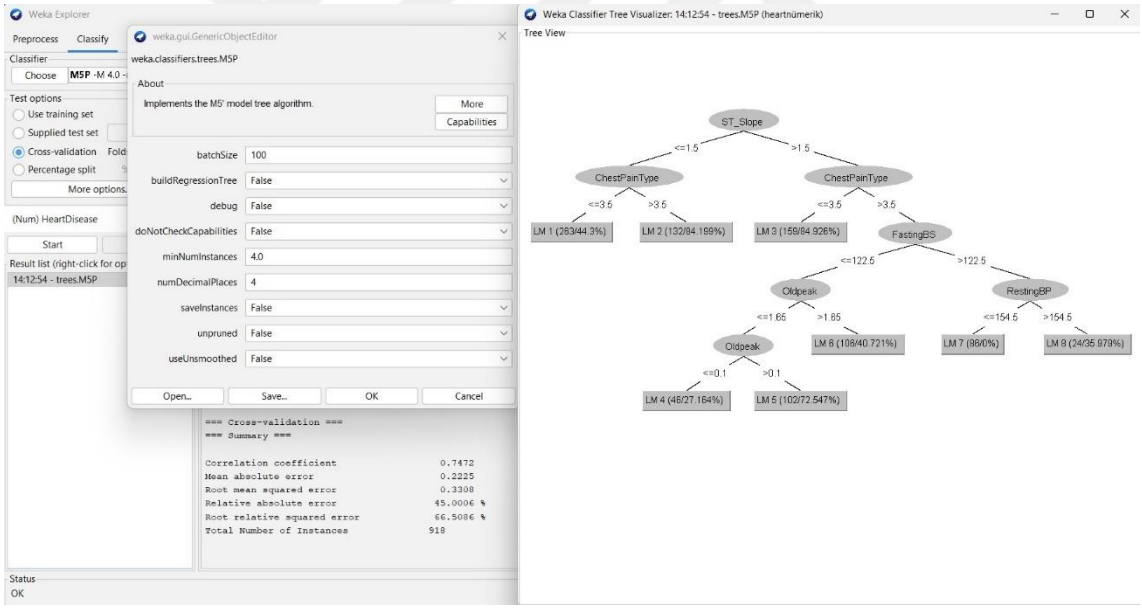


Şekil 4.7. Lineer regresyon algoritması ve Weka' daki parametreleri

4.3.1.2. M5p algoritması

M5P, bir regresyon modeli kurmak için kullanılan bir karar ağacı tabanlı bir algoritmadır. M5P algoritması, veri setindeki öznitelikler arasındaki ilişkileri modellemek için bir karar ağacı yapısı oluşturur. Karar ağacı, bir kök düğümden başlayarak bir takım iç içe geçmiş karar düğümü ve yaprak düğümünden oluşur. Ağaç yapısını oluştururken, her düğümde bir özniteliğin değerine bakılırsa veri kümesinin bölünmesi gerekmektedir. Bölme işlemi, veri setindeki özniteliklerin değerlerine

bakılarak belirlenen kriterlere dayanır. Bu kriterler çoğu zaman veri setindeki değişkenliği en çok düzeyde azaltacak halde seçilir. Ağaç yapısının yaprak düğümleri, bir regresyon modeli kurmak için kullanılır. Her yaprak düğümünde, veri kümesinin bir alt kümesi için bir regresyon modeli oluşturulur. Bu regresyon modelleri, yaprak düğümündeki verilerin değerlerine bakılarak tahminler yapmak için kullanılır. Bu durum ise, veri setindeki değişen ilişkileri daha iyi modellemek için algoritmanın parametrelerinin optimize edilmesini içerir. Oluşturulan ağaç yapısı kullanılarak, yeni veri noktaları için tahminler yapılabilir. Yeni bir veri noktası, ağaç yapısında bir takım karar düğümünden geçirilir ve netice olarak bir tahmin elde edilir. M5P algoritması, karar ağacı tabanlı bir regresyon modeli kurmak için kullanılan etkili bir yöntemdir. Veri setindeki öznitelikler arasındaki ilişkileri idrak etmek ve tahminler yapmak için kullanılabilir (Taşpınar, 2013, s.69). Şekil 4.8' de bu çalışmada kullanılan M5P algoritması ve Weka' daki parametreleri görülmektedir.

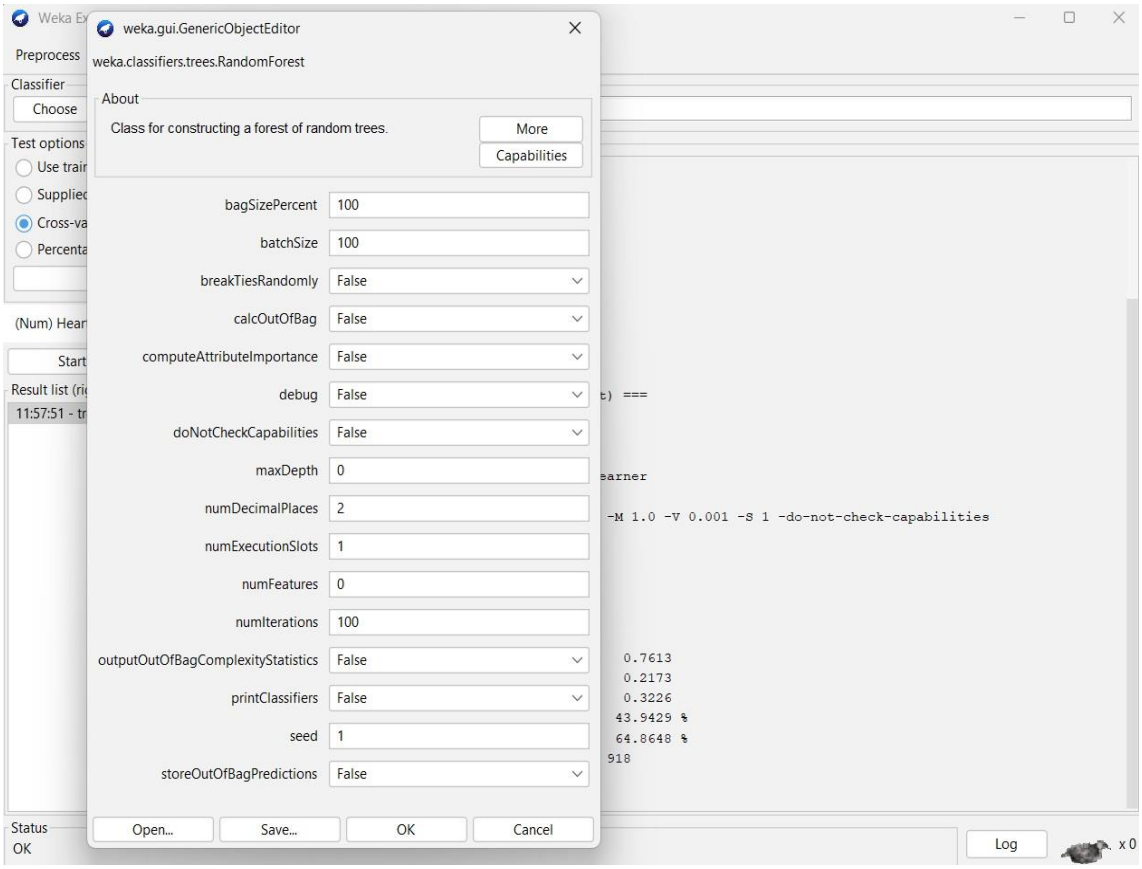


Şekil 4.8. M5P algoritması ve Weka' daki parametreleri

4.3.1.3. Random forest algoritması

Random Forest, Weka` da bulunan bir makine öğrenmesi algoritmasıdır ve çoğu zaman sınıflandırma ve regresyon problemleri için kullanılır. Random Forest algoritması, birden fazla karar ağacının birleşmesiyle oluşturduğu bir ormanı kullanır ve bu orman üstünden tahminler yaparak iyi bir genelleme performansı sunar. Ayrıca, her bir karar

ağacının rastgele seçilmiş alt kümelere dayalı olarak oluşturulması, overfitting' i (aşırı uydurma) önlemeye destek olabilir. Bu nedenle, Random Forest algoritması geniş bir uygulama yelpazesine sahiptir (Taşpınar, 2013i s.69). Şekil 4.9' da nümerik veri seti tercih edilerek çalışılan Random Forest algoritması ve Weka' daki parametreleri görülmektedir.

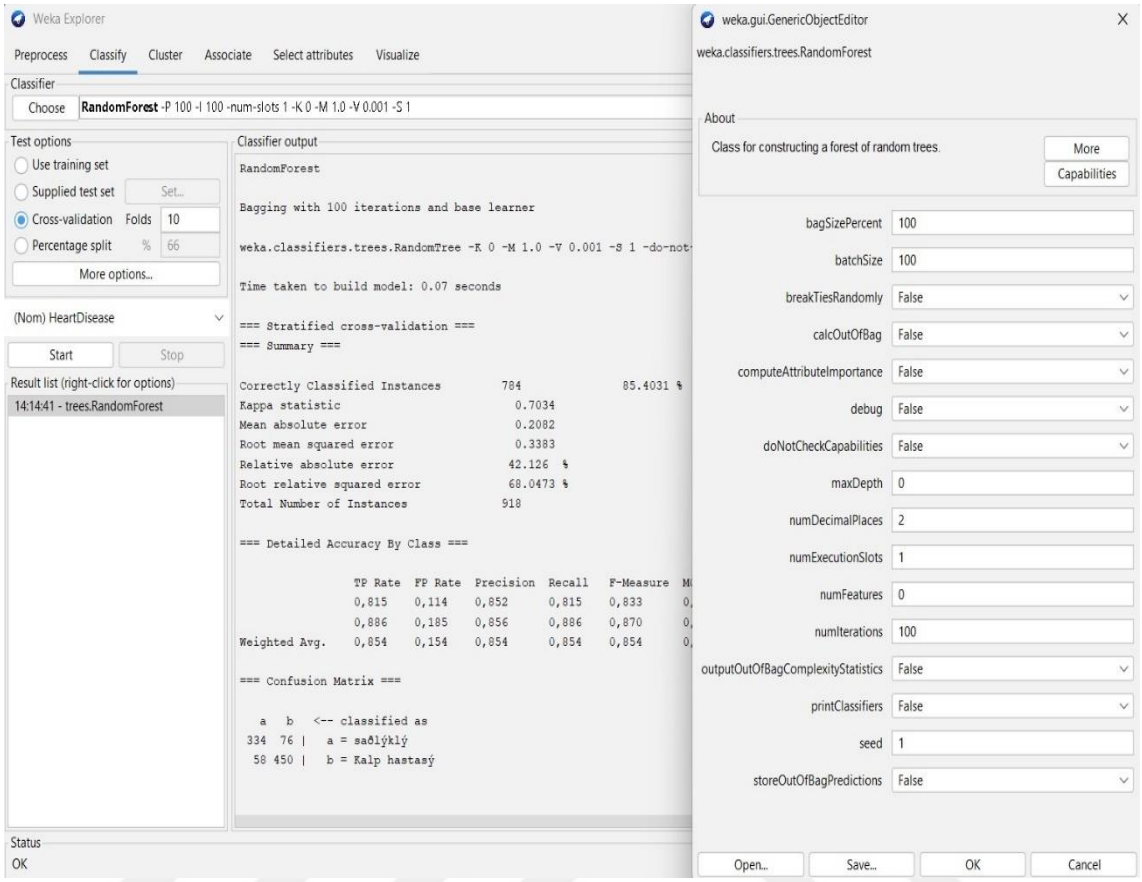


Şekil 4.9. Random forest algoritması (regresyon) ve Weka' daki parametreleri

4.3.2. Sınıflandırma algoritmaları

4.3.2.1. Random forest algoritması

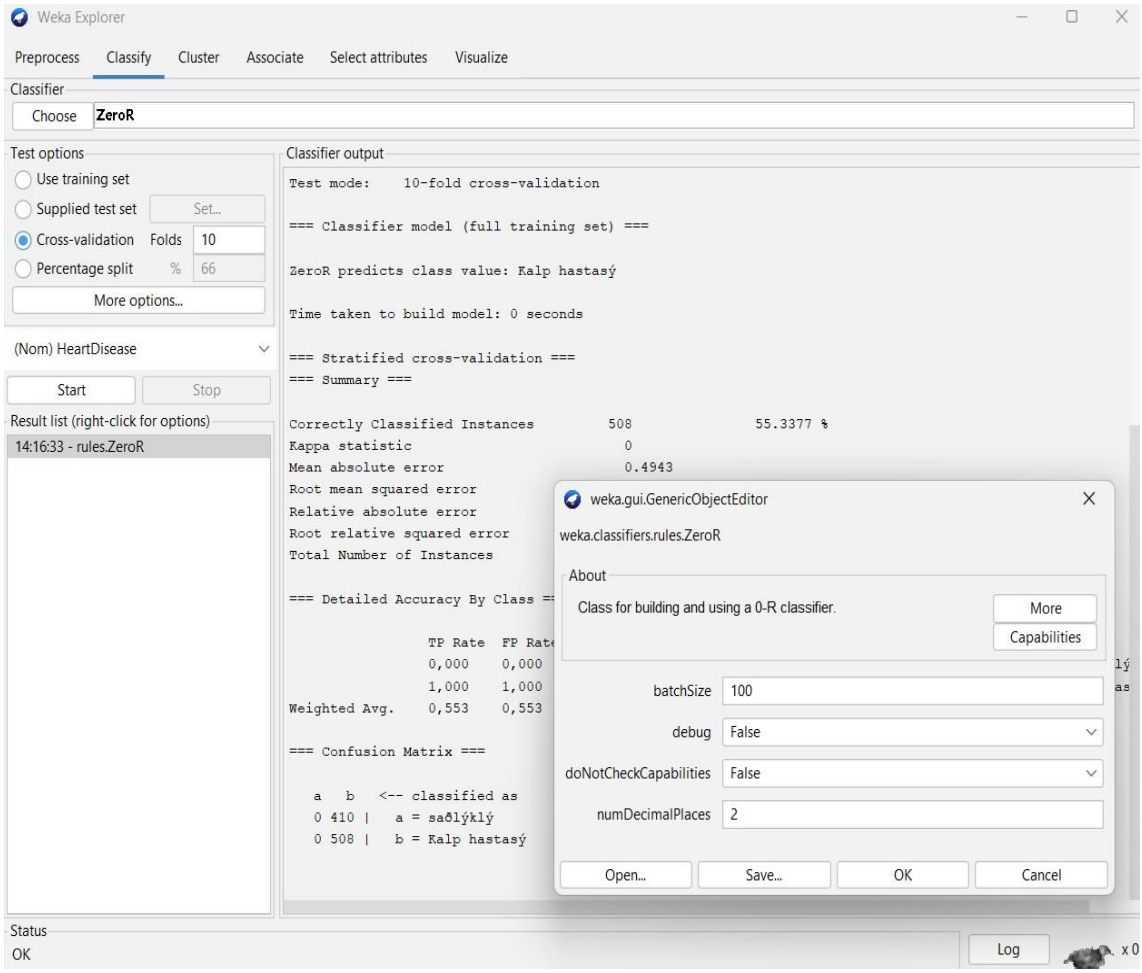
Random Forest algoritması bu veri setinde regresyon algoritması haricinde sınıflandırma algoritması olarak kullanılmıştır. Sınıflandırma algoritması şeklinde kullanıldığından veri seti nominal olarak düzenlenmiştir. Şekil 4.10' da nominal veri seti tercih edilerek çalışılan Random Forest algoritması ve Weka' daki parametreleri görülmektedir.



Şekil 4.10. Random forest algoritması (sınıflandırma) ve Weka' daki parametreleri

4.3.2.2. ZeroR algoritması

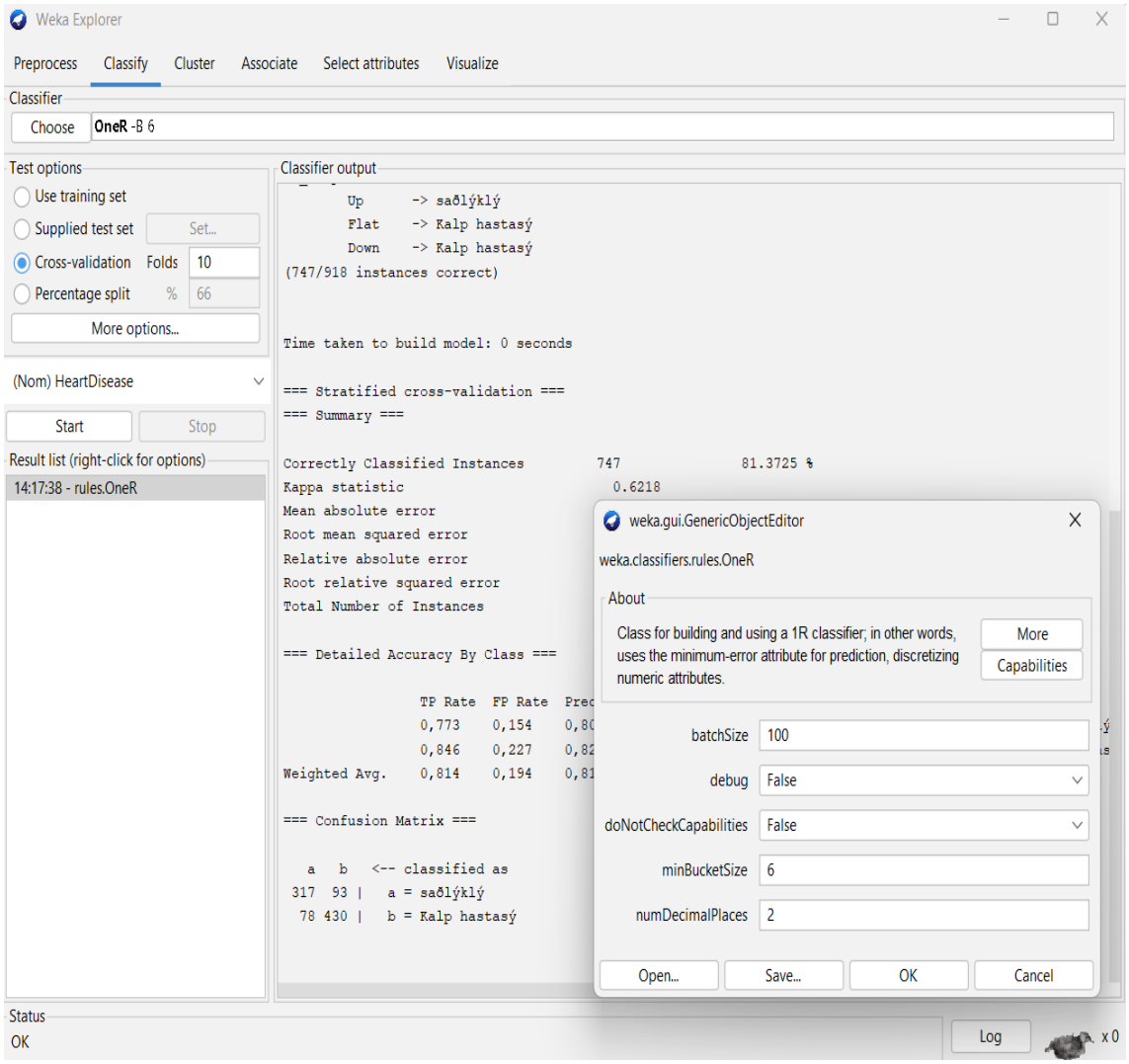
ZeroR algoritması, veri setindeki bağımlı değişkenin en yaygın sınıfını tahmin ederek çalışır. Bu tahmin, veri setindeki sınıfların dağılımına bakılarak belirlenir. İlk adımda, veri setindeki sınıfların sayısı ve dağılımı çözümlenir. En yaygın derslik, kısaca veri setinde en fazla bulunan derslik belirlenir. ZeroR algoritması, herhangi bir girdiye bağlı olmaksızın en yaygın sınıfı tahmin eder. Yani, algoritma herhangi bir özniteliğe ya da örüntüye dayalı olarak tahmin yapmaz; ancak en yaygın sınıfı döndürür. Algoritma, bütün girdiler için en yaygın sınıfı tahmin eder. Bu tahminler, sınıflandırma sorununun kolay bir çözümü olarak kabul edilir. Şekil 4.11' de bu çalışmada kullanılan ZeroR algoritması ve Weka' daki parametreleri görülmektedir.



Şekil 4.11. ZeroR algoritması ve Weka' daki parametreleri

4.3.2.3. OneR algoritması

OneR, sınıflandırma problemlerinin çözümünde kullanılır. Bu algoritma, veri setindeki her özniteliği bir bir inceleyerek en iyi kuralı belirler. OneR algoritması, basit bir kural tabanlı sınıflandırma yöntemidir, ancak birçok durumda etkin sonuçlar verebilir. Veri setinin karmaşıklığına ve öznitelikler arasındaki ilişkilere bağlı olarak, daha farklı sınıflandırma algoritmaları tercih edilebilir. Şekil 4.12' de bu çalışmada kullanılan OneR algoritması ve Weka' daki parametreleri görülmektedir.



Şekil 4.12. OneR algoritması ve Weka' daki parametreleri

4.3.2.4. NaiveBayes algoritması

NaiveBayes algoritması, bir veri örneğinin belirli bir sınıfa ait olma olasılığını hesaplamak için kullanılır. Bayes teoremi için uygulanan formül Eşitlik 4.1' de verilmiştir.

$$P(C\backslash X) = \frac{P(X \setminus C) \cdot P(C)}{P(X)} \qquad (4.1)$$

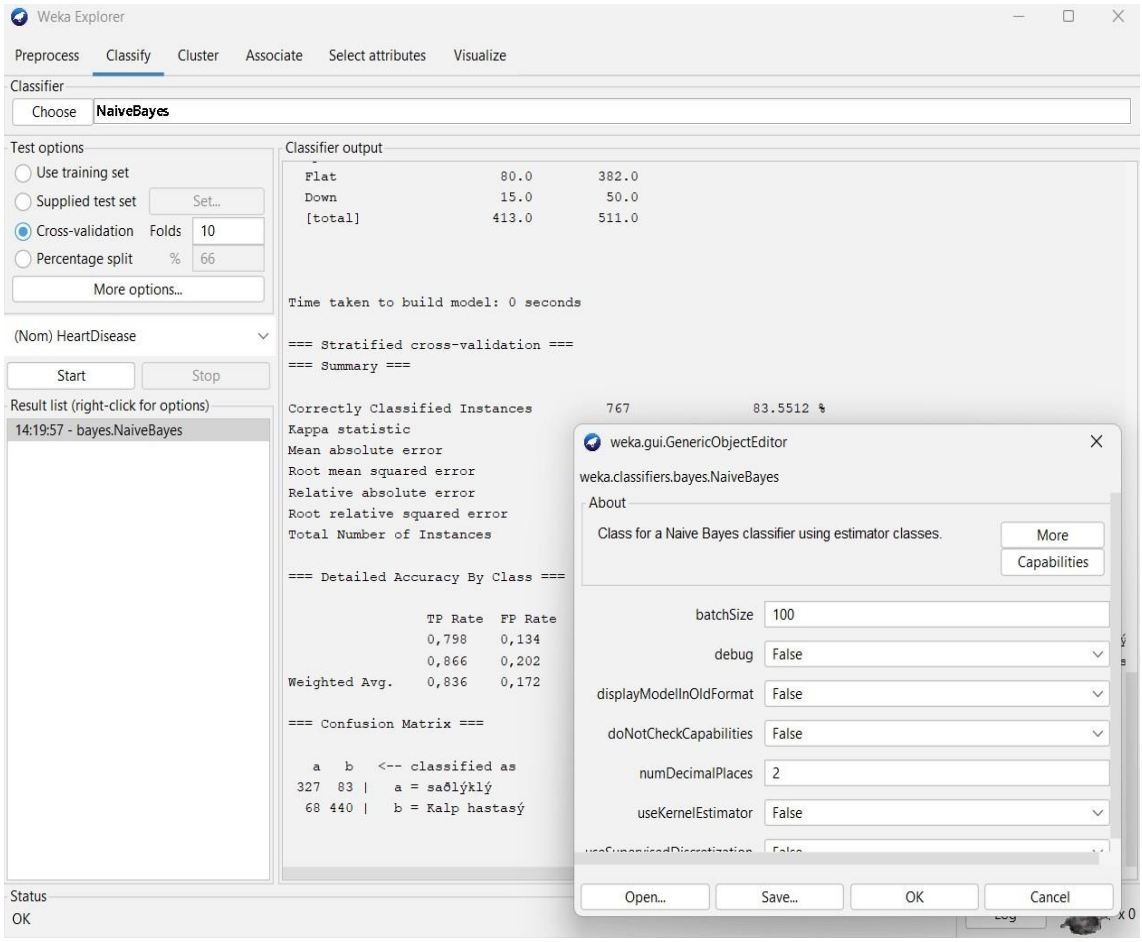
P(C\X): Örnek veri X olduğunda sınıf C' ye ait olma olasılığını temsil eder.

P(X\C): Örnek veri C olduğunda sınıf X' e ait olma olasılığını temsil eder.

P(C): Sınıf C' nin genel olasılığını temsil eder.

P(X): Örnek veri X' in genel olasılığını temsil eder.

Şekil 4.13' te bu çalışmada kullanılan NaiveBayes algoritması ve Weka' daki parametreleri görülmektedir.

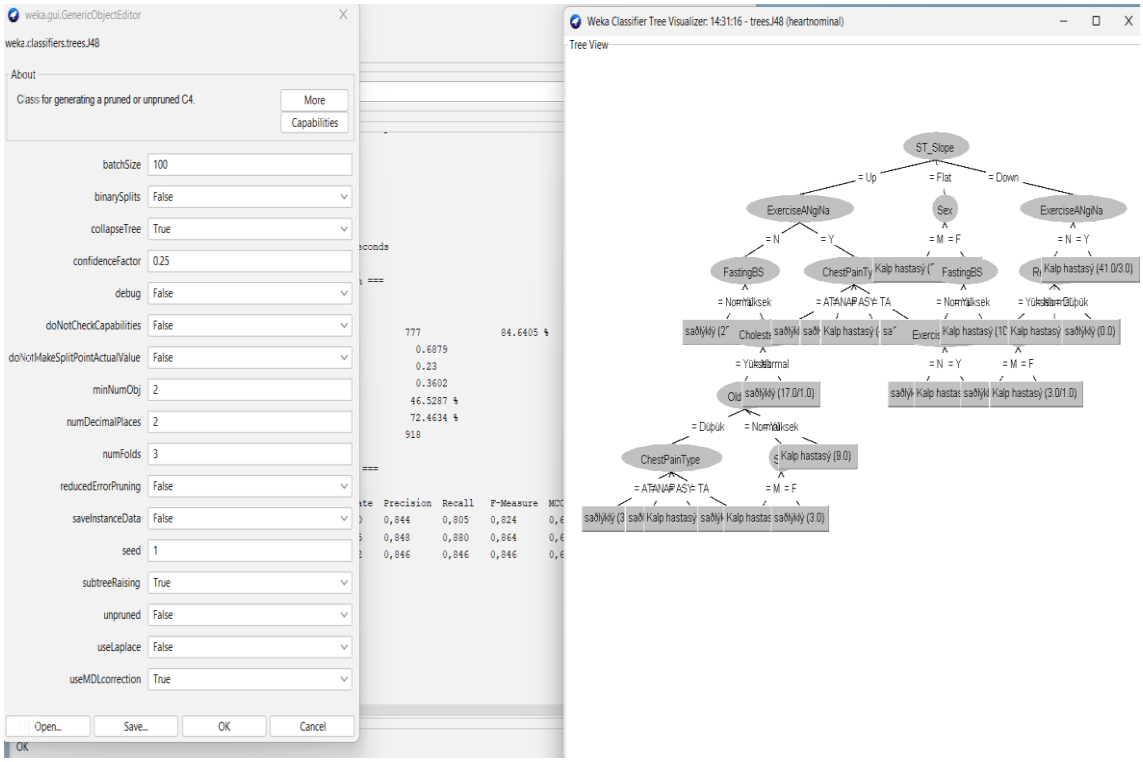


Şekil 4.13. NaiveBayes algoritması ve Weka' daki parametreleri

4.3.2.5. J48 algoritması

J48, sınıflandırma problemlerinin çözümünde kullanılan bir karar ağacı algoritmasıdır. J48, veri setindeki özelliklerin değerlerine göre bir karar ağacı oluşturur. Bu karar ağacı, veri setinin yapısını çözümleyerek her bir düğümdeki en iyi özniteliği seçer ve bu özniteliğe göre veriyi sınıflandırır. J48 algoritmasının, karar ağacını oluştururken yaptığı öznitelik tarzı, veri setindeki özniteliklerin önem derecelerini belirlemek için kullanılır. Karar ağacı oluşturmak işlemi, veri setindeki her bir öznitelik bir bir incelenir. Her düğümde, veri setindeki özniteliğe göre bölünmesiyle elde edilmiş alt kümeler içinde en uygun bölünmeyi seçer. Bu işlem, veri setindeki bütün öznitelikler tüketilene kadar tekrarlanır.

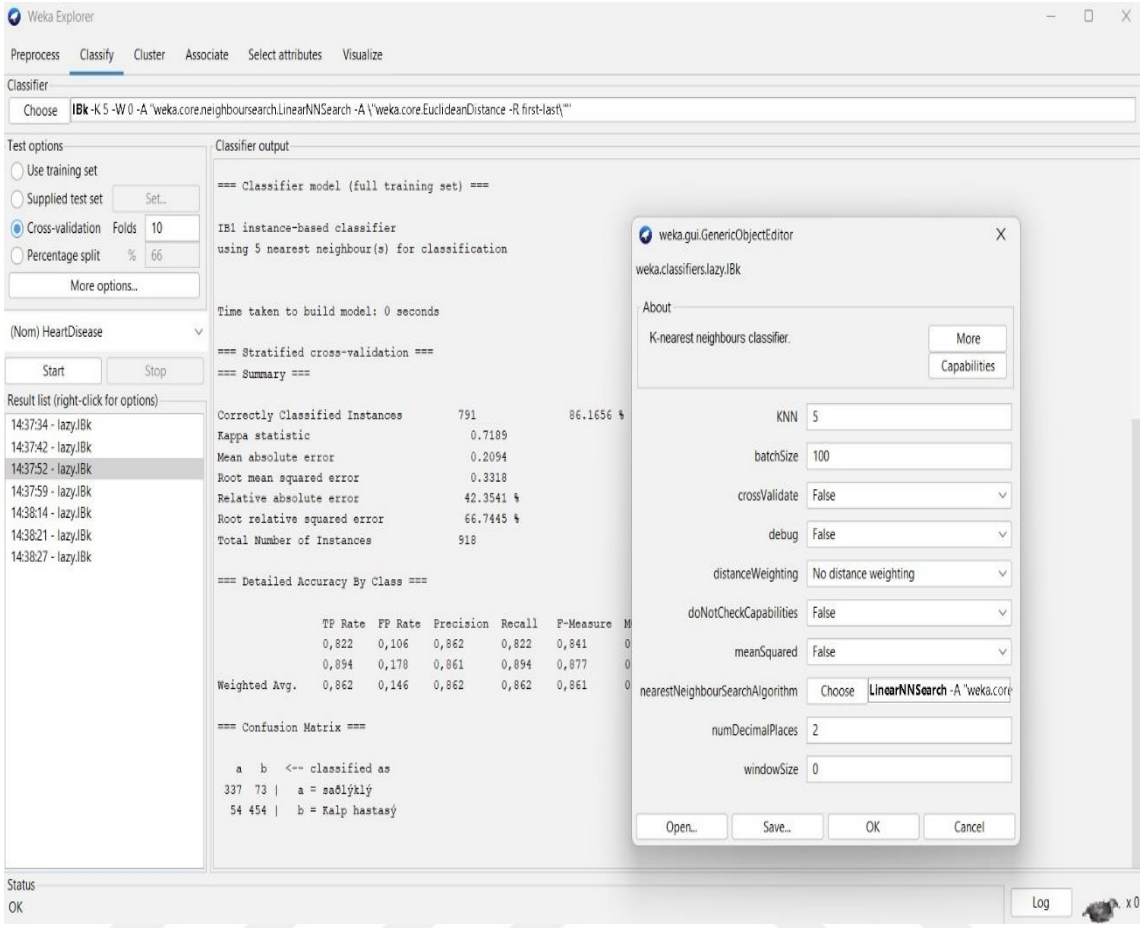
Sonuç olarak, J48 algoritması veri setindeki özelliklere göre bir karar ağacı oluşturur. Bu karar ağacı, sınıflandırma işlemlerinde yeni veri örneklerini sınıflandırmak için kullanılabilir. Karar ağacı, her bir özniteliğin değerine göre veriyi sınıflandırarak neticeleri açıklar. Bu sayede, J48 algoritması veri setinin yapısını algılamak ve sınıflandırmak için etkili bir metot sağlar. Şekil 4.14' te bu çalışmada kullanılan J48 algoritması ve Weka' daki parametreleri görülmektedir.



Şekil 4.14. J48 algoritması ve Weka' daki parametreleri

4.3.2.6. IBK algoritması

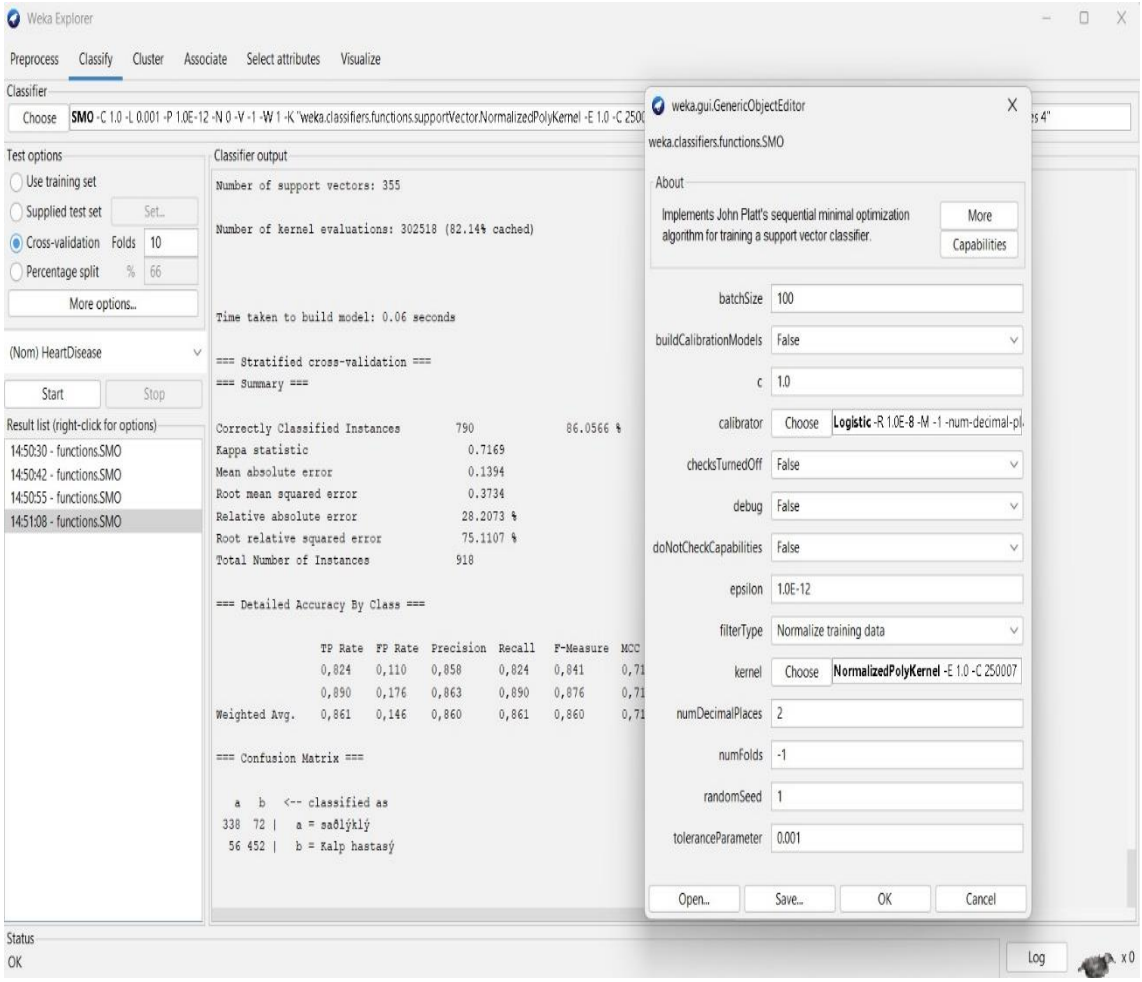
IBK algoritması, sınıflandırma problemlerini çözmek için kullanılan bir k-En Yakın Komşu algoritmasıdır. IBK algoritması, sınıflandırma problemleri için oldukça basit ve etkilidir. Ancak, k değerinin seçimi, algoritmanın performansını olumlu ya da olumsuz etkileyebilir. Küçük k değerleri, modele daha fazla esneklik katar ve eğitim verisine daha fazla uyum sağlar, ancak aşırı uydurma riskini artırabilir. Büyük k değerleri ise daha genelleştirici modeller sağlar, ancak aşırı basitleştirme riski taşır. Dolayısıyla, k değerinin doğru seçilmesi, algoritmanın başarımını önemli ölçüde etkiler. Şekil 4.15' te bu çalışmada kullanılan IBK algoritması ve Weka' daki parametreleri görülmektedir.



Şekil 4.15. IBK algoritması ve Weka' daki parametreleri

4.3.2.7. SMO algoritması

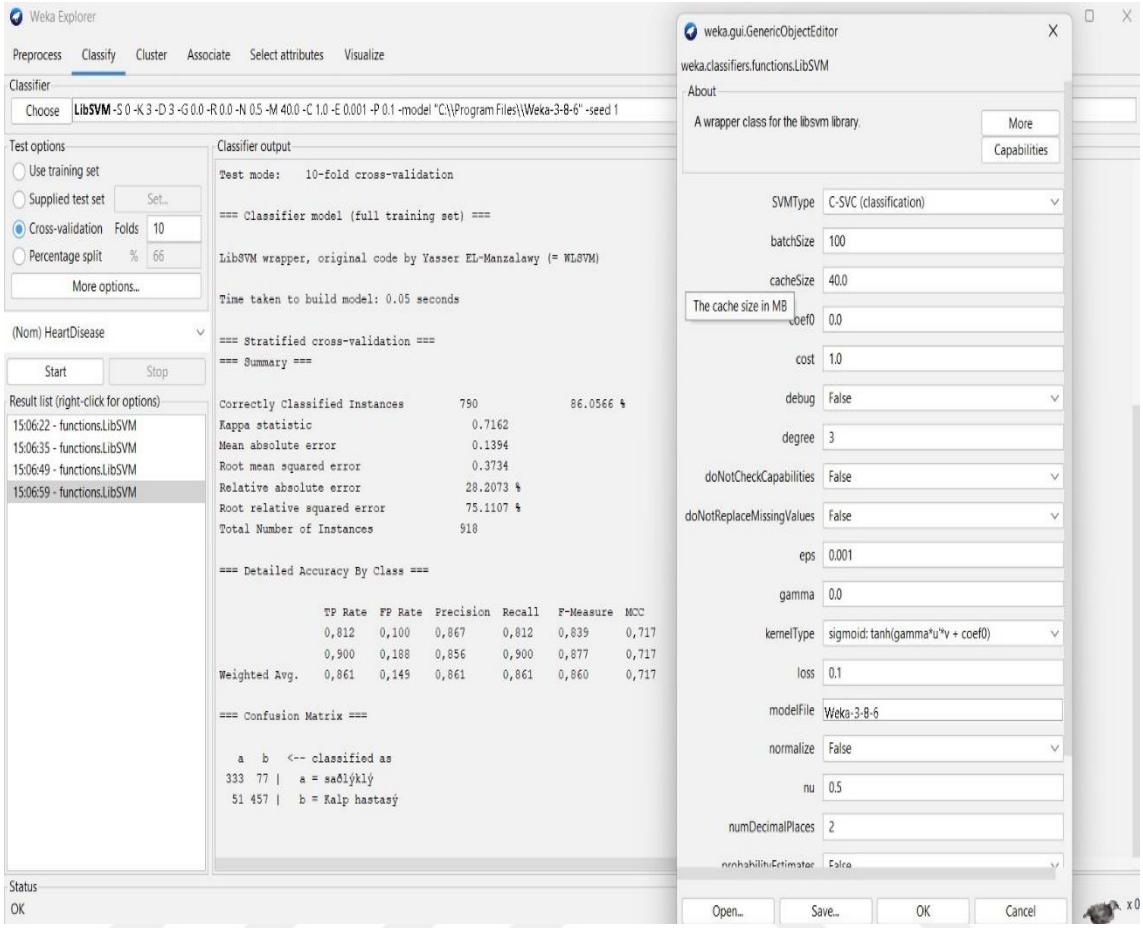
Sequential Minimal Optimization (SMO), destek vektör makineleri (SVM) için geliştirilmiş bir optimizasyon algoritmasıdır. SMO algoritması, SVM için çekirdek fonksiyonlarını (kernel functions) kullanır. Çekirdek fonksiyonları, veri noktalarını yüksek boyutlu özellik uzayına dönüştürerek, doğrusal olarak ayrılamayan veri kümesini doğrusal olarak ayrılabilir hale getirir. Bu, SVM' nin kompleks sınıflandırma sorunlarını çözebilmesini sağlar. Ancak, doğru çekirdek tipini seçmek, modelin performansını ve genelleme kabiliyetini etkileyebilir, bundan dolayı bu algoritmada doğru çekirdek tipinin seçilmesi önemlidir. Bu veri setinde uygulanan çekirdek tipleri içinde optimum performansın "NormalyzedPolyKernel" için olduğu görülmüştür. Şekil 4.16' da bu çalışmada kullanılan SMO algoritması ve Weka' daki parametreleri görülmektedir.



Şekil 4.16. SMO algoritması ve Weka' daki parametreleri

4.3.2.8. LibSVM algoritması

LibSVM (Library for Support Vector Machines), SVM için bir kütüphanedir ve sınıflandırma, regresyon gibi birçok makine öğrenmesi görevini gerçekleştirmek için kullanılır. LibSVM, SVM` nin kompleks sınıflandırma problemlerinin çözümü için etkendir. LibSVM algoritması, veri noktalarını yüksek boyutlu özellik uzayına dönüştürerek, bu uzayda doğrusal olarak ayrılabilir hale getirir. Bu dönüşüm, farklı çekirdek fonksiyonları kullanılarak gerçekleştirilir. LibSVM algoritması, yüksek boyutlu özellik uzayında kompleks sınıflandırma sorunlarını halletmek için etken bir araçtır. Doğru çekirdek fonksiyonunun seçilmesi ve modelin müsait biçimde ayarlanması, algoritmanın performansını büyük seviyede etkiler. Bu veri setinde uygulanan çekirdek tipleri içinde optimum performansın "Sigmoid" için olduğu görülmüştür. Şekil 4.17' de bu çalışmada kullanılan LibSVM algoritması ve Weka' daki parametreleri görülmektedir.



Şekil 4.17. LibSVM algoritması ve Weka' daki parametreleri

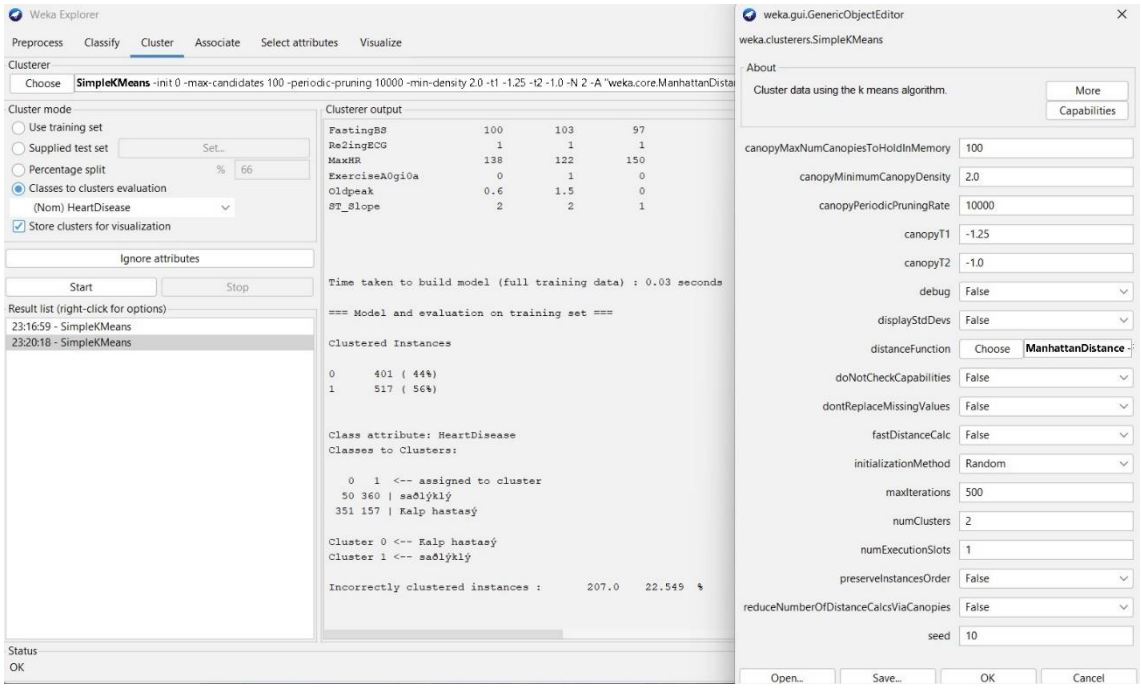
4.3.3. Kümeleme algoritmaları

4.3.3.1. K-Means algoritması

K-means algoritması, kümeleme analizinde sıkça kullanılan bir yöntemdir ve veri noktalarını belirli bir sayıda küme veya grup içinde kümelemek için kullanılır. Bu algoritma, başlangıçta belirli sayıda küme merkezi seçer, bu küme merkezleri çoğu zaman rastgele seçilir ya da veri noktalarından bazıları arasından rastgele seçilir. Her veri noktası, en yakın küme merkezine atanır. Bu atanma genellikle, Öklid mesafesi veya Manhattan mesafesi gibi bir uzaklık ölçüsü kullanılarak hesaplanır. Veri noktalarının kümeleme işlemi, belirli bir iterasyon sayısına veya küme merkezlerinin değişim miktarının belirli bir eşiğin altına düşmesine kadar devam eder. Sonlanma kriteri sağlandığında, algoritma sonlanır ve küme merkezleri ile kümeleme sonuçları elde edilir.

K-means algoritması, veri noktalarını belirli sayıda küme arasında kümelemek için yaygın olarak kullanılan bir yöntemdir. Ancak, küme sayısının önceden belirlenmesi

gerektiği ve algoritmanın başlangıç merkezlerine duyarlı olabileceği benzer biçimde birtakım dezavantajları bulunmaktadır. Bu veri setinde uygulanan uzaklık ölçüsüne göre optimum performansın "Manhattan" için olduğu görülmüştür (Jain ve Dubes, 1988, s.68). Şekil 4.18' de bu çalışmada kullanılan K-Means algoritması ve Weka' daki parametreleri görülmektedir.

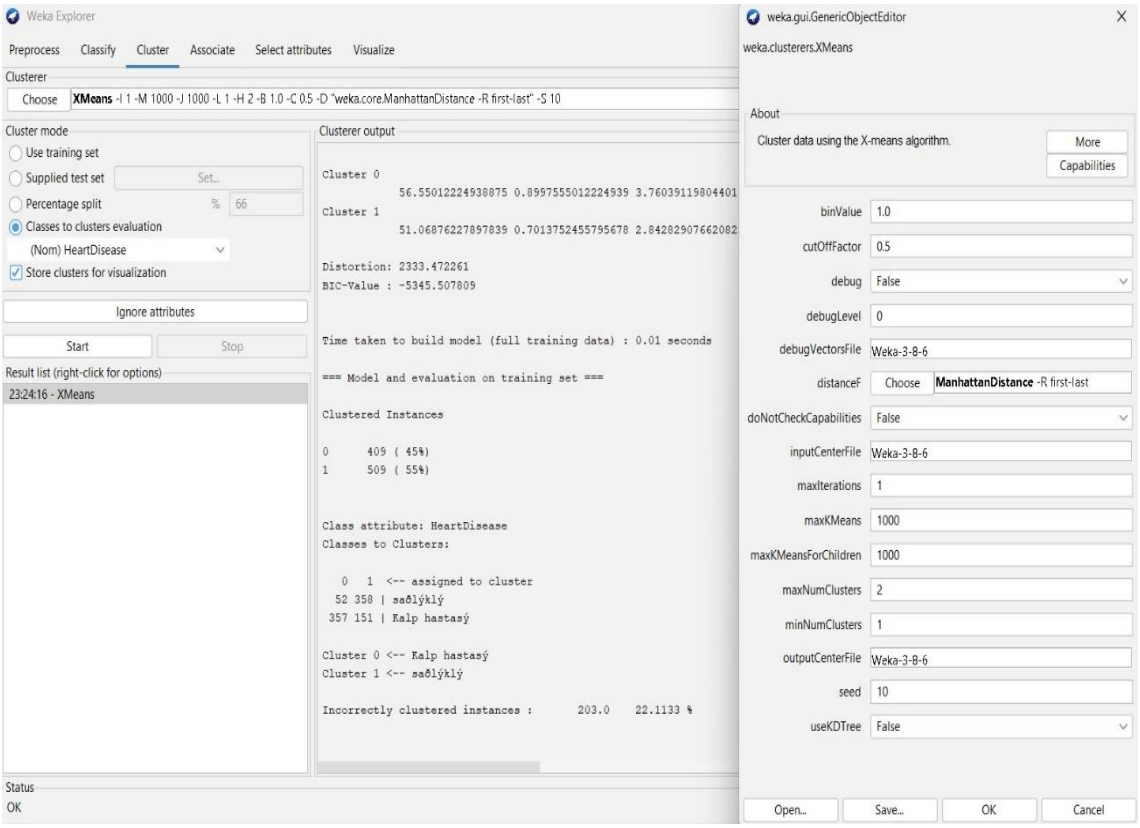


Şekil 4.18. K-Means algoritması ve Weka' daki parametreleri

4.3.3.2. X-Means algoritması

X-Means algoritması, K-Means algoritmasının genelleştirilmiş bir versiyonudur ve küme sayısını otomatik belirler. Bu algoritma, K-Means algoritmasının her iterasyonunda küme sayısını artırarak ve azaltarak çalışır, böylece veri setinin daha iyi bir modellemesini sağlar. Başlangıçta belirli sayıda küme merkezi seçilir. Her veri noktası, en yakın küme merkezine atanır. Bu atanma çoğu zaman, Öklid mesafesi ya da Manhattan mesafesi şeklinde bir uzaklık ölçüsü kullanılarak hesaplanır. X-means algoritması çoğu zaman küme merkezlerinin değişiklik miktarının belirli bir eşiğin altına düşmesi ya da belirli bir iterasyon sayısına ulaşılması şeklinde durumlarda sonlanır. X-Means algoritması, k-Means algoritmasının otomatik küme sayısı atama kabiliyetinden daha rahat bir yaklaşım sunar. Bu sayede, veri setlerindeki yapısal karmaşıklığı daha iyi modelleyebilir ve daha doğru kümeleme neticeleri elde edilir. Bu veri setinde uygulanan

uzaklık ölçüsüne göre optimum performansın “Manhattan” için olduğu görülmüştür (Jain ve Dubes, 1988, s.68). Şekil 4.19’ da bu çalışmada kullanılan X-Means algoritması ve Weka’ daki parametreleri görülmektedir.

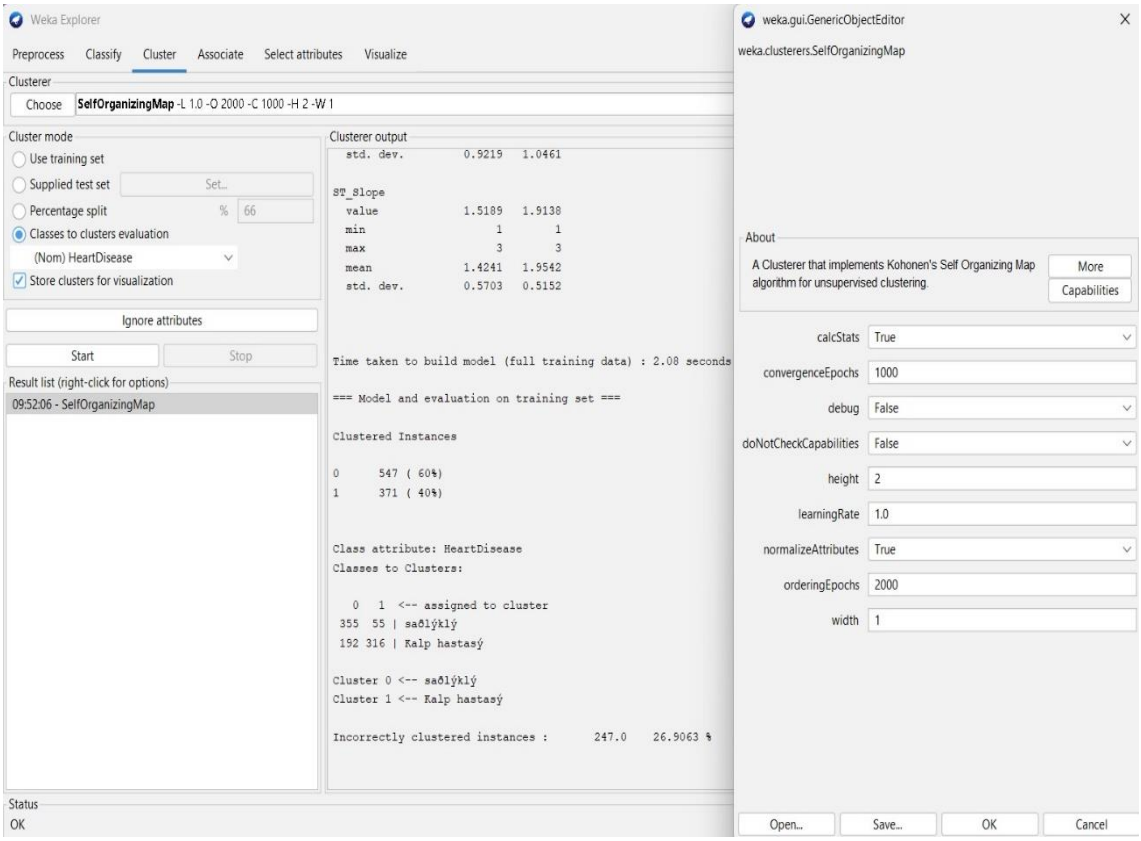


Şekil 4.19. X-Means algoritması ve Weka’ daki parametreleri

4.3.3.3. Self-Organizing Maps algoritması

Self-Organizing Maps (SOM) denetimsiz öğrenme algoritmalarından biridir ve temelde veri setlerindeki yapıyı ortaya çıkarmak için kullanılır. SOM, özellikle çok boyutlu veri setlerinin görselleştirilmesi ve kümeleme için etkili bir araçtır. SOM algoritması, genellikle bir iki boyutlu bir ızgara olarak düşünülen bir ağ oluşturur. Bu ızgara, düğümler veya nöronlar denilen hücrelerden oluşur. Her düğüm, bir vektör olarak temsil edilen bir ağırlık vektörüne sahiptir. Başlangıçta, her düğümün ağırlık vektörleri genellikle rastgele belirlenir. Ağırlık vektörleri, veri noktalarının boyutuna uygun uzunluktadır ve genellikle 0 ile 1 arasında rastgele değerler alır. Her veri noktası, ağırlık vektörleri arasındaki mesafe kullanılarak en yakın düğüme atanır. Her veri noktası için en yakın düğüm, kazanan düğüm olarak belirlenir. Bu düğüm, veri noktasına en yakın olan ve ona atanmış olan düğümdür. Bu süreç, belirli bir iterasyon sayısına ya da belirli

bir durum gerçekleşene kadar tekrarlanır. SOM, veri setlerindeki yapıyı görselleştirmek ve analiz etmek için yaygın olarak kullanılan bir yöntemdir. Özellikle boyutsal azaltma ve veri keşfi alanlarında önemli bir araçtır (Jain ve Dubes, 1988, s.68). Şekil 4.20' de bu çalışmada kullanılan SOM algoritması ve Weka' daki parametreleri görülmektedir.

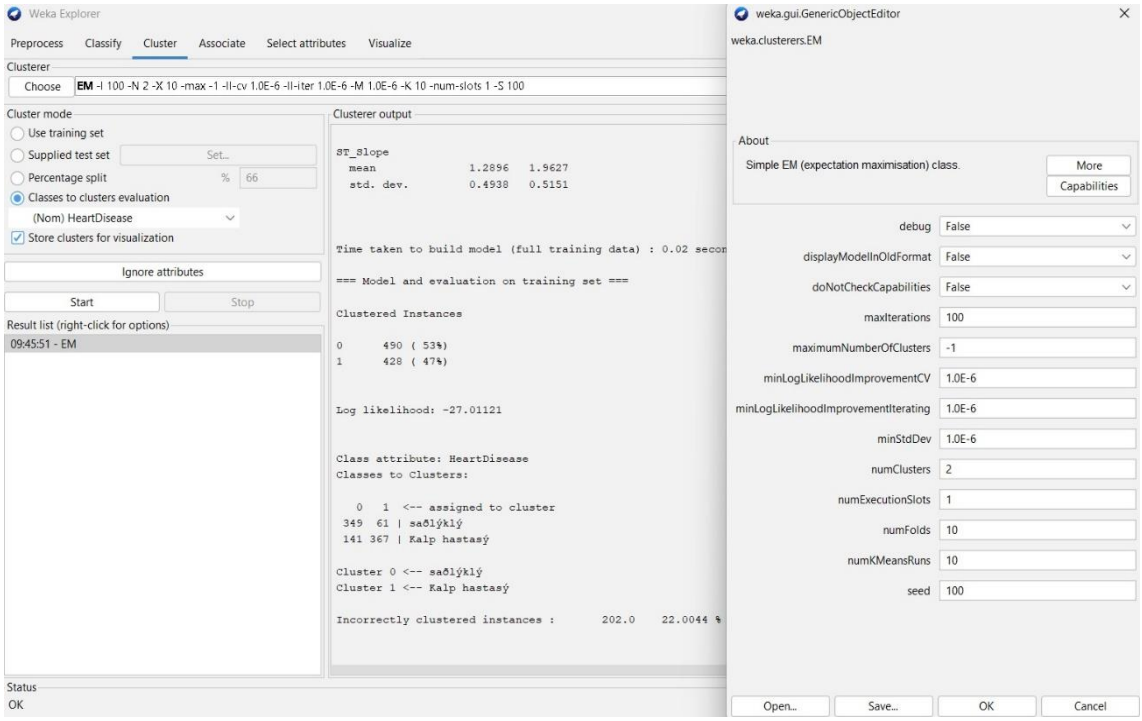


Şekil 4.20. SOM algoritması ve Weka' daki parametreleri

4.3.3.4. EM algoritması

EM algoritması başlangıçta modelin parametrelerini rastgele seçer veya kullanıcı tarafından belirlenir. Beklenti adımında, eksik verilerin veya gizli değişkenlerin olasılık dağılımları tahmin edilir. Bu adımda, eksik verilere veya gizli değişkenlere ilişkin olasılık dağılımları, mevcut parametre tahminleri ve veri noktaları kullanılarak hesaplanır. Maksimizasyon adımında, modelin parametreleri, beklenen değerler kullanılarak güncellenir. Bu adımda, parametrelerin maksimum olabilirlik tahminleri elde edilir. Parametreler, mevcut veri ve beklenen değerler kullanılarak yeniden tahmin edilir. Beklenti ve maksimizasyon adımları, belirli bir kriter karşılanana kadar veya belirli bir iterasyon sayısına ulaşana kadar tekrarlanır. Yakınsama, parametrelerin belirli bir

hassasiyet veya kararlılık seviyesine ulaşması durumunda gerçekleşir. EM algoritması, veri setlerinde eksik verilerin olduğu veya gizli değişkenlerin bulunduğu istatistiksel modellerin tahmin edilmesi için yaygın olarak kullanılan bir yöntemdir. Özellikle kümeleme, gizli Markov modelleri ve karmaşık karışık modeller gibi birçok alanda kullanılır. EM algoritması, genellikle parametrik istatistiksel modelleme problemlerinde kullanılan bir araçtır ve geniş bir uygulama alanına sahiptir (Jain ve Dubes, 1988, s.68). Şekil 4.21' de bu çalışmada kullanılan EM algoritması ve Weka' daki parametreleri görülmektedir.

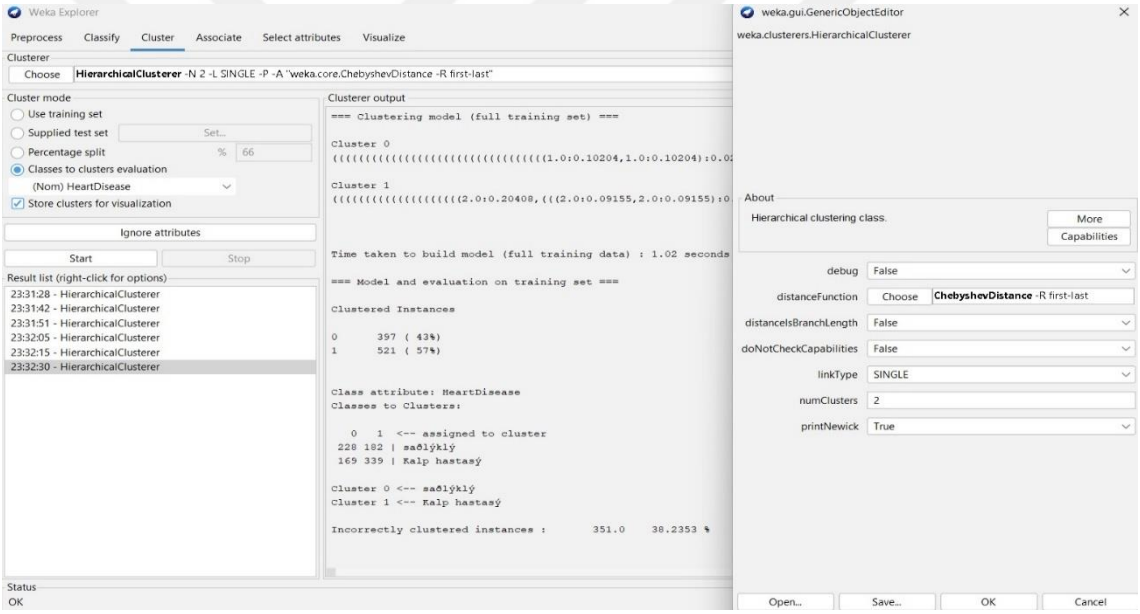


Şekil 4.21. EM algoritması ve Weka' daki parametreleri

4.3.3.5. Hiyerarşik kümeleme algoritması

Hiyerarşik kümeleme algoritmasında başlangıçta, her veri noktası bir küme olarak kabul edilir. İki kümenin birbirine ne kadar benzediğini belirlemek için Öklid, Manhattan vb. gibi bir benzerlik ölçüsü seçilir, . Başlangıçta, her veri noktası bir kümedir. Her veri noktası kendisini içeren bir küme olarak başlar. Benzerlik ölçüsü kullanılarak, en yakın iki küme birleştirilir. Birleştirme işlemi, belirlenen benzerlik ölçüsüne göre en yakın iki kümenin bir araya getirilmesiyle gerçekleşir. Bu adım, belirlenen bir kriter sağlanana kadar veya belirli bir küme sayısına ulaşılana kadar tekrarlanır. Her bir birleştirme adımıyla, yeni bir küme oluşturulur ve bu işlem hiyerarşik bir ağaç yapısını oluşturur.

Ağaç yapısı, veri noktalarının birbirleriyle olan benzerliklerine dayalı olarak kümeleme düzeylerini gösterir. Kümeleme işlemi, belirli bir kriter karşılanana kadar devam eder. Bu kriter, belirli bir küme sayısına ulaşma, belirli bir benzerlik düzeyine ulaşma veya belirli bir uzaklık eşiğine ulaşma gibi olabilir. Kümeleme işlemi sona erdikten sonra, küme sayısı seçilen bir kesme düzeyine göre belirlenir. Hiyerarşik kümeleme, veri noktalarını birbirine benzerlikleri temelinde gruplandıran ve bu grupları hiyerarşik bir yapıda sunan bir kümeleme yöntemidir. Ağaç yapısı, veri setindeki kümeleme düzeylerini görselleştirmek için kullanılır ve farklı kesme düzeyleri seçilerek farklı küme sayıları elde edilebilir (Jain ve Dubes, 1988, s.68). Bu veri setinde uygulanan uzaklık ölçüsüne göre optimum performansın "Chebyshev" için olduğu görülmüştür. Şekil 4.22' de bu çalışmada kullanılan EM algoritması ve Weka' daki parametreleri görülmektedir.



Şekil 4.22. Hiyerarşik kümeleme algoritması ve Weka' daki parametreleri

4.4. Performans Ölçütleri

4.4.1. Kappa istastiği

Kappa istatistiği, bir değerlendirici veya gözlemcinin iki kategori arasındaki uyumunu değerlendirmek için kullanılan istatistiksel bir ölçüttür. İki değişken arasındaki uyumu ölçerken, $P_r(a)$ ve $P_r(e)$ olasılıkları üzerinden Cohen'in kappa istatistiği için uygulanan formül Eşitlik 4.2' de verilmiştir.

$$K = \frac{P_r(a) - P_r(\mathrm{e})}{1 - P_r(\mathrm{e})} \qquad (4.2)$$

$P_r(a)$: İki değerlendirici arasındaki gözlemlenen uyum oranını temsil eder.

$P_r(e)$: Şansa bağlı olarak beklenen uyum oranını ifade eder.

Bu formül, Kappa istatistiğini hesaplarken gözlemlenen uyumu şansa bağlı uyumdan çıkartarak düzeltmeye imkan tanır. Elde edilen Kappa değeri -1 ile +1 arasında değişir. 0' dan küçük bir değer uyumun olmadığını, 1' e yaklaşan bir değer ise tam uyumu gösterir. Kappa değeri ne kadar yüksekse, değerlendiriciler arasındaki uyum o kadar güçlüdür.

Kappa istatistiği, aynı veri kümesini birbirinden farklı gözlemcilerin yorumlama ve değerlendirme şekillerine tabi tutulduğunda kullanışlıdır. Özellikle tıbbi teşhislerde ve toplumsal bilimlerde sınıflandırma sistemlerinin değerlendirilmesinde sıklıkla kullanılır. Bu istatistik, gözlemciler arasındaki anlaşma düzeyini daha objektif bir halde değerlendirmeye imkan sağlar (Aydemir, 2018, s.65).

4.4.2. Korelasyon katsayısı

Korelasyon katsayısı iki değişken arasındaki ilişkiyi ölçen ve ilişkinin yönü ile gücünü belirleyen bir istatistiksel ölçüttür. Değer genellikle -1 ile +1 arasında değişir ve bu değer, ilişkinin doğasını algılamak için kullanılır.

Korelasyon katsayısı +1' e yaklaşıyorsa: İki değişken içinde güçlü bir pozitif ilişki vardır.

Korelasyon katsayısı -1' e yaklaşıyorsa: İki değişken içinde güçlü bir negatif ilişki vardır.

Korelasyon katsayısı 0' a yaklaşıyorsa: İki değişken içinde bariz bir ilişki yoktur ya da ilişki çok zayıftır.

Korelasyon katsayısı için uygulanan formül Eşitlik 4.3' te verilmiştir.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\Sigma x^2 - (\sum x)^2][n\Sigma y^2 - (\sum y)^2]}} \qquad (4.3)$$

r : Korelasyon katsayısı

n : Veri noktalarının sayısı

x : Değişkenin ölçümleri

y : Diğer değişkenin ölçümleri

$\sum xy$: x ve y' nin çarpımlarının toplamı

$\sum x$: x değerlerinin toplamı

$\sum y$: y değerlerinin toplamı

Σx^2 : x değerlerinin karelerinin toplamı

Σy^2 : y değerlerinin karelerinin toplamı

Korelasyon katsayısının pozitif değerleri, iki değişken arasında doğrusal bir ilişkinin bulunduğunu gösterir. Bu durumda, bir değişkenin değeri artarken öteki değişkenin değeri de artar. Negatif değerler ise iki değişken arasında ters yönlü bir ilişkinin olduğunu gösterir; şöyle ki bir değişkenin değeri artarken diğer değişkenin değeri azalır.

Korelasyon katsayısı, veri setindeki değişkenlerin birbirleriyle olan ilişkisini algılamak ve bu ilişkiyi nicel olarak ifade etmek için önemlidir. Özellikle istatistik, ekonomi, psikoloji ve diğer toplumsal bilimlerdeki incelemelerde sıkça kullanılır. Ancak korelasyon, nedensel ilişkiyi doğrulamaz, yalnızca değişkenler arasındaki ilişkinin gücünü ve yönünü belirler (Marapelli, 2019, s.67).

4.4.3. Ortalama mutlak hata

Burada tahminde tam başarı elde edilebilmesi için, tahmin hatasının sıfıra eşit olması beklenir. Fakat bu durum gerçekte çok nadir ortaya çıkar. Ortalama mutlak hata(OMH) için uygulanan formül Eşitlik 4.4' te verilmiştir.

$$\text{OMH} = \frac{\sum_{i=1}^{n} |\hat{\theta}_i - \theta_i|}{n} \qquad (4.4)$$

Eşitlik 4.4, Eşitlik 4.5, Eşitlik 4.6 ve Eşitlik 4.7' de kullanılan terimler aşağıdaki gibidir.

n = örneklem sayısı

θ_i= i sıra numaralı gerçek talep

$\hat{\theta}_i$= i sıra numaralının tahmin edilen talebi

i = örneklem sırası

OMH, bir tahmin modelinin gerçek değerlerle ne kadar uyumlu olduğunu ölçen bir ölçüttür. OMH, tahmin edilen değerlerin gerçek değerlerden ne kadar sapma gösterdiğini belirtir ve bu sapmaların mutlak değerlerinin ortalamasıdır.

Matematiksel olarak, bir modelin OMH' si, her gözlem için tahmin edilen değer ile gerçek değer arasındaki farkın mutlak değerinin ortalaması olarak hesaplanır. OMH' nin küçük olması, modelin daha doğru tahminler yaptığı anlamına gelirken, büyük OMH değerleri modelin daha düşük bir tahmin doğruluğuna sahip olduğunu gösterir.

OMH, özellikle regresyon modellerinin performansını değerlendirmede yaygın olarak kullanılır. Örneğin, bir ev fiyatı tahmin modelinin OMH değeri, modelin ev fiyatlarını ne kadar doğru tahmin ettiğini belirtir. Daha düşük OMH değerleri, modelin daha iyi tahminler yaptığını ve daha yüksek doğruluk seviyelerine sahip olduğunu gösterir (Adalıer, 2008, s.65).

4.4.4. Hataların karelerinin ortalamasının karekökü

Bir diğer hata ölçüsü hataların karelerinin ortalamasının karekökü hesaplamasıdır. Bu durum aşağıdaki denklem aracılığıyla hesaplanır. Hataların karelerinin ortalamasının karekökü için uygulanan formül Eşitlik 4.5' te verilmiştir.

$$HKOK = \sqrt{\frac{\sum_{i=1}^{n} \left(\hat{\theta}_i - \theta_i\right)^2}{n}} \qquad (4.5)$$

Hataların Karelerinin Ortalamasının Karekökü (HKOK), bir tahmin modelinin gerçek değerlerle ne kadar uyumlu olduğunu ölçen bir ölçüttür. HKOK, modelin tahminlerinin gerçek değerlerden ne kadar sapma gösterdiğini belirtir ve bu sapmaların karelerinin ortalamasının kareköküdür.

Matematiksel olarak, bir modelin HKOK' si, her gözlem için tahmin edilen değer ile gerçek değer arasındaki farkın karesinin ortalamasının karekökü olarak hesaplanır. HKOK' nin küçük olması, modelin daha doğru tahminler yaptığı anlamına gelirken, büyük HKOK değerleri modelin daha düşük bir tahmin doğruluğuna sahip olduğunu gösterir (Alpaydın, 2011, s.65).

4.4.5. Göreceli mutlak hata

Tahmin ortalamasının, talepte oluşan değişikliklere ne kadar uyum sağladığını görebilmek için göreceli mutlak hata kullanılır. Göreceli mutlak hata için uygulanan formül Eşitlik 4.6' da verilmiştir.

$$GMH = \frac{\sum_{i=1}^{n} |\hat{\theta}_i - \theta_i|}{\sum_{i=1}^{n} |\bar{\theta} - \theta_i|} \quad (4.6)$$

Burada $\bar{\theta}$, gerçek taleplerin ortalamasıdır. $\bar{\theta}$ değeri Eşitlik 4.7' de de kullanılacaktır.

Matematiksel olarak, bir modelin Göreceli Mutlak Hatası (GMH), her gözlem için tahmin edilen değer ile gerçek değer arasındaki farkın mutlak değerinin gerçek değere bölünmesi ve ardından bu değerin ortalamasının alınmasıyla hesaplanır. Bu, bir tahminin gerçek değere göre yüzde olarak ne kadar sapma gösterdiğini ifade eder. Genellikle regresyon modellerinin performansını değerlendirmek için kullanılır. GMH yapılan tahminin gerçek değerlerle olan yakınlığını izlemenin bir yoludur (Aydemir, 2018, s.65).

4.4.6. Göreceli mutlak hata karekökü

Tahmin hatasını izlemede kullanılan bir başka yol da Göreceli Mutlak Hata Karekökü' dür. Göreceli mutlak hata karekökü için uygulanan formül Eşitlik 4.7' de verilmiştir.

$$GMHK = \sqrt{\frac{\sum_{i=1}^{n} (\hat{\theta}_i - \theta_i)^2}{\sum_{i=1}^{n} (\bar{\theta} - \theta_i)^2}} \quad (4.7)$$

Göreceli Mutlak Hata Karekökü (GMHK), bir tahmin modelinin gerçek değerlerle uyumluluğunu ölçen bir ölçüttür. Bu ölçüt, modelin tahminlerinin gerçek değerlere göre ne kadar doğru olduğunu yüzde olarak ifade ederken, aynı zamanda bu sapmaların karelerinin ortalamasının karekökünü alarak hesaplanır.

Matematiksel olarak, bir modelde GMHK, her gözlem için tahmin edilen değer ile gerçek değer arasındaki farkın karesinin alınması, ardından bu karelerin ortalamasının alınması ve son olarak bu değerin karekökünün alınmasıyla hesaplanır. Bu, bir tahminin gerçek değere göre yüzde olarak ne kadar sapma gösterdiğini ve bu sapmaların karelerinin ortalamasının karekökünün ne kadar olduğunu belirtir.

GMHK, genellikle regresyon modellerinin performansını değerlendirmek için kullanılır. Daha düşük GMHK değerleri, modelin daha az hata yaptığını ve daha iyi tahminler gerçekleştirdiğini gösterir. Bu ölçüt, modelin tahminlerinin ne kadar doğru olduğunu daha net bir şekilde belirlemeye yardımcı olur (Aydemir, 2018, s.65).

4.4.7. Doğru pozitif oranı

Doğru Pozitif Oranı (True Positive Rate), tıbbi teşhislerden makine öğrenimi modellerine kadar birçok alanda kullanılan önemli bir performans ölçütüdür. Bu ölçüt, bir modelin gerçek pozitif vakaları doğru bir şekilde tanımlama yeteneğini değerlendirir.

Doğru Pozitif Oranı, aynı zamanda hassasiyet veya duyarlılık olarak da bilinir. Bir sınıflandırma modelinde, doğru pozitif oranı, gerçek pozitif vakaların tüm pozitif vakalara oranını ifade eder. Matematiksel olarak, doğru pozitif oranı için uygulanan formül Eşitlik 4.8' de verilmiştir.

$$TP\ Rate = \frac{dp}{(dp + yn)} \tag{4.8}$$

Gerçek değeri pozitif olup pozitif olarak tahmin edilenler doğru pozitiftir ve burada "dp" olarak kısaltılacaktır. Gerçek değeri pozitif olup negatif olarak tahmin edilenler yanlış negatiftir ve "yn" olarak kısaltılacaktır. Pozitif olarak tahmin edilmiş fakat gerçek değeri negatif olanlar yanlış pozitiftir ve "yp" olarak kısaltılacaktır. Negatif olarak tahmin edilmiş ve gerçek değeri negatif olanlar doğru negatiftir ve "dn" olarak kısaltılacaktır.

Doğru Pozitif Oranı, bir modelin hastalığı veya bir durumu doğru bir şekilde tanımlama yeteneğini belirler. Özellikle tıbbi teşhislerde, bu ölçüt hastalıkların erken teşhisinde ve tedavi planlarının belirlenmesinde kritik öneme sahiptir. Yüksek bir doğru pozitif oranı, modelin hastalığı doğru bir şekilde tanımlama yeteneğinin yüksek olduğunu gösterirken, düşük bir doğru pozitif oranı ise modelin bu yeteneğinin zayıf olduğunu gösterir. Bu nedenle, doğru pozitif oranı, bir sınıflandırma modelinin performansını değerlendirmede önemli bir ölçüttür (Aydemir, 2018, s.65).

4.4.8. Yanlış pozitif oranı

Yanlış Pozitif Oranı (False Positive Rate), sınıflandırma modellerinin performansını değerlendirmede kullanılan önemli bir ölçüttür. Bu ölçüt, bir modelin gerçek negatif vakaları yanlış bir şekilde pozitif olarak tanımlama oranını ifade eder.

Yanlış Pozitif Oranı, bir modelin özellikle negatif sınıfa ait verileri yanlış bir şekilde pozitif olarak sınıflandırma yeteneğini gösterir. Matematiksel olarak, yanlış için uygulanan formül Eşitlik 4.9' da verilmiştir.

$$FP\ Rate = \frac{yp}{(yp + dn)} \quad (4.9)$$

Weka platformu özellikle doğru pozitif oranı ve yanlış pozitif oranı değerlerini, hem her bir kategori için oranlarını hem de tüm kategoriler için ağırlıklı ortalama sonucunu hesaplamaktadır. Performans analizi yapılırken ağırlıklı ortalama sonuçlar dikkate alınmıştır.

Yanlış Pozitif Oranı, bir modelin özellikle belirli bir durumu veya hastalığı yanlış bir şekilde tanımlama riskini gösterir. Yüksek bir yanlış pozitif oranı, modelin negatif sınıfa ait verileri yanlış bir şekilde pozitif olarak sınıflandırma eğiliminde olduğunu gösterirken, düşük bir yanlış pozitif oranı ise modelin bu eğiliminin daha az olduğunu gösterir. Bu nedenle, yanlış pozitif oranı, bir sınıflandırma modelinin spesifikliğini değerlendirmede önemli bir ölçüttür (Aydemir, 2018, s.65).

4.4.9. Kesinlik

Kesinlik, sınıflandırma modellerinin performansını değerlendirmek için kullanılan önemli bir ölçüttür. Bu ölçüt, modelin pozitif olarak sınıflandırdığı verilerin ne kadarının gerçekten pozitif olduğunu gösterir. Kesinlik, modelin yanlış pozitiflerin sayısını ne kadar az yaptığını ve gerçek pozitifleri ne kadar doğru bir şekilde saptadığını belirler.

Matematiksel olarak, kesinlik için uygulanan formül Eşitlik 4.10' da verilmiştir.

$$Kesinlik = \frac{dp}{(dp + yp)} \quad (4.10)$$

Burada, "Doğru Pozitifler", modelin doğru bir şekilde pozitif olarak sınıflandırdığı örneklerin sayısını, "Yanlış Pozitifler" ise modelin yanlışlıkla pozitif olarak sınıflandırdığı örneklerin sayısını temsil eder.

Kesinlik değeri 0 ile 1 arasında değişir, ve daha yüksek bir kesinlik değeri, modelin pozitif olarak sınıflandırdığı örneklerin gerçekten pozitif olma olasılığının daha yüksek olduğunu gösterir. Yüksek kesinlik değerleri, modelin pozitif sınıflandırmalarında güvenilir olduğunu ve yanlış pozitiflerin nadir olduğunu gösterir. Kesinlik ölçütü, özellikle dengesiz veri kümelerinde (örneğin, bir sınıfın diğerinden çok

daha fazla örneğe sahip olduğu durumlarda) model performansını değerlendirmede önemli bir rol oynar (Platt, 1998, s.68).

4.4.10. Hassasiyet

Hassasiyet, sınıflandırma modellerinin performansını değerlendirmek için kullanılan önemli bir ölçüttür. Bu ölçüt, gerçek pozitiflerin model tarafından ne kadar başarılı bir şekilde tespit edildiğini gösterir. Hassasiyet, modelin gerçek pozitifleri ne kadar doğru bir şekilde saptadığını ve kaç pozitif örneği kaçırdığını belirler.

Matematiksel olarak, hassasiyet için uygulanan formül Eşitlik 4.11' de verilmiştir.

$$Hassasiyet = \frac{dp}{(dp + yn)} \tag{4.11}$$

Burada, "Doğru Pozitifler", modelin doğru bir şekilde pozitif olarak sınıflandırdığı örneklerin sayısını, "Yanlış Negatifler" ise modelin yanlışlıkla negatif olarak sınıflandırdığı (ancak gerçekte pozitif olan) örneklerin sayısını temsil eder.

Hassasiyet değeri 0 ile 1 arasında değişir ve daha yüksek bir hassasiyet değeri, modelin gerçek pozitifleri tespit etme yeteneğinin daha yüksek olduğunu gösterir. Yüksek hassasiyet değerleri, modelin pozitif sınıflandırmalarında güvenilir olduğunu ve pozitif örnekleri kaçırma olasılığının düşük olduğunu gösterir. Hassasiyet ölçütü, özellikle tüm pozitif örneklerin doğru bir şekilde tespit edilmesinin önemli olduğu durumlarda model performansını değerlendirmede önemli bir rol oynar (Quinlan, 1986, s.68).

4.4.11. F-Ölçüsü

Kesinlik ve hassasiyet bir modelin performansını değerlendirmek için yaygın olarak kullanılır. Ancak, bu ölçütler tek başına yeterli değildir ve genellikle birlikte değerlendirilir. Bu bağlamda, F-Ölçüsü devreye girer. F-Ölçüsü, kesinlik ve hassasiyetin harmonik ortalamasını temsil eder ve bu sayede her iki ölçütü birleştirerek daha kapsamlı bir performans değerlendirmesi sağlar (Işık ve Ulusoy, 2021, s.66). F-ölçüsü, için uygulanan formül Eşitlik 4.12' de verilmiştir.

$$F - \text{Ölçüsü} = 2x\frac{pr}{p + r} \tag{4.12}$$

4.4.12. Alıcı işlem karakteristikleri

Doğru pozitif oranları ile yanlış pozitif oranlarının bir grafik üzerine yerleştirilmesi ile AİK eğrileri oluşturulur. Alıcı işlem karakteristikleri denilen bu eğrilerde dp'nin 1' e yaklaştıkça daha iyi tahminler elde edildiği, yp'nin ise 0' a yaklaştıkça daha az hatalı tahminler elde edildiği bilinmektedir. Grafiğin altında kalan alan AİK alanı olarak ifade edilir ve bu alan ne kadar büyükse o kadar başarılı bir sınıflandırma olduğu söylenebilir (Aydemir, 2018, s.65).

4.4.13. Kesinlik hassasiyet eğrisi alanı

Hassasiyet ve Kesinlik Değerlerinin bir grafik üzerine yerleştirilmesi ile KHE oluşturulur. KHE grafiği Kesinlik- Hassasiyet Eğrisi denilen bu eğrilerde çizilen grafiğin altında kalan alan KHE Alanı olarak adlandırılır. Buradaki değerin 1' e yaklaşması yüksek hassasiyet ile daha başarılı sonuçlar ürettiğini gösterir (Aydemir, 2018, s.65).

5. ANALİZ VE BULGULAR

Bu çalışmada, kalp hastalığı teşhisi konulmuş bireylerin belirlenmesi amacıyla makine öğrenmesi uygulamalarının performans analizi gerçekleştirilmiştir. Analiz kapsamında Weka platformu kullanılarak regresyon, sınıflandırma ve kümeleme algoritmaları olmak üzere üç farklı türde toplam 16 algoritma değerlendirilmiştir.

Regresyon algoritmaları olarak Lineer Regresyon, M5P ve Random Forest kullanılmıştır. Bu algoritmaların kalp hastalığı sınıflandırılmasındaki tahminlerinin performansı Weka' nın sağladığı kriterlere göre karşılaştırılmıştır. Sonuçlar incelendiğinde, Random Forest algoritmasının korelasyon katsayısı açısından diğerlerine göre daha başarılı olduğu belirlenmiştir. Ayrıca, hata oranını değerlendiren metriklerde (OMH, HKOK, GMH, GMHK), Random Forest algoritmasının diğerlerine göre daha düşük hata oranlarına sahip olduğu tespit edilmiştir. Çizelge 5.1' de Excel ile düzenlenen veri seti regresyon algoritmalarıyla çalıştırıldığında başarı ve hata oranı parametreleri görülmektedir.

Çizelge 5.1. Regresyon algoritmalarının başarı ve hata oranı parametreleri (Excel)

Yöntem	KK	OMH	HKOK	GMH	GMHK
Lineer Regresyon	0.7105	0.2733	0.3499	55.2741	70.3430
M5P	0.7472	0.2225	0.3308	45.0006	66.5086
Random Forest	**0.7613**	**0.2173**	**0.3226**	**43.9429**	**64.8648**

Çizelge 5.2' de Weka ile düzenlenen veri seti regresyon algoritmalarıyla çalıştırıldığında başarı ve hata oranı parametreleri görülmektedir.

Çizelge 5.2. Regresyon algoritmalarının başarı ve hata oranı parametreleri (Weka)

Yöntem	KK	OMH	HKOK	GMH	GMHK
Lineer Regresyon	0.7395	0.2463	0.3347	49.8159	67.2936
M5P	0.7261	0.2326	0.3425	47.041	68.8728
Random Forest	**0.7575**	**0.2164**	**0.3247**	**43.7569**	**65.2778**

Sınıflandırma algoritmalarının performansı incelendiğinde veri setindeki özelliklerin ve algoritmaların kalp hastalığının doğru sınıflandırılmasındaki etkisi değerlendirilmiştir.

Analiz kapsamında kullanılan algoritmalar arasında IBK, SMO, LibSVM, Random Forest, J48, NaiveBayes, OneR ve ZeroR bulunmaktadır. Bu algoritmaların performansları, doğruluk oranı, Kappa istatistiği, hata oranları (OMH, HKOK, GMH, GMHK), dp oranı, yp oranı, kesinlik, hassasiyet, F-Ölçüsü, AİK Alanı ve KHE Alanı gibi çeşitli metrikler kullanılarak değerlendirilmiştir.

Elde edilen bulgulara göre, IBK algoritması birçok performans ölçütü açısından en başarılı algoritma olarak öne çıkmaktadır. Özellikle doğru tahminlenen veri oranı, Kappa istatistiği ve hata oranları gibi parametrelerde IBK algoritmasının diğer algoritmaları geride bıraktığı gözlemlenmiştir. Bununla birlikte, LibSVM ve SMO gibi diğer algoritmalar da genel olarak iyi bir performans sergilemiştir. Çizelge 5.3' te ve Çizelge 5.4' te Excel ile düzenlenen veri seti sınıflandırma algoritmalarıyla çalıştırıldığında başarı, hata oranı parametreleri ve performans ölçütleri görülmektedir.

Çizelge 5.3. Sınıflandırma algoritmalarının başarı ve hata oranı parametreleri (Excel)

Yöntem	Doğruluk Oranı	KI	OMH	HKOK	GMH	GMHK
Random Forest	%85.4031	0.7034	0.2082	0.3383	42.126	68.0473
ZeroR	%55.3377	0	0.4943	0.4971	100	100
OneR	%81.3725	0.6218	0.1863	0.4316	37.6832	86.815
NaiveBayes	%83.5512	0.6661	0.18	0.3437	36.4154	69.1269
J48	%84.6405	0.6879	0.23	0.3602	46.5287	72.4634
IBK	**%86.1656**	**0.7189**	0.2094	**0.3318**	42.3541	**66.7445**
SMO	%86.0566	0.7169	**0.1394**	0.3734	**28.2073**	75.1107
LibSVM	%86.0566	0.7162	**0.1394**	0.3734	**28.2073**	75.1107

Çizelge 5.4. Sınıflandırma algoritmalarının diğer performans ölçütleri (Excel)

Yöntem	dp	yp	Kesinlik	Hassasiyet	F-Ö	AİK	KHE
Random Forest	0.854	0.154	0.854	0.854	0.854	0.911	0.905
ZeroR	0.553	0.553	tanımsız	0.553	tanımsız	0.498	0.505
OneR	0.814	0.194	0.813	0.814	0.813	0.810	0.755
NaiveBayes	0.836	0.172	0.835	0.836	0.835	**0.917**	**0.916**
J48	0.846	0.162	0.846	0.846	0.846	0.851	0.809
IBK	**0.862**	**0.146**	**0.862**	**0.862**	**0.861**	0.916	0.914
SMO	0.861	0.146	0.860	0.861	0.860	0.857	0.809
LibSVM	0.861	0.149	0.861	0.861	0.860	0.856	0.809

Çizelge 5.5' te ve Çizelge 5.6' da Weka ile düzenlenen veri seti sınıflandırma algoritmalarıyla çalıştırıldığında başarı, hata oranı parametreleri ve performans ölçütleri görülmektedir.

Çizelge 5.5. Sınıflandırma algoritmalarının başarı ve hata oranı parametreleri (Weka)

Yöntem	Doğruluk Oranı	KI	OMH	HKOK	GMH	GMHK
R.Forest	%79.7386	0.5858	0.3544	0.3962	71.6975	79.6862
ZeroR	%55.3377	0	0.4943	0.4971	100	100
OneR	%81.3725	0.6218	0.1863	0.4316	37.6832	86.8150
NaiveBayes	%85.0763	0.6983	0.168	0.3457	33.9808	69.5366
J48	%81.3725	0.6216	0.3003	0.3891	60.7598	78.2749
IBK	**%87.1460**	**0.7387**	0.1899	**0.3198**	38.4077	**64.3362**
SMO	%85.7298	0.7104	0.1427	0.3778	28.8685	75.9858
LibSVM	%85.8388	0.7105	**0.1416**	0.3763	**28.6481**	75.6952

Çizelge 5.6. Sınıflandırma algoritmalarının diğer performans ölçütleri (Weka)

Yöntem	dp	yp	Kesinlik	Hassasiyet	F-Ö	AİK	KHE
R.Forest	0.7970	0.2180	0.7980	0.7970	0.7960	0.8680	0.8610
ZeroR	0.5530	0.5530	Tanımsız	0.5530	Tanımsız	0.4980	0.5050
OneR	0.8140	0.1940	0.8130	0.8140	0.8130	0.8100	0.7550
NaiveBayes	0.8510	0.1520	0.8510	0.8510	0.8510	0.9120	0.9060
J48	0.8140	0.1950	0.8130	0.8140	0.8130	0.7960	0.7600
IBK	**0.8710**	**0.1360**	**0.8720**	**0.8710**	**0.8710**	**0.9240**	**0.9250**
SMO	0.8570	0.1490	0.8570	0.8570	0.8570	0.8540	0.8060
LibSVM	0.8580	0.1550	0.8600	0.8580	0.8570	0.8520	0.8030

Kümeleme algoritmalarının performansı incelendiğinde yanlış kümelenmiş verilerin oranı üzerinden algoritmaların performansları karşılaştırılmıştır. Excel ve Weka ile düzenlenen veri setlerine göre çalıştırılan algoritmalarda, Weka hata oranlarına göre değerlendirme yapılmıştır.

Buna göre, EM algoritmasının diğer kümeleme algoritmalarına göre daha başarılı olduğu görülmektedir. Bu durum, EM algoritmasının, veri setindeki yapıyı daha doğru bir şekilde tanımlayabildiğini ve daha tutarlı kümeler oluşturabildiğini göstermektedir. Öte yandan, Hiyerarşik algoritmanın en yüksek hata oranına sahip olması, bu algoritmanın performansının diğerlerine göre daha zayıf olduğunu göstermektedir. Çizelge 5.7' de Excel ile düzenlenen veri setine göre kümeleme algoritmalarının

performans parametreleri görülmektedir.

Çizelge 5.7. Kümeleme algoritmalarının performans parametreleri (Excel)

Yöntem	Yanlış Kümelenmiş Verilerin Oranı
K-Means	% 22.5490
X-Means	% 22.1133
Hiyerarşik	% 38.2353
EM	**% 22.0044**
Self-Organizing Maps	% 26.9063

Çizelge 5.8' de Weka ile düzenlenen veri setine göre kümeleme algoritmalarının performans parametreleri görülmektedir.

Çizelge 5.8. Kümeleme algoritmalarının performans parametreleri (Weka)

Yöntem	Yanlış Kümelenmiş Verilerin Oranı
K-Means	% 32.6797
X-Means	% 22.1133
Hiyerarşik	% 38.4532
EM	**% 19.1721**
Self-Organizing Maps	% 26.9063

Bu sonuçlar, Random Forest regresyon modelinin, IBK sınıflandırma algoritmasının ve EM kümeleme algoritmasının bu çalışma bağlamında en iyi performansı sergilediğini göstermektedir. Ancak, her bir algoritmanın avantajları ve dezavantajları göz önünde bulundurularak uygun uygulama alanları için seçilmesi gerekmektedir. Ayrıca, sonuçların farklı veri setleri ve analiz yöntemleriyle doğrulanması önemlidir. Bu sonuçlar, IBK algoritmasının kalp hastalığı sınıflandırılmasında önemli bir potansiyele sahip olduğunu ve diğer algoritmalarla kıyaslandığında daha etkili bir sınıflandırıcı olduğunu göstermekte olup LibSVM ve SMO algoritmalarının performansının da dikkat çekici olduğu gözlemlenmektedir. Genel olarak bu veri seti için sınıflandırıcı algoritmaların daha başarılı olduğu görülmüştür.

Yapılan analizlerde sınıflandırma algoritmalarında 10 katlı çapraz doğrulama yöntemi kullanılmıştır. Çapraz doğrulama kat sayısının değiştirilmesi durumunda, sonuçlar üzerindeki etkisi test edilerek, algoritma performansları iyileştirmeye

çalışılmıştır. Bu nedenle kalp hastalığı tahmininde en başarılı modeller olan, IBK, LibSVM ve SMO yöntemleri için sınıflandırma performansları düzenlenen her iki veri setinde 5, 10, 15, 20, 25, 30, 35 ve 40 katlı çapraz doğrulama oranları ve %33' e %66 ayırma yöntemleri kullanılarak yeniden değerlendirilmiştir. Çizelge 5.5' te IBK algoritmasının farklı çapraz doğrulama oranları ile performansı Çizelge 5.6' da LibSVM algoritmasının farklı çapraz doğrulama oranları ile performansı ve Çizelge 5.7' de SMO algoritmasının farklı çapraz doğrulama oranları ile performansı görülmektedir.

Çizelge 5.9. IBK' nın farklı çapraz doğrulama oranları ile performansı (Excel)

Yöntem	Doğruluk Oranı	OMH	HKOK
5 Katlı Çapraz Doğrulama	85.2941	0.2119	0.3353
10 Katlı Çapraz Doğrulama	86.1656	0.2094	0.3318
15 Katlı Çapraz Doğrulama	85.8388	0.2097	0.3334
20 Katlı Çapraz Doğrulama	86.1656	**0.2065**	**0.3291**
25 Katlı Çapraz Doğrulama	85.9477	0.2090	0.3319
30 Katlı Çapraz Doğrulama	85.8388	0.2076	0.3311
35 Katlı Çapraz Doğrulama	**86.2745**	0.2078	0.3304
40 Katlı Çapraz Doğrulama	86.1656	0.2075	0.3303
%66'ya 33 Ayırma	84.6154	0.2133	0.3458

Çizelge 5.10. IBK' nın farklı çapraz doğrulama oranları ile performansı (Weka)

Yöntem	Doğruluk Oranı	OMH	HKOK
5 Katlı Çapraz Doğrulama	**87.2549**	**0.1871**	**0.3164**
10 Katlı Çapraz Doğrulama	87.1460	0.1899	0.3198
15 Katlı Çapraz Doğrulama	86.6013	0.1896	0.3206
20 Katlı Çapraz Doğrulama	87.1460	0.1883	0.3190
25 Katlı Çapraz Doğrulama	86.9281	0.1894	0.3210
30 Katlı Çapraz Doğrulama	86.9281	0.1902	0.3210
35 Katlı Çapraz Doğrulama	**87.2549**	0.1897	0.3206
40 Katlı Çapraz Doğrulama	86.9281	0.1892	0.3208
%66'ya 33 Ayırma	85.8974	0.2069	0.3419

Çizelge 5.11. LibSVM' nin farklı çapraz doğrulama oranları ile performansı (Excel)

Yöntem	Doğruluk Oranı	OMH	HKOK
5 Katlı Çapraz Doğrulama	85.8388	0.1416	0.3763
10 Katlı Çapraz Doğrulama	86.0566	0.1394	0.3734
15 Katlı Çapraz Doğrulama	85.9477	0.1405	0.3749
20 Katlı Çapraz Doğrulama	86.0566	0.1394	0.3734
25 Katlı Çapraz Doğrulama	85.7298	0.1427	0.3778
30 Katlı Çapraz Doğrulama	85.9477	0.1405	0.3749
35 Katlı Çapraz Doğrulama	86.0566	0.1394	0.3734
40 Katlı Çapraz Doğrulama	**86.1656**	**0.1383**	**0.3719**
%66' ya 33 Ayırma	84.9359	0.1506	0.3881

Çizelge 5.12. LibSVM' nin farklı çapraz doğrulama oranları ile performansı (Weka)

Yöntem	Doğruluk Oranı	OMH	HKOK
5 Katlı Çapraz Doğrulama	85.8388	0.1416	0.3763
10 Katlı Çapraz Doğrulama	85.8388	0.1416	0.3763
15 Katlı Çapraz Doğrulama	85.9477	0.1405	0.3749
20 Katlı Çapraz Doğrulama	**86.0566**	**0.1394**	**0.3734**
25 Katlı Çapraz Doğrulama	85.7298	0.1427	0.3778
30 Katlı Çapraz Doğrulama	85.7298	0.1427	0.3778
35 Katlı Çapraz Doğrulama	85.7298	0.1427	0.3778
40 Katlı Çapraz Doğrulama	85.8388	0.1416	0.3763
%66' ya 33 Ayırma	83.6538	0.1635	0.4043

Çizelge 5.13. SMO' nun farklı çapraz doğrulama oranları ile performansı (Excel)

Yöntem	Doğruluk Oranı	OMH	HKOK
5 Katlı Çapraz Doğrulama	86.6013	0.1340	0.3660
10 Katlı Çapraz Doğrulama	86.0566	0.1394	0.3734
15 Katlı Çapraz Doğrulama	**86.7102**	**0.1329**	**0.3646**
20 Katlı Çapraz Doğrulama	86.4924	0.1351	0.3675
25 Katlı Çapraz Doğrulama	86.4924	0.1351	0.3675
30 Katlı Çapraz Doğrulama	**86.7102**	**0.1329**	**0.3646**
35 Katlı Çapraz Doğrulama	86.4924	0.1351	0.3675
40 Katlı Çapraz Doğrulama	86.4924	0.1351	0.3675
%66' ya 33 Ayırma	85.2564	0.1474	0.3840

Çizelge 5.14. SMO' nun farklı çapraz doğrulama oranları ile performansı (Weka)

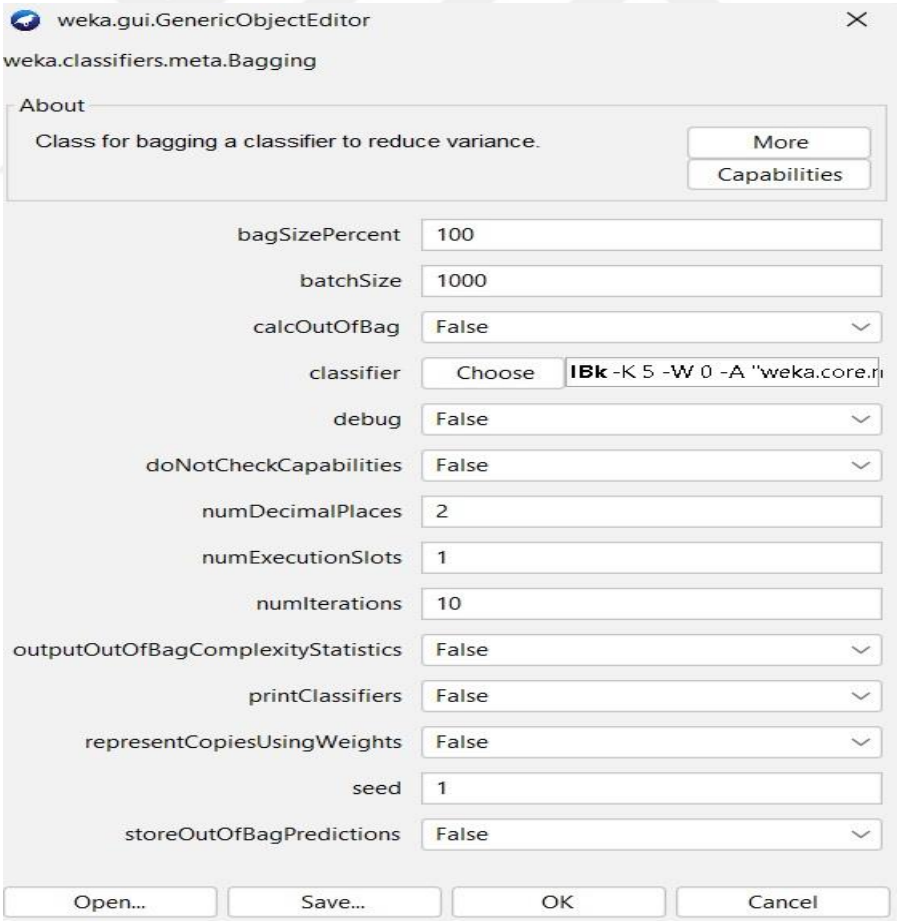
Yöntem	Doğruluk Oranı	OMH	HKOK
5 Katlı Çapraz Doğrulama	85.2941	0.1471	0.3835
10 Katlı Çapraz Doğrulama	85.7298	0.1427	0.3778
15 Katlı Çapraz Doğrulama	86.0566	0.1394	0.3734
20 Katlı Çapraz Doğrulama	**86.2745**	**0.1373**	**0.3705**
25 Katlı Çapraz Doğrulama	85.8388	0.1416	0.3763
30 Katlı Çapraz Doğrulama	**86.2745**	**0.1373**	**0.3705**
35 Katlı Çapraz Doğrulama	86.1656	0.1383	0.3719
40 Katlı Çapraz Doğrulama	**86.2745**	**0.1373**	**0.3705**
%66' ya 33 Ayırma	83.9744	0.1603	0.4003

Çizelge 5.10' da detaylı olarak incelenen sonuçlara göre, kalp hastalığı tahminlemesinde IBK algoritması için farklı katlı çapraz doğrulama oranlarıyla yapılan analizler dikkate alındığında, en yüksek başarı oranı Weka ile düzenlenen veri setinde (%87.2549) 5 ve 35 katlı çapraz doğrulama yöntemiyle elde edilmiştir. Ancak, veri setinin %66' lık ve %33' lük oranlarda ayrılmasıyla uygulanan yöntemde başarı oranında bir düşüş gözlemlenmiştir. Benzer şekilde, LibSVM algoritması için farklı katlı çapraz doğrulama yöntemleri Çizelge 5.11 ve Çizelge 5.12' de sunulmuş ve en yüksek başarı oranı Excel ile düzenlenen veri setinde (%86.1656) 40 katlı çapraz doğrulama yöntemiyle elde edilmiştir. Ancak, %66' lık ve %33' lük ayrılma oranları uygulandığında başarı oranında bir düşüş yaşanmıştır. Çizelge 5.13 ve Çizelge 5.14' te SMO algoritması için yapılan analizlerde ise en yüksek başarı oranı Excel ile düzenlenen veri setinde (%86.7102) 30 katlı çapraz doğrulama yöntemiyle elde edilmiştir. Ancak, %66' lık ve %33' lük ayrılma oranları uygulandığında başarı oranında bir düşüş gözlemlenmiştir.

Yeniden yapılan değerlendirmeler sonucunda, en başarılı üç algoritma için IBK algoritmasının 5 katlı çapraz doğrulama yöntemiyle uygulandığında en başarılı algoritma olduğu belirlenmiştir. 35 katlı çapraz doğrulama yöntemiyle aynı performansı gösterse de hata oranlarına göre daha az hata oranı sergilediği için 5 katlı yöntemin gerisinde kalmıştır. Ayrıca, bu analizde %66' lık ve %33' lük ayrılma yönteminin en kötü performansı sergileyen yöntem olduğu tespit edilmiştir.

Çalışmada, en yüksek sınıflandırma performansına sahip olan IBK algoritması

için 10 katlı çapraz doğrulama yöntemi sonucunda elde edilen %87.146' lık oranın farklı optimizasyon yöntemleriyle iyileştirilmesi hedeflenmiştir. Farklı katlarda çapraz doğrulama oranları ile performans yeniden değerlendirildiğinde, 5 katlı çapraz doğrulama yöntemi için sınıflandırma performansı %87.2549 olarak ölçülmüştür. Ayrıca, IBK yöntemi için sınıflandırma performansını arttırmak amacıyla Weka ortamında Bagging algoritması kullanılarak parametre optimizasyonu yapılmıştır. Bagging, temel algoritmanın birçok kopyasını eğiterek ve bu kopyaların tahminlerini bir araya getirerek birleştirerek daha güçlü ve istikrarlı bir model elde etmeyi amaçlar. Bagging, meta bir algoritma olarak kabul edilir, çünkü temel algoritma ile kullanılır. Weka' da Bagging uygularken, parametre optimizasyonu yapmak için bazı seçenekler bulunmaktadır. Bu seçenekler, kullanıcıların Bagging' in performansını artırmak için çeşitli ayarlar yapmasına izin verir. Şekil 5.1'de Bagging meta algoritmasının parametre ekranı görülmektedir.



Şekil 5.1. Weka bagging algoritması parametre ekranı

Bagging modeli, Şekil 5.1' de belirlenen parametrelere IBK algoritması üzerinde

test edilerek yeniden veri seti üzerinde çalıştırılmıştır. Modelin sınıflandırma performansını gösteren ekran Şekil 5.2' de sunulmuştur.

```
Correctly Classified Instances         802               87.3638 %
Kappa statistic                          0.743
Mean absolute error                      0.1895
Root mean squared error                  0.3175
Relative absolute error                 38.3317 %
Root relative squared error             63.8656 %
Total Number of Instances              918

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0,832     0,093     0,879       0,832   0,855      0,744    0,925     0,926     0
                 0,907     0,168     0,870       0,907   0,888      0,744    0,925     0,926     1
Weighted Avg.    0,874     0,134     0,874       0,874   0,873      0,744    0,925     0,926

=== Confusion Matrix ===

   a   b   <-- classified as
 341  69 |   a = 0
  47 461 |   b = 1
```

Şekil 5.2. Weka bagging modeli IBK algoritması sonuç ekranı

Şekil 5.2' deki veriler incelendiğinde, Bagging optimizasyonu sonucunda IBK algoritmasının %87.3638 doğru sınıflandırma performansı sergilediği gözlemlenmiştir. Elde edilen sonuçlar, Çizelge 5.15' te 10 katlı çapraz doğrulama oranı ve 5 katlı çapraz doğrulama oranı sonuçlarıyla karşılaştırılmıştır.

Çizelge 5.15. Kalp hastalığı tahmininde IBK algoritmasının iyileştirme sonuçları

Yöntem	**Doğru sınıflandırılan veri sayısı**	**Yanlış sınıflandırılan veri sayısı**	**Sınıflandırma Performansı**
IBK(10 Katlı)	800	118	87.1460
IBK (5 Katlı)	801	117	87.2549
Bagging IBK	**806**	**112**	**87.7996**

Bu çalışmada, kalp hastalığı hedef değişkeni için en yüksek sınıflandırma oranını sağlayan IBK algoritmasının farklı çapraz doğrulama oranları ve Bagging yöntemi ile sonuçlarının iyileştirilmesi amaçlanmıştır. Kalp hastalığının erken teşhisi için kullanılan risk faktörlerinin makine öğrenmesi yöntemleriyle değerlendirilmesi sonucunda, kullanılan yöntemlerin sınıflandırma performansları analiz edilmiştir. Çizelge 5.15 incelendiğinde, IBK yöntemi için yapılan Bagging optimizasyonu sonucunda 5 ve 10 katlı çapraz doğrulama yöntemine göre performans artışı gözlemlenmiştir. Sınıflandırma

sonucunda, 918 özniteliğin 806 tanesi doğru olarak tahmin edilmiştir. Optimizasyon işleminin ardından sınıflandırmanın performansı % 87.7996 olarak ölçülmüştür. Bagging optimizasyonu ile elde edilen sınıflandırma performansı diğer yöntemlere göre daha başarılı sonuçlar vermiştir. Bu durum, çalışmanın performansta iyileştirme sağladığını ve geliştirilebileceğini göstermektedir. Tez çalışması kapsamında, en iyi sınıflandırma performansını gösteren algoritmanın IBK algoritması olduğu belirlenmiştir. Yapılan parametre optimizasyonunun sınıflandırma performansı üzerindeki etkisi incelenmiş olup, kullanılan makine öğrenmesi yönteminin performansında artış sağlanmıştır.

6. SONUÇLAR

Çalışmada, kalp hastalığı teşhisi için makine öğrenmesi uygulamalarının performans analizinin yapılarak regresyon, sınıflandırma ve kümeleme olmak üzere üç farklı türde toplam 16 algoritma değerlendirilmiştir. Regresyon algoritmaları arasında Lineer Regresyon, M5P ve Random Forest kullanılmıştır. Analiz sonuçlarına göre, Random Forest algoritmasının korelasyon katsayısı açısından diğerlerine göre daha başarılı olduğu ve hata oranları açısından da diğer regresyon algoritmalarından daha düşük hata oranlarına sahip olduğu belirlenmiştir.

Sınıflandırma algoritmaları arasında IBK, SMO, LibSVM, Random Forest, J48, NaiveBayes, OneR ve ZeroR algoritmaları değerlendirilmiştir. IBK algoritmasının birçok performans ölçütü açısından en başarılı algoritma olduğu tespit edilmiştir. Ayrıca, LibSVM ve SMO gibi diğer algoritmalar da genel olarak iyi bir performans sergilemiştir.

Kümeleme algoritmaları incelendiğinde, EM algoritmasının diğer kümeleme algoritmalarına göre daha başarılı olduğu, Hiyerarşik algoritmanın ise en yüksek hata oranına sahip olduğu görülmüştür.

Yeniden yapılan değerlendirmeler sonucunda, en başarılı üç algoritma için IBK algoritmasının 5 katlı çapraz doğrulama yöntemiyle en başarılı algoritma olduğu belirlenmiştir. Bagging optimizasyonu ile IBK algoritmasının sınıflandırma performansının daha da arttığı gözlemlenmiştir.

Bu tez çalışmasında kalp hastalığı teşhisi için kullanılan makine öğrenmesi algoritmalarının performansları değerlendirilmiş ve özellikle IBK algoritmasının yüksek performans gösterdiği tespit edilmiştir. Elde edilen bu sonuçlar doğrultusunda, gelecekte yapılacak çalışmalar için çeşitli öneriler sunulmaktadır. Öncelikle, farklı coğrafi bölgelerden daha geniş ve çeşitlendirilmiş veri setlerinin kullanılması, modellerin genellenebilirliğini artıracaktır. Veri ön işleme aşamasında eksik verilerin tahmini, veri normalizasyonu ve özellik seçimi gibi tekniklerin daha etkin uygulanması, model performansını artırabilir. Ayrıca, derin öğrenme yöntemleri ve hibrit modeller gibi daha ileri makine öğrenmesi tekniklerinin kalp hastalığı teşhisindeki performanslarının araştırılması önemlidir. IBK algoritmasının performansını artıran Bagging gibi optimizasyon tekniklerinin diğer algoritmalarda da uygulanması performans artışı sağlayabilir. Makine öğrenmesi modellerinin klinik uygulamalarla entegrasyonu, modellerin gerçek dünya ortamlarında test edilmesini ve geçerliliklerinin doğrulanmasını

sağlayacaktır. Bu bağlamda, sağlık profesyonelleri ile iş birliği yapılarak geliştirilen modellerin klinik kullanıma uygunluğu değerlendirilmeli ve gerekli düzenlemeler yapılmalıdır. Sağlık verilerinin kullanımı sırasında hasta mahremiyeti ve veri güvenliği konularına dikkat edilmesi gerekmektedir. Yapay zeka etiği ve veri güvenliği konusunda da detaylı analizler yapılarak uygun çözümler geliştirilmelidir.

Son olarak, çalışmanın bulgularının başka araştırmacılar tarafından da doğrulanması ve genişletilmesi için sonuçlar ve kullanılan yöntemler detaylı bir şekilde raporlanarak, açık veri setleri ve algoritmalar paylaşılmış ve diğer araştırmacıların erişimine sunulmuştur. Bu öneriler doğrultusunda yapılacak çalışmalar, kalp hastalığı teşhisi için daha etkili ve güvenilir makine öğrenmesi modellerinin geliştirilmesine katkı sağlayacak ve hastaların yaşam kalitesini artıracaktır.

7. KAYNAKÇA

Adalıer, O. (2008). Yapay zekâ yöntemleri ile yazılım projelerinde maliyet kestirimi. Doktora Tezi. İzmir: Ege Üniversitesi, Fen Bilimleri Enstitüsü.

Akay, A. (2015). Veri madenciliği teknikleri ve uygulamaları: geçmişten günümüze bir inceleme. Yüksek Lisans Tezi. İstanbul: Boğaziçi Üniversitesi, Fen Bilimleri Enstitüsü.

Alpaydın, E. (2011). Yapay Öğrenme. İstanbul: Boğaziçi Üniversitesi Yayınevi.

Altıntaş, T. (2006). Veri madenciliği metotlarından kümeleme algoritmalarının uygulamalı etkinlik analizi. Yüksek Lisans Tezi. Sakarya: Sakarya Üniversitesi, Fen Bilimleri Enstitüsü.

Aydemir, E. (2018). Weka ile Yapay Zeka. Ankara: Seçkin Yayınevi.

Aydın, E., ve Uzunboylu, H. (2019). Teknoloji bağımlılığı ve sağlık üzerine etkileri. Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi, 20(3), 955-978.

Bahrami, B. and Shirvani M. H. (2015). Prediction and diagnosis of heart disease by data mining techniques. Journal of Multidisciplinary Engineering Science and Technology (JMEST), vol. 2, pp. 164-168.

Beltrame, J. F., and Crea, F. (2018). The discovery of coronary microvascular dysfunction. A Story of Three Decades. Journal of the American College of Cardiology, 72(19), 2339-2340.

Bilgin, M., ve Eren, F. (2019). Yüksek kolesterol ve kalp hastalığı riski arasındaki ilişki. Ankara Üniversitesi Tıp Fakültesi Mecmuası, 72(2), 149-154.

Boden, M. A. (2016). AI: Its Nature and Future. London: Oxford University Press.

Can, F., ve Kaya, M. (2019). Makine öğrenimi ve veri analizi: güncel yaklaşımlar ve uygulamalar. Uluslararası Bilgisayar Bilimleri Konferansı Bildirileri, 25-31.

Chawla, N. V., and Davis, D. A. (2013). Bringing big data to personalized healthcare: A patient-centered framework. Journal of General Internal Medicine, 28(3), 660-665.

Cheng, P. M., and Montagnon, E. (2018). Deep learning for lung cancer detection: Tackling the challenges of clinical implementation. Journal of Thoracic Disease, 10(7), 867-868

Copeland, B. J. (2004). The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life. London: Oxford University

Press.

Çelik, A., ve Erdoğan, O. (2020). Sigara içmenin kalp hastalığı üzerindeki etkileri ve risk faktörleri. Gülhane Tıp Dergisi, 62(1), 46-52.

Çelik, İ., ve Öztürk, S. (2019). Veri madenciliği uygulamalarında Weka kullanımı. Akademik Bilişim Konferansı. Malatya: İnönü Üniversitesi.

Dalkıran, İ., ve Ozan, M. (2022). Derin öğrenme teknikleri kullanılarak borsadaki hisse değerlerinin tahmin edilmesi. Avrupa Bilim ve Teknoloji Dergisi, (39), 143-148.

Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78-87.

Demir, A. (2019). Veri madenciliği araçları ve uygulamaları: açık kaynak kodlu ve ücretli yazılımlar üzerine bir inceleme. Yüksek Lisans Tezi. Ankara: Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü.

Erem, C. (2018). Türkiye'de hipertansiyon epidemiyolojisi. Hipertansiyon ve Böbrek Hastalıkları Dergisi, 27(3), 157-165.

Erkan, K. (2017). Yapay zekâ: İnsan zekâsını taklit eden teknolojiler. TÜBİTAK Bilim ve Teknik Dergisi, 59(1), 24-29.

Eroğlu, K. (2019). Kalp hastalıklarında yaşlanma ve risk faktörleri. Ankara Üniversitesi Tıp Fakültesi Mecmuası, 72(2), 141-148.

Esteva, A. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, 17(3), 37-54.

French, R. M. (2000). The Turing test: The First fifty years. Trends in Cognitive Sciences, 4(3), 115-122.

Gregg, E. W., Sattar, N., and Ali, M. K. (2016). The changing face of diabetes complications. The Lancet Diabetes and Endocrinology, 4(6), 537-547.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Heidelberg: Springer Science and Business Media.

Holzinger, A. (2014). Biomedical informatics: Discovering knowledge in big data. Communications of the ACM, 57(1), 70-79.

Işık, K., & Ulusoy, S. K. (2021). Determining the factors that affect the production time in metal industry utilizing data mining methods. Journal of the Faculty of

Engineering and Architecture of Gazi University, 36(4), 1949-1962.

Jindal, H., Agrawal, S., Khera, R., Jain, R., and Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. IOP Conference Series: Materials Science and Engineering, 1022(1).

Kandemir, Ö., and Turhan, H. (2015). Türkiye'de kardiyovasküler hastalıkların epidemiyolojisi ve erken teşhis stratejileri. Türk Kardiyoloji Dergisi, 42(1), 95-102.

Kaplan, A., and Haenlein, M. (2019). Siri, Siri, in my Hand: Who's the smartest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horizons, 62(1), 15-25.

Kaya, H., ve Ertürk, S. (2019). Yapay zeka ve disiplinler arası etkileşim. İstanbul Üniversitesi İktisat Fakültesi E-Dergisi, 69, 217-234.

Konuralp, S. (2018). Veri madenciliği yöntemleri ve uygulamaları: bir literatür taraması. Yüksek Lisans Tezi. Ankara: Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü.

Kretschmann, J., Becker, M., and Ketterlin, A. (2015). Protein sequence analysis using weka: The practical part of the Weka course. BMC Bioinformatics, 16(5), 156.

Kumdereli, Ü. C. (2012). Tıp bilişimi ve veri madenciliği uygulamaları: EEG sinyallerindeki epileptiform aktiviteye veri madenciliği yöntemlerinin uygulanması. Yüksek Lisans Tezi. Edirne: Trakya Üniversitesi, Fen Bilimleri Enstitüsü.

Li, Y., Guo, Y., Chen, Y. and He, Y. (2019). A deep learning-based radiomics model for differentiating benign and malignant renal tumors. Translational Andrology and Urology, 8(6), 618–619.

Mahmoodabadi, Z. and Abadeh, M.S. (2014). CADICA: Diagnosis of coronary artery disease using the imperialist competitive algorithm. Journal of Computing Science and Engineering, 8, 87-93.

Maini E., Venkateswarlu B., Maini B., and Marwaha D. (2021). Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India. Medical Journal Armed Forces India, 77(3), 302–311.

Marapelli, B. (2019). Software development effort duration and cost estimation using linear regression and k-nearest neighbors machine learning algorithms. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 9(2), 2278-3075.

McCarthy, J. (2022). Machine Learning and the City: Applications in Architecture and Urban Design. New Jersey: Wiley-Blackwell.

Nattel, S. and Dobrev, D. (2017). Electrophysiological and molecular mechanisms of paroxysmal atrial fibrillation. Nature Reviews Cardiology, 14(10), 575-590.

Jain, A. K. and Dubes, R. C. (1988). Algorithms for clustering data. New Jersey: Prentice Hall.

Özdemir, R. and Arslan, D. (2018). Kalp sağlığına yönelik risk faktörleri ve önlemler. Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi, 19(1), 345-357.

Özkan, E. (2017). Veri ambarları ve veri madenciliği uygulamaları. Yüksek Lisans Tezi. İstanbul: İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü.

Özlen, T. (2022). Servikal kanserlerin teşhisinde kullanılan makine öğrenmesi algoritmalarının karşılaştırmalı analizi. Yüksek Lisans Tezi. İstanbul: İstanbul Aydın Üniversitesi, Fen Bilimleri Enstitüsü.

Öztürk, M. ve Genç, Z. (2019). Makine öğrenmesi ve yapay zekâ uygulamaları. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı, 44-49.

Piatetsky-Shapiro, G., and Frawley, W. J. (1991). Knowledge discovery in real databases: A report on the IJCAI-89 Workshop. AI Magazine, 13(2), 57-70.

Platt, J. C. (1998). Sequential minimal optimization: A Fast algorithm for training support vector machines. Washington: Microsoft Research.

Puyalnithi, T. and Vankadara, M. (2017). Performance analysis of classification algorithms on a novel unified clinical decision support model for predicting coronary heart disease risks. International Journal of Intelligent Engineering and Systems, 10(3), 210-217.

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106.

R. Katarya and S. K. Meena. (2021). Machine learning techniques for heart disease prediction: A comparative study and analysis. Health Technology, 11(1), 87–97.

R. R. Sanni and H. S. Guruprasad. (2021) Analysis of performance metrics of heart failured patients using Python and machine learning algorithms. Global Transitions Proceedings, 2(2), 233–237.

Russell, S. J., and Norvig, P. (2016). Artificial Intelligence: A Modern Approach. London: Pearson.

Srinivas, K., Rani, B. K. and Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on

Computer Science and Engineering (IJCSE), 2(2), 250–255.

Şen, S. (2018). Makine öğrenmesi ve veri madenciliği teknikleri ile finansal tahminler. Doktora Tezi. Ankara: Orta Doğu Teknik Üniversitesi, Fen Bilimleri Enstitüsü.

Şeyranlıoğlu, O. (2022). Şirket değerlemesinde makine öğrenmesi algoritmalarının kullanımı: holding şirketleri üzerine bir araştırma. Doktora Tezi. Giresun: Giresun Üniversitesi, Fen Bilimleri Enstitüsü.

Taşçılar, M. E. (2021). Genetik faktörlerin kalp hastalıkları üzerindeki etkileri. Ankara Üniversitesi Tıp Fakültesi Mecmuası, 74(3), 367-375.

Taşpınar, M. (2013). Makine Öğrenmesi ve Uygulamaları. İstanbul: Nobel Akademik Yayıncılık.

Tobler, R., Rohrlach, A. B., Soubrier, J., Bover, P. and Llamas, B. (2015). Variable melting temperatures of megafaunal DNA samples from Southern Australia: Implications for Palaeogenomics. Scientific Reports, 5(5), 15568.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.

Tüfekçi, P. ve Doğan, İ. (2020). Veri ön işleme süreci ve makine öğrenimi algoritmalarına etkisi. Bilgisayar Mühendisliği ve Veri Bilimi Dergisi, 8(2), 87-102.

Venkatalakshmi B. and Shivsankar M. (2014). Heart disease diagnosis using predictive data mining, International Journal of Innovative Research in Science, Engineering and Technology, 3, 1873-1877.

Witten, I. H., Frank, E. and Hall, M. A. (2011). Data mining: Practical machine learning tools and techniques. Proceedings of the 4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 123-130.

Yılmaz, B. ve Karahan, E. (2020). Yapay zeka ve insan benzeri hareket ve düşünme yetenekleri. Bilim ve Teknik Dergisi, 592, 44-49.

http-1:https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases(cvds)

http-2:https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

http-3:https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

ÖZGEÇMİŞ

Kişisel Bilgiler	
Adı Soyadı	**Bekir Can TELKENAROĞLU**

Eğitim	
Lisans	2014, Yıldız Teknik Üniversitesi, Gemi İnşa ve Denizcilik Fakültesi, Gemi İnşaatı ve Gemi Makineleri Mühendisliği
Yüksek Lisans	2024, Bakırçay Üniversitesi, Lisansüstü Eğitim Enstitüsü, Akıllı Sistemler Mühendisliği Anabilim Dalı

Yayın Listesi	
Makale	Köse, B., Telkenaroğlu, B. C., & Demirtürk, B. (2024). Rüzgar Enerjisi Güç Yoğunluğu Tahmininde Optimum Weibull Olasılık Dağılım Parametrelerinin Elde Edilmesi İçin İstatistik, Matematik Ve Fizik Tabanlı Algoritmaların Karşılaştırmalı Analizi: Loras Ve Foça Örnekleri. Isı Bilimi ve Tekniği Dergisi, 44(1), 47-58.
	Köse, B., Telkenaroğlu, B. C., & Demirtürk, B. (2024). Makine Öğrenmesi Algoritmaları ile Kalp Hastalığı Tespitinin Performans Karşılaştırmaları. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi(Yayına Sunuldu).
Bildiri	Demirtürk, B., & Telkenaroğlu, B. C. (2023, July). A Performance Analysis Comparison of Machine Learning Algorithms In Detection Of Heart Disease. In 9th International IFS and Contemporary Mathematics and Engineering Conference (p. 220), Tarsus/Mersin/TÜRKİYE