ARAŞTIRMA MAKALESİ/Research Article

Sağlık Bilimlerinde Yapay Zeka Dergisi

Journal of Artifical Intelligence in Health Sciences

CPV Tahmininde Makine Öğrenme Yöntemlerinin Performanslarının Değerlendirilmesi: Farklı Bölgeler ve Parametreler ile Büyük Veri Uygulaması

Evaluating the Performance of Machine Learning Methods in CPV Prediction: Big Data Application Wiih Different Regions and Parameters

Gözde ZABZUN[1], Meltem SEVİNÇ[2], Çınar DALKILIÇ[3], Kıvanç Ege ÇAM[4]

Received / Geliş	**06.12.2022**
Accepted / Kabul	**27.12.2022**
Publication Date	**30.12.2022**

***Sorumlu Yazar**
Corresponding Author

Gözde ZABZUN[1]

Karşıyaka Aydoğan Yağcı Bilim ve Sanat Merkezi
ÇİĞLİ/İZMİR

ORCID iD: 0000-0002-9502-8756
e posta: gzabzun@gmail.com

Meltem SEVİNÇ[2]

Karşıyaka Aydoğan Yağcı Bilim ve Sanat Merkezi
ÇİĞLİ/İZMİR

ORCID iD: 0000-0001-6351-3370
e posta: sevincmeltem2007@gmail.com

Çınar DALKILIÇ[3]

Karşıyaka Aydoğan Yağcı Bilim ve Sanat Merkezi
ÇİĞLİ/İZMİR

ORCID iD: 0000-0002-1881-56840
e posta: cinar.dalkilic@gamil.com

Kıvanç Ege ÇAM[4]

Karşıyaka Aydoğan Yağcı Bilim ve Sanat Merkezi
ÇİĞLİ/İZMİR

ORCID iD: 0000-0002-5951-7245
e posta: kivanccam2@gmail.com

Yazar Katkıları
Motivasyon / Konsept: GZ
Çalışma Tasarımı: GZ
Kontrol / Gözetim: GZ
Veri Toplanması ve / veya İşlemesi: MS
Analiz ve / veya Yorum: MS
Literatür incelemesi: ÇD
Makalenin Yazılması: GZ, MS, KEÇ
Eleştirel İnceleme: GZ, KEÇ

ABSTRACT

Scopa: Nowadays, the interpretation, classification, storage and extraction of big data in different fields, which are rapidly increasing in regular and irregular areas, and making them useful again are among the subjects that are intensively studied. The correct interpretation of big data in the field of health is of vital importance as it enables fast and accurate diagnosis. In the project, machine learning methods that can interpret health data have been applied specifically to Canine parvovirus infection. While CPV can be diagnosed based on clinical findings, it needs to be supported by laboratory findings to distinguish it from other infections. Correct diagnosis is vital to distinguish CPV from other infections with bloody diarrhoea, which can result in death in puppies. For this reason, by analysing the virus together with other data that may be affected by the virus, the methods of making the most accurate decision were compared and evaluated.

Purpose: In this study, it was aimed to interpret CPV, which is considered to be one of the most important infectious agents of dogs, popularly known as mad-head disease, using K-NearestNeighbour (KNN), RandomForest (RF), Logistic Regression and NaiveBayes classification algorithms in terms of different parameters. When the total accuracy values were examined, the accuracy rates decreased in logistic regression and RF methods when the insignificant variable was removed in the model.

Result: RF method made the best predictions when Platelets, Platelet (PLT) variable was in the model.In cases where we do not want to remove this variable from the model, it can give us very efficient results. KNN method gives better results when the number of variables decreases. Especially when the data size increases, it has been observed that the machine learning method gives more efficient results with better performance.

Keywords: *Big Data, Data Mining, Big Data in Health, Prediction Algorithms, Canine parvovirus (CPV)*

Özet

Kapsamı: Günümüzde hızla artan düzenli/düzensiz farklı alanlardaki büyük verilerin yorumlanması, sınıflandırılması, depolanması ve ayıklanarak tekrar işe yarar hale getirilmesi, üzerinde yoğun çalışılan konuların arasındadır. Sağlık alanındaki büyük verilerin doğru yorumlanması ise hızlı ve doğru teşhis konulmasını sağladığından hayati öneme sahiptir. Projede sağlık verilerinin yorumlanabileceği makine öğrenme yöntemleri, Canine parvovirüsü enfeksiyonun özelinde uygulanmıştır. CPV klinik bulgulara dayanılarak teşhis konulabilirken; diğer enfeksiyonlardan ayırt edilebilmesi için ise laboratuvar bulguları ile desteklenmesi gerekir. CPV, yavru köpeklerde ölümle sonuçlanabilen, kanlı ishalle seyreden diğer enfeksiyonlardan ayırt edilebilmesi için doğru teşhis hayati önem taşır. Bu sebeple, virüsün etkilenebileceği diğer verilerle beraber incelenmesi yapılarak, en doğru kararın alma yöntemleri karşılaştırılarak değerlendirilmiştir.

Amaç: Çalışmada, halk arasında delibaş hastalığı olarak bilinen, köpeklerin en önemli enfeksiyöz etkenlerinden birisi olarak kabul edilen CPV farklı parametreleri açısından En Yakın komşu Algoritması (KNN), Rastgele Orman (RF), Lojistik Regresyon ve NaiveBayes sınıflandırma algoritmaları kullanarak yorumlamayı hedeflemiştir.

Sonuç/Bulgular: Toplam doğruluk değerleri incelendiğinde anlamsız değişken modelde çıkarıldığında lojistik regresyon ve RF yöntemlerinde doğruluk oranları düşmüştür. RF yöntemi Platelets,Trombosit (PLT) değişkeni modelde iken en iyi tahminleri yapmıştır. Bu değişkeni modelden çıkarmak istemediğimiz durumlarda bize çok verimli sonuçlar verebilmektedir. KNN yöntemi değişken sayısı azaldığında daha iyi sonuçlar vermektedir. Özellikle veri boyutu arttığında makine öğrenmesi yöntemi daha iyi performans ile daha verimli sonuçlar verdiği gözlemlenmiştir.

Anahtar kelimeler: *Büyük Veri, Veri Madenciliği, Sağlıkta Büyük Veri, Tahmin Algoritmaları, Canine parvovirüsü (CPV)*

Amaç

Projenin amacı, büyük veri olan sağlık verilerinin yorumlanmasında makine öğrenmesi kullanarak farklı yöntemler karşılaştırılarak değerlendirilmesidir. Araştırmada sağlık verilerinin değerlendirilmesinde dört farklı yöntem kullanılarak, sonuçlar karşılaştırılmış ve farklı yönlerden en iyi sonuçların elde edilme koşulları araştırılmıştır. Çalışmada, halk arasında delibaş hastalığı olarak bilinen, yavru köpeklerin en önemli enfeksiyonlarından biri olarak görülen Canine parvovirüsü (CPV) farklı parametreleri açısından En yakın komşu algoritması (KNN), Rastgele Orman, Lojistik Regresyon ve NaiveBayes sınıflandırma algoritmaları kullanarak yorumlamayı hedeflemiştir. Proje, elde edilen modellerin en doğru tahmini yapabilme koşulları karşılaştırmalı olarak değerlendirerek, faydalı olanların kullanılabilirliğini artırmayı amaçlar.

Giriş

Biyolojik problemlerin tanımlanması ve çözümlerinde gerçek yaşama uygun doğru modellemelerin yapılabilmesi için büyük veri setlerinin kullanılması önemlidir. Biyolojik verilerin derlenmesi, işlenmesi, depolanması ve analiz edilmesi biyoinformatik bilim dalının konu alanıdır. (1).Bu alandaki büyük veriler, biyoinformatik açısından incelendiğinde hipotez odaklı araştırmalardan uzaklaşıp, veri odaklı modeller bulma arayışına yönlendirmektedir (2).Veri madenciliği yaklaşımları ise çok büyük miktarlardaki biyolojik verilerin biyoinformatik uygulamalar için analizinde idealdir.

Veri madenciliği tekniklerinin etkili ve doğru kullanılması ile biyolojik verilerden bilimsel çıkarım yapılmasını sağlayabilmektedir (3;1).21.yüzyıla kadar biyolojik verilerin depolanması ve veri tabanlarının analiz edilebilecek araçlar ulaşılabilir değildi (4) Ancak günümüz teknolojisi, bilim insanlarının çeşitli sağlık problemlerini çözebilecek büyük verileri yönetmelerine ve analiz etmelerine olanak sağlamaktadır (5). Biyolojik verilerin işlenmesi ve kullanılabilir modeller üretilmesi için sağlanan bu araçlar veri madenciliği konu alanına girmektedir.

Araştırma ekibi, biyolojik verilerin veri madenciliği ile büyük verilerin tahmin edilebilirliğini ve dahası doğru karar alınabilirliği için farklı yöntemleri denemiştir. Yapılan çalışmalar incelenmiş ve en çok tercih edilen dört farklı makine öğrenme yöntemi daha önce uygulanmamış bir veri setinde uygulanmış ve farklı açılardan değerlendirilmiştir. Araştırma ekibi, özellikle yavru köpekleri etkileyen, sonu ölümle sonuçlanabilen CPV'yi farklı değişkenler bazında incelemiştir. Halk arasında delibaş hastalığı olarak bilinen CPV,1978 yılında ilk kez dünyada tanımlanmış ve yavru köpeklerde en önemli enfeksiyöz olarak kabul edilmiştir. Ardından birçok ülkede (6;7) ve Türkiye'de (8; 9) de CPV'nin etkileri görülmeye başlanmıştır. CPV enfeksiyonunun teşhisi klinik bulgulara dayanılarak konulabilse de kesin teşhisin özellikle köpek yavrularında kanlı ishalle seyreden diğer enfeksiyonlardan ayırt edilmesi açısından laboratuvar bulguları ile doğrulanması gerekir. Bu sebeple, projede virüsün etkilenebileceği diğer verilerle beraber incelenmesi yapılarak, en doğru kararın alma yöntemleri karşılaştırılarak değerlendirilmiştir.

Sağlıkta Büyük Veri

"Sağlıkta Büyük Veriler, sağlık ve sağlık sistemi performansını artırmak amacıyla elektronik olarak yakalanan ve saklanan, rutin veya otomatik olarak toplanan büyük veri kümelerini ifade eder. Büyük Veri çok amaçlı veri anlamında yeniden kullanılabilir ve var olan veri tabanlarının birleştirilmesi ve bağlantısını içerir" (10). Günümüzde büyük veri

analizi sağlık sektöründe AR-GE çalışmalarında ön plana çıkmaktadır. Özellikle, ilaç sektöründe ve sigorta şirketleri için çok önemli olan veriler işlenerek piyasaya yön verebilmektedir (11;12). Hekimler ve potansiyel hastalar açısından ise teşhis, tanı ve tedavinin en hızlı ve güvenilir sonuçlarını elde edilmesi açısından bu alandaki büyük verilerin analizi hayati bir önem sahiptir.

Günümüzde, dünyadaki sağlık için yapılan harcamanın yaklaşık %80'i tedaviye yöneliktir (13). Sağlık alanındaki büyük veri analizi ile hastalıkların önlenmesi öngörülmektedir (14,15). Bu durumda tedavi aşamasına geçmeden, koruyucu hekimlik önemli hale gelecek ve doğru tedaviler için zaman kazanılmış olacaktır (16;13). Hatta kişiselleştirilmiş büyük verilerin analizleri yapılarak hastalığa ve hastaya göre en az yan etki ile en etkin tedavi sağlayabilecektir (14; 15). Araştırmada kullanılan örnek veri seti Türkiye'de üç bölgede yer alan (Marmara, Karadeniz ve İç Anadolu Bölgeleri) farklı yaşlardaki (günlük yaş) köpekleri kandaki Ortalama Eritrosit Hemoglobin Konsantrasyonu (MCHC) (g/dL)

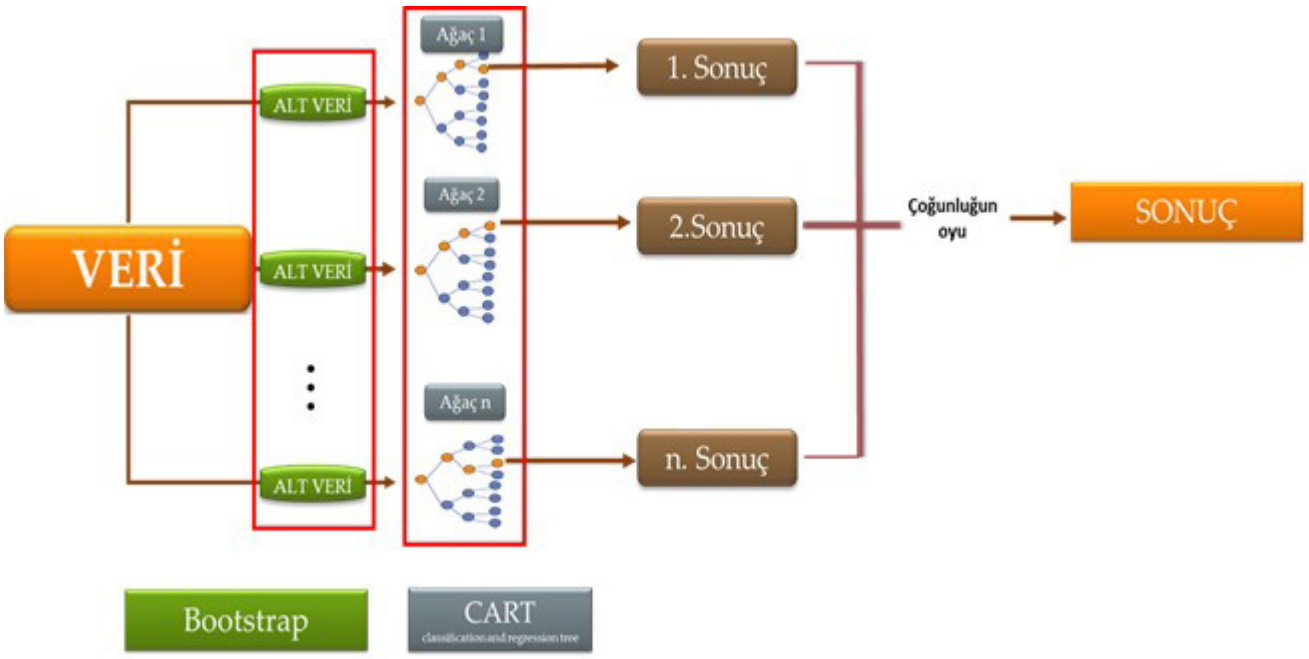
Platelets, trombosit (PLT) (x10³/µL) değerlerine göre hastalığa sebep olan parvovirüsün varlığı hakkında tahmin algoritmaları incelenmiştir.

Makine Öğrenmesi

Makine öğrenmesi, mevcut bilgilerden hareketle tutarlı ve en doğru tahminleri yapmaya sağlayan öngörücü bir disiplin olarak tanımlanır. Sağlıklı toplanmış büyük verileri tahmin algoritmaları ile işleyerek yeni veriler elde edilmesini sağlar (17). Mevcut verinin türü ve boyutu göz önünde bulundurularak, eldeki sorunu çözmek için uygun algoritmayı seçmek çok önemlidir. Yüksek kaliteli ve niceliğe sahip veri kümeleri, çoğunlukla ML modellerinin doğruluğunu artırabilir (18)

En yakın komşu algoritması**:** Çalışma mantığı, bir veri seti içerisinde bağımsız değişkenleri sınıflandırarak en yakın komşuların hangi sınıfta olduğunu tahmin etmeye dayanır (2). Araştırmada, bağımsız değişkenlerin CPV varlığı hakkındaki tahminlerinin doğruluk değerleri incelenmiştir.

Rastgele Orman Algoristması: Hem regresyon hem de sınıflandırma gerektiren verilerin analizinde kullanılır. Veriye ait her bir düğümden rastgele alınmış özelliklerin en iyisini seçerek elde edilen bütün düğümleri dallara ayırarak verileri işler (19). Karar ağaçlarının oluşumunda budama işleminin olmaması ve rastgele tercihi kullanmasından dolayı diğer pek çok algoritmaya göre daha doğru ve hızlı sonuç vermektedir (20). Özellikle karmaşık ve çok boyut içeren büyük verilerin analizinde iyi derecede performans göstermesinden dolayı, Rastgele Orman tercih edilen öğrenme algoritmalarından biridir (21;22). Bu özelliklerinden dolayı, araştırmada, kandaki farklı değerlerin, yaş ve yaşadıkları bölgeler ile ilişkileri CPV özelinde yorumlanmasında kullanılmıştır.



Şekil 1: Rastgele Orman Sınıflama Algoritması Çalışma Mantığı

LojistikRegresyon: Türkçe karşılığı istatistiksel analiz de olan yöntem, genellikle tahmine dayalı olasılıkları modelleme için kullanılan bir makine öğrenimi uygulamasıdır. Yöntemde, bağımlı değişken sonlu veya kategoriktir (22;19). Araştırmada kullanılan veri seti de kategorik ve sonlu olmasından dolayı bu yöntem tercih edilmiştir.

NaiveBayes sınıflandırma algoritmaları: Hızlı ve kolay uygulanabilir sınıflandırma yöntemleri arasında tercih edilen bir yöntemdir. Sınıflandırmanın temeli Bayes teoremine dayanmaktadır. Olasılık hesabının yapıldığı Bayes teorimi verilerin bağımsız olması koşulu ile daha iyi sonuçlar vermektedir. Farklı A ve B olayları için;

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

Bu durum aynı zamanda yöntemin en büyük dezavantajı olarak görülse de istatiksel bağımsızlık şartları esnetilerek karmaşık yapay sinir ağları gibi yöntemlerle karşılaştırılabilir sonuçlar elde edilebilmektedir. Araştırmada değişkenlerin bağımsız olup olmama durumuna göre sonuç tahmini incelenmiştir.

Sağlık verilerinin analizlerinin makine öğrenmesinin farklı yöntemleri ile son yıllarda yapılmış çalışmalar Tablo 1'degösterilmiştir.

Literatürde yer alan çalışmalar, hastalığın durumuna göre, araştırmacıların tercihine göre ve bağımsız değişkenlere göre farklı makine öğrenmesi yöntemleri gösterilmiştir. Bu araştırma ise literatürde sık karşılaştığımız dahası veri setimiz ile çalışmaya uygun farklı yöntemleri deneyerek sonuçları karşılaştırmayı hedefler.

Tablo 1: Sağlık Verileri ve ML hakkındaki çalışmalar. ML: Makine Öğrenmesi; SVM: Destek Vektör Makineleri; DT: Karar Ağaçları; NB:NaiveBayes; ANN: Yapay Sinir Ağları; LR: Lojistik Regresyon; RF: Rastagele Orman; KNN: k-En Yakın Komşu

Yazar/lar	Kullanılan Yöntem	Açıklama
Bollig N ve ark (2020)	SVM, RF, DT, Baging, Boosting, LR	Atopik dermatiti olan köpeklerde tedavi başarısını tahmin etmişlerdir.
Saberioon M ve ark (2018)	RF, SVM, LR, KNN.	Gökkuşağı Alabalığı diyetler türleri başarıları üzerinde bir sınıflandırma çalışması yapmışlardır.
Bollig N ve ark (2020)	LR, SVM, DT, RF, Boosting	Ölüm raporlarına göre ölüme neden olan hasatlıkların tahmini için Makine Öğrenmesi (ML) algoritmaları kullanılmıştır.
Liang R et al. (2020)	NB, ANN, SVM, RF, AdaBoost, C4.5	Afrika domuz hastalığını ML yöntemlerini kullanarak tahmin etmişlerdir.
Romero MP et al. (2021).	RF	İngiltere'deki sığır tüberkülozu hastalığı kontrolünü desteklemek için iki makine öğrenimi tahmin modelinin değerinin karşılaştırmışlardır.
Bates AJ ve Saldias B. (2019)	LR, KNN, DT, ANN, RF	İneklerde vücut kondisyon puanı ve başvuru oranı arasındaki ilişkinin ML ve lojistik regresyon kullanarak modellenmişlerdir.
Cihan P ve ark. (2019)	Literatür Bilgisi	Veterinerlik Alanında ML yöntemleri ile tanı teşhis belirleyen çalışmalar hakkında literatür taraması yapmışlardır.
Dilwani AAR. (2019)	LR, RF, Boosting, KNN	Acil servis hastalarını triyaj durumuna göre değerlendirmiştir
Bizal Ö. (2014)	NB, KNN, DT, LR, SVM	Parkinson hastalığının belirlenmesinde makine öğrenmesi tekniklerini kullanmışlardır.

Yöntem

Veri Toplama Araçları: Proje kapsamında kullanılan büyük veri seti (EK 1)için Selçuk Üniversitesi Veteriner Fakültesi Deney Hayvanları Üretim ve Araştırma Merkezi Etik Kurulu'ndan (SÜVDAMEK), 13.02.2018 gün ve 2018/14 sayılı karar ile izin alınmıştır

Kullanılacak terimler:

Tanı: Parvovirüs var / Parvovirüs yok

Merkez: Marmara Bölgesi / Karadeniz Bölgesi / İç Anadolu Bölgesi

Günlük Yaş: Köpeğin gün olarak yaşı

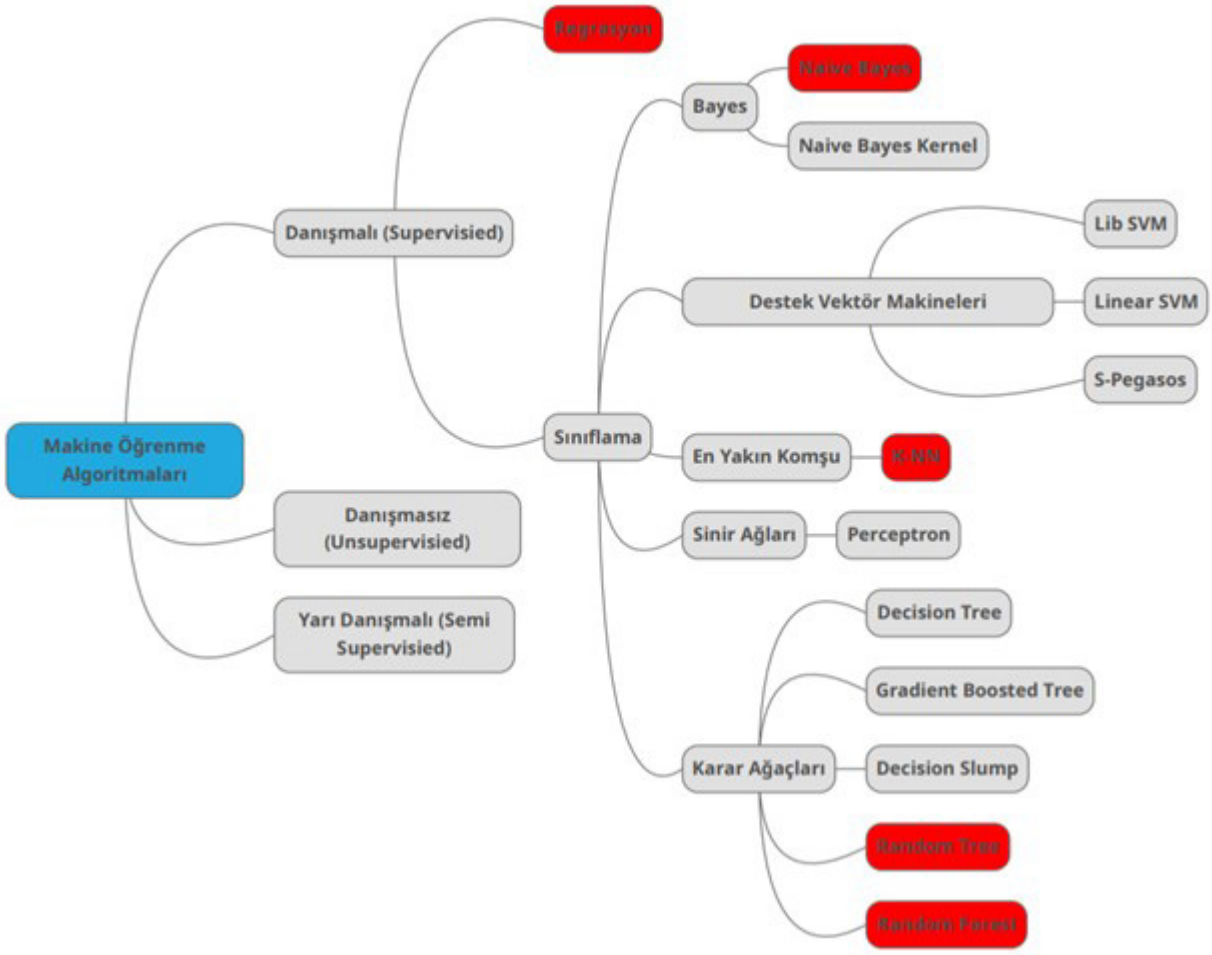
MCHC: Ortalama Eritrosit Hemoglobin Konsantrasyonu (g/dL)

PLT: Platelets, trombosit (x10³/µL)

Araştırma Yöntemi:

Araştırmada Veri Madenciliğinin alt çalışma alanlarından olan Makine öğrenmesinin En Yakın Komşu (KNN), Rastegele Orman,ve NaiveBayes sınıflandırma algoritmaları kullanılmıştır.

Tablo 1: Sağlık Verileri ve ML hakkındaki çalışmalar. ML: Makine Öğrenmesi; SVM: Destek Vektör Makineleri; DT: Karar Ağaçları; NB:NaiveBayes; ANN: Yapay Sinir Ağları; LR: Lojistik Regresyon; RF: Rastgele Orman; KNN: k-En Yakın Komşu

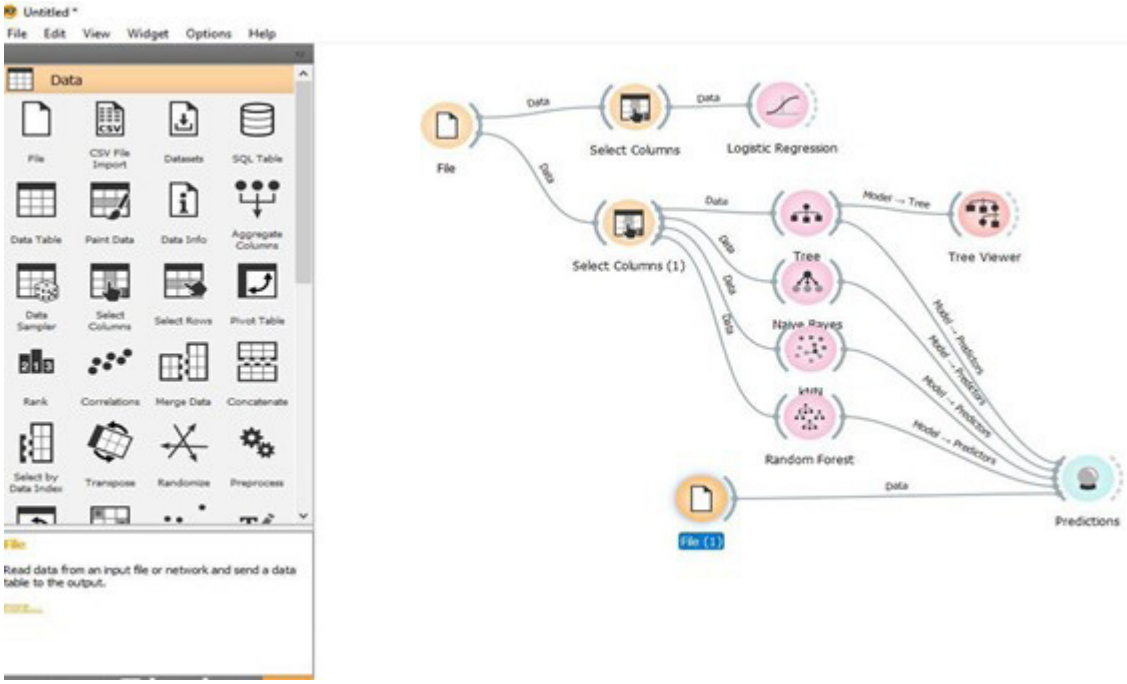


Araştırmada Orange Data Mining ve Jasp (JASP Team (2020). JASP (Version 0.14) [Computer software]) programı kullanılarak gerçekleştirilmiştir. Sonuçları daha iyi gözlemleyebilmek ve yöntem karşılaştırmalarının daha iyi görselleştirilmesi için araştırma ekibi tarafından tercih edilmiştir.

Bulgular

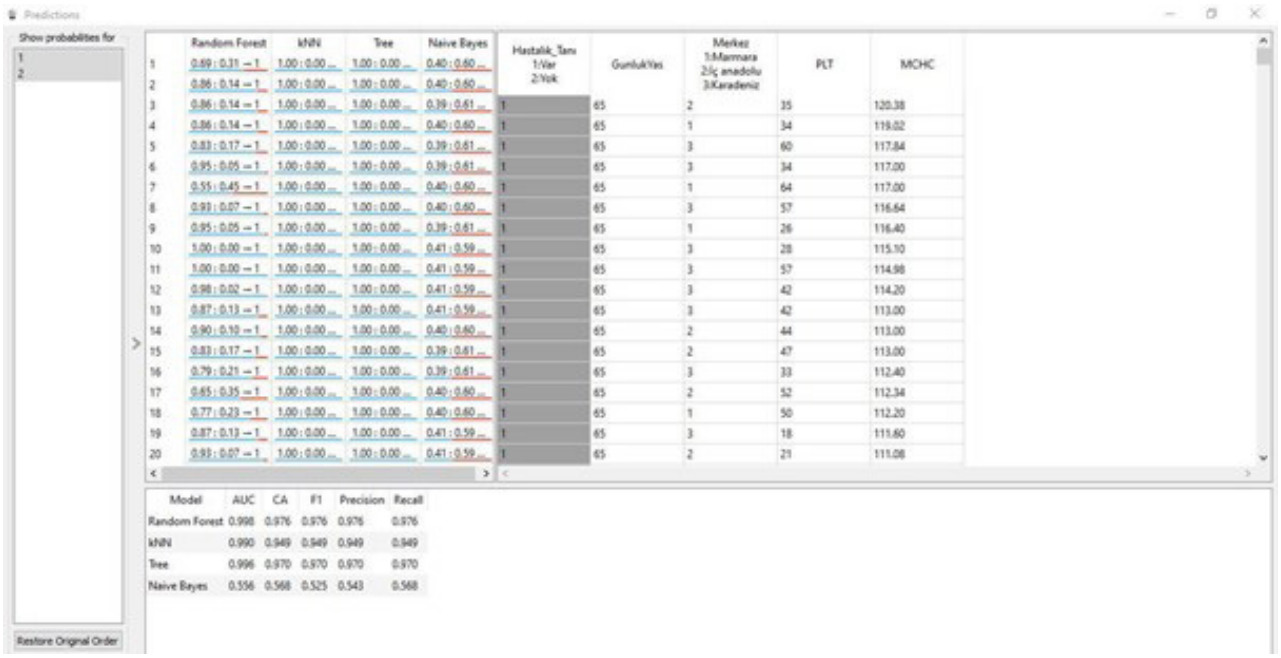
Araştırmada kullanılan veri seti öncelikle Orange Data Mining Programında incelenmiştir. Lojistik Regresyon, KNN, Rastgele Orman ve NaiveBayes Yöntemleri veri setinde uygulanmış (Tablo 2) ve doğruluk oranları incelenmiştir (Tablo 3)

Tablo 2: Orange Data Mining Programında Modellerin Uygulanması



Orange veri madenciliği programında modelin kurulması için öncelikle karar ağacı, NaiveBayes, KNN ve Rastgele Orman modelleri seçilerek ön görüsel bir modelleme kurulmuştur. Ayrıca sınıflandırma performansını doğrulayabilmek için lojistik regresyon modeli kurulmuştur.

Tablo 3:ML yöntemlerinin doğruluk oranları



Doğru sınıflandırma kriterlerine bakıldığında eğri altında kalan alan (AUC) değerlendirildiğinde %99,8 yüzdeyle en yüksek performansı Rastgele Orman modeli ardından %99,6 oranıyla karar ağacı modeli ve %99 başarıyla KNN yöntemi göstermiştir. NaiveBayes yöntemi ise öngörü olasılığının önsel bir olasılığa dayandırıldığı için %55,6 doğru sınıflandırma oranıyla diğer 3 modelden oldukça sınırlı şekilde çalışmıştır.

Bu sonuçlara göre parvovirüs olup olmama durumu, köpeklerin günlük yaşları, MCHC (g/dL) ve PLT (x10³/µL) he-

matolojik parametreleri ile Rastgele Orman, KNN ve karar ağaçlarıyla modelleriyle doğru bir kesinlikle sınıflandırılabilir.

Lojistik Regresyon Uygulaması

Veri setine lojistik regresyon uygulaması sonucunda elde edilen çıktılar değerlendirilmiştir.

Çizelge 1: Lojistik regresyon model değerlendirme

Model Summary - Hastalık_Tanı

Model	Deviance	AIC	BIC	df	X^2	p	McFadden R^2	Nagelkerke R^2	Tjur R^2	Cox & Snell R^2
H_0	60742.534	60744.534	60753.239	44591						
H_1	59067.591	59079.591	59131.823	44586	1674.943	< .001	0.028	1.000	0.037	1.000

Analiz gerçekleştirildikten sonra ilk olarak modelin anlamlılık değerleri değerlendirilir. Bu değerlendirmeler için McFadden R^2 ve TjurR^2 değerleri incelendiğinde çoklu regresyonda elde edilen R^2 'ye göre oldukça küçük değerler alma eğiliminde olup 1'e yaklaştıkça modelin açıklama gücü artarken 0'a yaklaştıkça azalır. Çıktılar sonucunda elde ettiğimiz değer McFaddenR^2 için 0,028 ve TjurR^2 için 0,037 olarak bulunmuş ve bu değer yeterli olduğu görülmüştür. NagelkerkeR^2 ve Cox&SnellR^2 sonuçlarına göre bağımsız değişkenler bağımlı değişkenin tamamını açıklamaktadır. Bu sonuçlar doğrultusunda modelin mükemmel uyum sahip olduğu söylenilir (=1674,943 p<0,001).

Çizelge 2: Lojistik regresyon model katsayı tahminleri

						Wald Test			95% Confidence interval	
	Estimate	Standard Error	Standardized*	Odds Ratio	z	Wald Statistic	df	p	Lower bound	Upper bound
(Intercept)	−0.179	0.093	0.837	0.836	−1.921	3.689	1	0.055	−0.362	0.004
GunlukYas	0.048	0.003	0.547	1.050	13.847	191.752	1	< .001	0.042	0.055
MCHC	−0.047	0.003	−0.648	0.954	−16.108	259.473	1	< .001	−0.053	−0.042
PLT	0.002	0.001	0.017	1.002	1.782	3.175	1	0.075	−0.000	0.003
Merkez (İç Anadolu Bölgesi)	−0.167	0.032	−0.167	0.846	−5.291	27.993	1	< .001	−0.229	−0.105
Merkez (Karadeniz Bölgesi)	−0.829	0.027	−0.829	0.437	−30.201	912.101	1	< .001	−0.882	−0.775

Note. Hastalık_Tanı level 'Yok' coded as class 1.

* Standardized estimates represent estimates where the continuous predictors are standardized (X-standardization).

Model için katsayı değerleri incelendiğinde Günlük yaş, MCHC ve Merkez değişkenleri model için anlamlı etkiye sahip iken PLT ve sabit (Intercept) değerlerinin model üzerinde anlamlı etkisi yoktur.

Günlük Yaşta bir birimlik artış parvovirüs bulunma olasılığını 1.050 kat arttırmaktadır. (%10). MCHC değerindeki bir birimlik artış parvovirüs bulunma olasılığını 0,954 kat azaltmaktadır. İç Anadolu'dan alınan örneklemdeki köpeklerin Marmara bölgesine göre parvovirüs bulunma olasılığını 0,846 kat daha azdır. Karadeniz bölgesinden alınan köpeklerin ise parvovirüs bulunma olasılığını 0,437 kat daha azdır.

Çizelge 3: Lojistik regresyon hata matrisi

	Predicted	
Observed	Var	Yok
Var	0.043	0.379
Yok	0.039	0.539

Çizelge 4: Lojistik regresyon performans metrikleri

Performance metrics

	Value
AUC	0.659
Sensitivity	0.940
Precision	0.582
F-measure	0.719

Oluşturulan model sonucunda tahmin değerleri ile gerçek değerlerin değerlendirmesi hata matrisi verilmiştir. Bu değerlendirme sonucunda CPV olan köpekleri %4,3 olmayan köpekleri ise %53,9 oranında doğru tahmin etmiştir. Toplamda model %58,2 doğru tahmin yaparken %41,8 yanlış tahmin yapmaktadır. Bu sonuçlar model anlamlı olsa da sonuçlarda büyük oranda hata olduğu görülmektedir.

Performans metrikleri değerlendirildiğinde eğri altında kalan alan 0,659, doğruluk 0,940, kesinlik 0,582 ve F ölçütü 0,719 olarak elde edilmiştir.

Rastgele Orman Algoritması Uygulaması

Veri setinde Rastgele Orman algoritması uygulaması sonucunda elde edilen çıktılar değerlendirilmiştir.

Çizelge 5: Rastgele Orman model değerlendirme

Random Forest Classification ▼

Trees	Predictors per split	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy	OOB Accuracy
96	2	28539	7135	8918	0.959	0.955	0.963

Note. The model is optimized with respect to the *out-of-bag accuracy*.

Modelde 100 ağaç üretilmiş ve bu ağaçlardan 96. ağacın en uygun sonucu verdiği tespit edilmiştir. Bu ağaçtaki doğrulama setindeki açıklayıcı değişkenler ile hedef değişken arasında uyum %96, test setindeki uyum %95 ve modelin test ve eğitim seti arasındaki doğruluk oranı %96 olarak bulunmuştur. Bu sonuçlara göre eğitim veri seti ile test veri seti arasında yüksek bir doğruluk oranı elde edilmiş ve modelimiz iyi bir uyuma sahip olduğu söylenebilir.

Çizelge 6: Rastgele Orman model hata matrisi

		Predicted	
		Var	Yok
Observed	Var	0.4	0.03
	Yok	0.02	0.56

Oluşturulan model sonucunda tahmin değerleri ile gerçek değerlerin değerlendirmesi hata matrisinde verilmiştir. Bu değerlendirme sonucunda CPV olan köpekleri %40 olmayan köpekleri işe %56 oranında doğru tahmin etmiştir. Toplamda model %96 doğru tahmin yaparken %4 yanlış tahmin yapmaktadır. Bu sonuçlar Rastgele Orman algoritmasının mükemmel bir doğrulukla tahmin yaptığı tespit edilmiştir.

Performans metrikleri değerlendirildiğinde eğri altında kalan alan 0,987, doğruluk 0,954, kesinlik 0,954 ve F ölçütü 0,954 olarak elde edilmiştir.

Doğruluktaki ortalama düşüş (Meandecrease in accuracy), her bir özelliğin Rastgele Orman modelinin doğruluğu üzerindeki etkisini ölçmek için değerlendirildi. Bunun sonucunda Rastgele Orman modelinin doğruluğu için en önemli değişken RDW olarak belirlendi. Sınıf ayırma önemi (Total increase in nodepurity) ise hangi değişkenin sınıfları belirlemede daha önemli olduğunu gösterir. Bunun sonucunda sınıf belirlemedeki en önemli değişken de RDW olarak belirlendi.

Çizelge 7: Rastgele Orman metrikleri

Evaluation Metrics

	Precision	Recall	F1 Score	Support	AUC
Var	0.951	0.939	0.945	3784	0.987
Yok	0.956	0.965	0.960	5134	0.987
Average / Total	0.954	0.954	0.954	8918	0.987

Note. Area Under Curve (AUC) is calculated for every class against all other classes.

Çizelge 8: Rastgele Orman değişken önemleri

Variable Importance ▼

	Mean decrease in accuracy	Total increase in node purity
RDW	0.345	0.390
HCT	0.261	0.329
Şehir	0.010	0.008
MCHC	0.003	0.003

KNN Algoritması Uygulaması

KNN algoritması uygulaması sonucunda elde edilen çıktılar değerlendirilmiştir.

Veri setinde 44592 köpek değerlendirilmiştir. Bu veri setinden Hold-out yöntemi ile %20 (8918) test veri seti için rastgele örneklem seçilmiştir. 44592 köpekten %20'si test veri setinin kullanıldığı için kalan %80 (35674) kişiden %20 (7135) oranında örneklem alınarak bu örneklemi de doğrulama için kullanılmıştır. Kalan veri (28539) eğitim için kullanılmıştır.

Çizelge 9: KNN model değerlendirme

K-Nearest Neighbors Classification

Nearest neighbors	Weights	Distance	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy
2	rectangular	Euclidean	28539	7135	8918	0.732	0.847

Note. The model is optimized with respect to the *validation set accuracy*.

Çizelge 10: KNN model hata matrisi

		Predicted	
		Var	Yok
Observed	Var	0.34	0.08
	Yok	0.07	0.51

Oluşturulan model sonucunda tahmin değerleri ile gerçek değerlerin değerlendirmesi hata matrisi verilmiştir. Bu değerlendirme sonucunda CPV olan köpekleri %34 olmayan köpekleri işe %51 oranında doğru tahmin etmiştir. Toplamda model %85 doğru tahmin yaparken %15 yanlış tahmin yapmaktadır. Bu sonuçlar En Yakın Komşu algoritmasının iyi bir doğrulukla tahmin yaptığı tespit edilmiştir.

Çizelge 11: KNN performans metrikleri

Evaluation Metrics

	Precision	Recall	F1 Score	Support	AUC
Var	0.823	0.807	0.815	3721	0.865
Yok	0.864	0.876	0.870	5197	0.865
Average / Total	0.847	0.847	0.847	8918	0.865

Note. Area Under Curve (AUC) is calculated for every class against all other classes.

Tablo 4: LR, RF ve KNN değerleri

	PLT Değişkeni	Toplam Doğruluk (%)	AUC	Duyarlılık	Kesinlik	F ölçütü
Lojistik Regresyon (LR)	Var	58,2	0,659	0,94	0,582	0,719
	Yok	57,6	0,679	0,933	0,583	0,717
Random Forest (RF)	Var	96	0,99	0,955	0,955	0,955
	Yok	68	0,626	0,688	0,699	0,69
KNN	Var	85	0,847	0,853	0,853	0,853
	Yok	97	0,994	0,964	0,965	0,965

Performans metrikleri değerlendirildiğinde eğri altında kalan alan 0,865, doğruluk 0,847, kesinlik 0,847 ve F ölçütü 0,847 olarak elde edilmiştir.

Toplam doğruluk değerleri incelendiğinde anlamsız değişken modelde çıkarıldığında lojistik regresyon ve RF yöntemlerinde doğruluk oranları düşmüştür.

RF yöntemi PLT değişkeni modelde iken en iyi tahminleri yapmıştır. Bu değişkeni modelden çıkarmak istemediğimiz durumlarda bize çok verimli sonuçlar verebilmektedir.

KNN yöntemi değişken sayısı azaldığında daha iyi sonuçlar vermektedir. Bu sebeple değişken sayısını azaltmak istediğimizde bu yöntemi kullanabiliriz.

NaiveBayes modeli önsel olasılığın doğru kesinlikle sınıflandırılamadığı için değerlendirilmemiştir.

Sonuç ve Tartışma

Sonuç olarak, toplam doğruluk değerleri incelendiğinde anlamsız değişken modelde çıkarıldığında lojistik regresyon ve RF yöntemlerinde doğruluk oranları düşmüştür. RF yöntemi PLT değişkeni modelde iken en iyi tahminleri yapmıştır. Bu değişkeni modelden çıkarmak istemediğimiz durumlarda bize çok verimli sonuçlar verebilmektedir. KNN yöntemi değişken sayısı azaldığında daha iyi sonuçlar vermektedir. Bu sebeple değişken sayısını azaltmak istediğimizde bu yöntemi kullanabiliriz.

Bu çalışmada, tanı testlerinin performanslarını değerlendirirken duyarlılık, seçicilik ve Roc eğrisi altında kalan alan önemli ölçütler olmuştur. Uygun kesim noktasının AUC değerlerine bağlı olduğu görülmüştür. Duyarlılık ve seçicilik ne kadar yüksek ise AUC değerleri de o kadar yüksek olmuştur. Bunun sonucunda yüksek AUC değerleri de uygun kesim noktasını vermiştir. Bu çalışma ile, parvoviralenterit hastalığının teşhisi için, uygulanan istatistiksel hesaplamalar ile modellendiği ortaya konmuştur. Prognoz sürecinde deneğin MCHC, Platelet ve hangi bölgeden olduğu bilindiğinde parvoviralenterit hastalığının teşhisi için Rastgele Orman, KNN ve karar ağaçları modelleriyle doğru bir seçim yapılabilir. Sonuçlar, ileride yapılacak bölgede/ülkede yayılım gösteren CPV suşlarını içeren aşıların üretilmesi ve CPV'nin antijenik varyantlarının belirlenmesine dönük moleküler karakterizasyon çalışmalarına kaynak olacağı düşünülmektedir. Teşhis koymaya çalışan hekimler, bu değişkenler ile elde edilen sınıflandırma modellerini kullanarak yararlanabilirler. Sonuç olarak; köpeklerin parvoviralenterit hastalığının teşhisinde kan hemogram değerlerinden MCHC, Platelet ve yaş parametrelerinin kesim noktaları göz önünde bulundurularak büyük doğrulukla karar vermede yardımcı olacağı ortaya çıkmıştır. İleride yapılacak çalışmalara da kaynak olabilecektir

KAYNAKLAR

1. Karim, M. R., Beyan, O., Zappa, A., Costa, I. G., Rebholz-Schuhmann, D., Cochez, M., &Decker, S. Deeplearning-basedclusteringapproachesforbioinformatics. *Briefings in Bioinformatics*, (2021), *22*(1), 393-415.
2. Li, H.,Xue, Y., &Zeng, X. Investigation of dataminingtechniqueandartificialintelligencealgorithm in microflorabioinformatics. In *E3S Web of Conferences* (2021). (Vol. 267, p. 01040). EDP Sciences.
3. De Mauro, A.,Greco, M., &Grimaldi, M. What is bigdata? A consensualdefinitionand a review of keyresearchtopics. In *AIP conferenceproceedings* (2015), (Vol. 1644, No. 1, pp. 97-104). AmericanInstitute of Physics.
4. Pathak, R. K., Singh, D. B., & Singh, R. Introductiontobasics of bioinformatics. (2022).In *Bioinformatics* (pp. 1-15). AcademicPress.
5. Ebrahimi, F.,Asemi, A., Shabani, A., &Nezarat, A. Developing a Prediction Model for Author Collaboration in Bioinformatics-Research Using GraphMiningTechniquesandBig Data Applications. *International Journal of Information Scienceand Management* (2021). *(IJISM)*, *19*(2), 1-18.
6. Sakulwira, K.,Vanapongtipagorn, P., Theamboonlers, A., Oraveerakul, K., &Poovorawan, Y. Prevalence of canine coronavirusandparvovirusinfections in dogswithgastroenteritis in Thailand. *VeterinarniMedicina*, (2003).*48*(6), 163..
7. Filipov, C.,Decaro, N., Desario, C., Amorisco, F., Sciarretta, R., &Buonavoglia, C. . Canine parvovirusepidemiology in Bulgaria. *Journal of VeterinaryDiagnosticInvestigation*, (2011), *23*(1), 152-154.

8. ÖZKUL, A., KELEŞ, İ., KARAOĞLU, T., ÇABALAR, M., & BURGU, İ. Detectionandrflpanalysis of canine parvovirus (cpv) dnabypolymerasechainreaction (pcr) in a dog. *TurkishJournal of VeterinaryandAnimalSciences*, (2002). *26*(5), 1201-1203.
9. Sellers, R. F., & Pedgley, D. E. (1985). Possible windborne spread to western Turkey of bluetongue virus in 1977 and of Akabane virus in 1979. *Epidemiology & Infection*, *95*(1), 149-158.
10. Sterzing, F., Kratochwil, C., Fiedler, H., Katayama, S., Habl, G., Kopka, K., ... & Giesel, F. L. (2016). 68Ga-PSMA-11 PET/CT: a new technique with high potential for the radiotherapeutic management of prostate cancer patients. *European journal of nuclear medicine and molecular imaging*, *43*(1), 34-41.
11. Khan, N. T. (2018). Data Mining–Basics of Bioinformatics. *Transcriptomics*, *6*(142), 2.
12. Mahapatro, P. S. AssociationRuleMining in HealthCare: A Study. Studies in IndianPlaceNames, (2020). 40(53), 87–91.
13. Dinov, I. D. Volume andvalue of bighealthcaredata. *Journal of medicalstatisticsandinformatics*, (2016). *4*.
14. Polat, M., & KARAHAN, A. (2009). Multidisipliner yeni bir bilim dalı: biyoinformatik ve tıpta uygulamaları. *SDÜ Tıp Fakültesi Dergisi*, *16*(3), 41-50.
15. Snyder, L. V., Atan, Z., Peng, P., Rong, Y., Schmitt, A. J., & Sinsoysal, B. (2016). OR/MS models for supply chain disruptions: A review. *Iie Transactions*, *48*(2), 89-109.
16. Zeynep, ÖZEL ve DEMİRSÖZ, M. . Makine Öğrenmesi Yöntemleri İle Covid-19 Verilerinin İncelenmesi: Türkiye Örneği: An Analysis of Covid-19 Data With Machine Learning Methods: The Case of Turkey. *Sağlık Bilimlerinde Yapay Zeka Dergisi (Journal of ArtificialIntelligence in HealthSciences) ISSN: 2757-9646*, *1*(2), 1-7.(2021)
17. Kochetkova, O. V.,&Shiryaeva, E. V. Perspectivearchitecture of dairyfarmingenterprises, using modern digitaltechnologiesforsustainabledevelopment. In *IOP Conference Series: Earth andEnvironmentalScience* (2022), (Vol. 965, No. 1, p. 012062). IOP Publishing.
18. Ghaffarian, S.,van der Voort, M., Valente, J., Tekinerdogan, B., & de Mey, Y. . Machine learning-basedfarm risk management: A systematicmappingreview. *ComputersandElectronics in Agriculture*, (2022), *192*, 106631.
19. Bollig N, DeBoer D, Döpfer D Learning Tutorial for Veterinarians: Examples Using Canine Atopic Dermatitis, (2020)
20. Yazdanbakhsh, O.,Zhou, Y., &Dick, S. An intelligentsystemforlivestockdiseasesurveillance. *Information Sciences*, (2017).*378*, 26-47.
21. Kılınçalp, S., Ekiz, F., Başar, Ö., Ayte, M. R., Çoban, Ş., Yılmaz, B., ... & Yüksel, O. (2014). Mean platelet volume could be possible biomarker in early diagnosis and monitoring of gastric cancer. *Platelets*, *25*(8), 592-594.
22. Saberioon M, Císař P, Labbé L, Souček P, Pelissier, P ve Kerneis T. Comparativeperformanceanalysis of supportvectormachine, randomforest, logisticregressionand k-nearestneighbours in rainbowtrout (oncorhynchusmykiss) classificationusingimage-basedfeatures. Sensors,(2018). 18(4), 1027.