Veri Madenciliği: Hastalık Tahminlerinde Sınıflandırma Algoritmalarının Karşılaştırılması

Nazlı Nazan AVCI[1]

[1] Bilgisayar Mühendisliği, Mehmet Akif Ersoy Üniversitesi, Burdur, Türkiye
naznazanavci@gmail.com

Özet: Son yıllarda, bilgi ve bilişim teknolojilerindeki yenilikler ve gelişmeler, analiz edilmesi gereken veri miktarını önemli ölçüde artırmıştır. Veri Madenciliği bir çok alanda olduğu gibi sağlık verilerinde de kullanılmaktadır. Veri madenciliği yöntemlerinden biri olan sınıflandırma, en çok kullanılan veri madenciliği yöntemidir. Ulaşılmak istenen bilginin elde edilmesi için uygulanan sınıflandırma algoritmaları, içerdiği verinin ortak özelliğine göre veri setinin belirli sınıflara ayrılmasını sağlamaktadırlar. Bu işlemin ardından bir sınıflandırma modeli elde edilir. Elde edilen sınıflandırma modeli veri seti üzerinde uygulanarak, model ile belirlenmiş olan sınıfların veri seti içindeki benzerliği araştırılır. Yapılan çalışmada Veri Madenciliğinde sınıflandırma süreci ele alınarak, k-en yakın komşu, naive bayes, karar ağaçları, rastgele ormanlar, destek vektör makineleri ve lojistik regresyon isimli altı farklı sınıflandırma algoritması ile üç veri seti üzerinde farklı uygulamalar gerçekleştirilmiştir. Sınıflandırma çalışmalarında en önemli kriter yüksek başarımlı bir sınıflandırıcı model oluşturabilmektir. Bu amaçla Python dili ile uygulamalar yapılarak sınıflandırma modelinin tahmin değerlerinin doğruluğuyla ilgili performans ölçüm değerleri elde edilmiştir. Bu makalenin amacı, sağlıkta Veri Madenciliğinin kullanımını yaygınlaştırmak ve sağlık profesyonellerine sağlık sektöründe Veri Madenciliğinin kullanımı ile ilgili örnekler sunarak karar verme süreçleri açısından yeni bir bakış açısı kazandırmaktır.

Anahtar Kelimeler- Veri Madenciliği, Sınıflandırma, Lojistik Regresyon, Naive Bayes, K-En Yakın Komşu, Karar Ağaçları, Rastgele Orman, Karışıklık Matrisi

Abstract: In recent years, innovations and developments in information and information technologies have significantly increased the amount of data that needs to be analyzed. Data Mining is used in health data as well as in many other fields. Classification, one of the data mining methods, is the most widely used data mining method. The classification algorithms applied to obtain the desired information enable the data set to be divided into certain classes according to the common feature of the data it contains. After this process, a classification model is obtained. By applying the obtained classification model on the data set, the similarity of the classes determined by the model in the data set is investigated. In the study, by considering the classification process in Data Mining, different applications were carried out on three data sets with six different classification algorithms named k-nearest neighbor, naive bayes, decision trees, random forests, support vector machines and logistic regression. The most important criterion in classification studies is to create a high-performance classifier model. For this purpose, performance measurement values related to the accuracy of the prediction values of the classification model were obtained by making applications with the Python language. The aim of this article is to popularize the use of Data Mining in healthcare and to provide healthcare professionals with a new perspective in terms of decision-making processes by presenting examples of the use of Data Mining in the healthcare sector.

Key Words- Data Mining, Classification, Logistic Regression, Naive Bayes, K-Nearest Neighbor, Decision Trees, Random Forest, Confusion Matrix

1.Giriş

Hastalıklar yaşamı olumsuz etkileyen etmenlerdir. Hastalıkların erken teşhisi Tıp bilimi ve insan hayatı için elzemdir. Özellikle kanser gibi bazı hastalıkların önceden tespiti tedavinin olumlu sonuç vermesi için şarttır. Böyle hayati hastalıkların verilere göre sınıflandırma ile önceden tahmin edilmesi önemlidir. Hastalıkların teşhisi işlemi her hastalığın ayrıca incelenmesi sonucunda ortaya çıkaracaktır. Pek çok hastalığın tek tek incelenerek sınıflandırılması zaman ve maliyet açısından uygun olmayan bir durumdur. Bu işlem için veriler toplanarak bilgisayar ortamında otomatik olarak teşhis edilmesi ve sınıflandırılmanın yapılması zaman ve maliyet açısından oldukça kolaylık sağlayacaktır. Sosyal medya üzerinden yapılan yorumlar, hastanelerde yapılan testler, günlük çekilen fotoğraf ve videolar gibi oluşan çok büyük veriden daha sonra makine öğrenmesi modellerinde kullanılmak üzere veri tabanları oluşturulmaktadır. Bu büyük miktardaki ham verinin işlenerek yararlı bilgiye dönüştürülmesi için veri madenciliği tekniklerinden faydalanılmaktadır. Veri Madenciliği, pek çok analiz aracı kullanımıyla veri içerisinde örüntü ve ilişkileri keşfederek, bunları geçerli tahminler yapmak için kullanan bir süreçtir. Makine öğrenmesi modeli oluşturulurken öğrenmenin ne oranda gerçekleştiği yani modelin ne kadar doğru tahminlerde bulunduğu büyük öneme sahiptir. Model oluşturulurken karar sınıfına ait azınlık veri gurubu eğitim veri seti içerisinde düşük miktarda ya da hiç yer almayacaktır. Bu da çoğu zaman modelin performansını ölçerken yanılgıya düşülmesine sebep olmaktadır.[1] Önemli olan geçmişe ait olaylara dair gizli bilgilerin keşfedilmesi, ileriye yönelik durumsal öngörüler veren modeller ile önceden tedbir almamızı sağlayacak bir yönetim anlayışına geçmek ve olası kayıpları öngörebilmektir.

Sağlık politika ve kararlarının amaçlara uygun ve etkin olabilmesi güvenilir, güncel ve doğru veriye bağlıdır. Sağlık bilgi sistemlerinin amacı büyük miktardaki sağlık verilerinden faydalı bilgi üretmektir. Bu bilgiler hasta düzeyinde daha iyi sağlık hizmeti sunumu, sağlık kurumlarının daha iyi yönetilmesi, kaynakların etkin kullanımı ve sağlık politikalarının oluşturulması amaçları ile kullanılmaktadır. [2]

Sağlık alanında ve sınıflandırma algoritmalarında literatür taraması yapıldığında konu ile ilgili yapılan çalışmaların sayısında son yıllarda bariz bir artış tespit edilmiştir. İlk olarak MEDLINE veri tabanında bulunan 25962 makale özeti ile yapılan çalışmanın genel amacı, öncelikle metin madenciliği yöntemleriyle kanser üzerine yazılmış makaleleri birbirinden otomatik sınıflandırma amaçlı geliştirilmiş metodun başarımını analiz etmektir.[3] R Dili ile Bir Uygulama Sınıflandırma Yöntemi çalışmasında Sınıflandırma Yöntemiyle model geliştirilmiştir. Başlıca algoritmalar, entropi tabanlı sınıflandırma (C4.5 algoritması, C5.0 algoritması) ve Karar Ağaçları, k-en yakın komşu algoritması, Rastgele Orman vb. şeklindedir. Elde edilen sonuçlar üzerinde karışıklık matrisi doğrultusunda performans değerlendirmesi yapılmıştır. En iyi performans ölçüm değeri C5.0 algoritması olarak sonuçlanmıştır.[4] Glokom hastalığının göz sinirleri zarar görmeden önce teşhis edilebilmesi çalışmasında üç farklı makine öğrenmesi modeli kullanmış ve destek vektör makineleri ile oluşturulan modelin en iyi performansı elde ettiği sonucuna ulaşmıştır.[5] Sınıflandırma İlişkisinde Performansa Etki Eden Faktörlerin Değerlendirilmesinde, diğer çalışmalardan farklı olarak farklı dağılımlara sahip 32 farklı veri seti incelenerek dokuz farklı sınıflandırma algoritması uygulanmıştır. Bu sınıflandırma algoritmaları; k-en yakın komşu, karar ağaçları, rasgele orman, naive bayes, lojistik regresyon, destek vektör makineleri, gradyan artırma, adaboost ve yapay sinir ağlarıdır. Daha sonra ise karışıklık matrisinden faydalanılarak doğruluk, F1 ölçütü ve alıcı işlem karakteristikleri altında kalan alan metrikleri kullanılmıştır. Ayrıca bu metriklerin ne ifade ettiği, temel düzeyde belirtilmiştir.[1] Kalp ameliyatı sırasında hastaya ait hayati riskin makine öğrenmesi yöntemleriyle belirlenmesi çalışmasında dört farklı model kulanmış ve farklı metrikler kullanarak karar ağacı modelinin bu veri seti için en iyi model olduğunu tespit etmiştir .[6] Başka bir çalışmada 108 kişiden alınan sol el parmak görüntülerinden kişilerin 2D:4D parmak oran bilgileri bir veri madenciliği aracı olan WEKA yazılım ortamında işlenmiştir. Kişilerin demografik bilgileri ve parmak oranları dikkate alınarak, karar ağaçlarının J48, NBTree, LADTree sınıflandırma algoritmaları ile bireylerin kişilik durumlarına göre sınıflandırılma yapılmıştır. Uygulama sonucunda en başarılı sınıflandırma % 100 oranla LADTree algoritması ile gerçekleştirilmiştir. % 97.222 oranla NBTree ve % 94.444 oranla J48 algoritması başarılı bir sınıflandırma gerçekleştirmiştir.[7] Bitki yapraklarının sınıflandırılması

çalışmasında bilgisayarlı görü işlemi sınıflandırma ve tanımlama işlemlerinde oldukça iyi sonuçlar vermektedir . Yaprak sınıflandırma işlemi içinde bilgisayarlı görü sistemi kullanılmış ve oldukça iyi sonuçlar vermiştir.[8]

Bu çalışmalara bakıldığında genelde tek veri seti ile inceleme yapıldığı ve başarım kriteri olarak da genellikle doğruluk metriği kullanıldığı gözlemlenmiştir. Yapılan çalışmalarda ise sonuçlar sayısal olarak verilmekte, sonuçlar hakkında yeterince değerlendirme yapılmadığından dolayı sonuçların anlamı yetersizdir.

Bu makalenin amacı, sağlık profesyonellerine sağlık sektöründe Veri madenciliğinin kullanımı ile ilgili örnekler sunarak karar verme süreçleri açısından yeni bir bakış açısı kazandırmaktır. Bundan dolayı, bu çalışmada sınıflandırma yöntemi olan karar ağaçları, rastgele ormanlar, naive bayes, k-en yakın komşu, destek vektör makineleri ve lojistik regresyon yöntemleri kullanılmıştır. Yöntemin uygulanma aşamasında veri setinin belirli bir oranı rastgele seçilerek eğitim veri seti olarak ele alınmış ve kalan kısmı ise test veri seti olarak değerlendirilmiştir. Söz konusu veri seti üzerinde sınıflandırma algoritmalarından biri uygulanarak bir sınıflandırma modeli oluşturulur ve ardından test veri setindeki örnekler için sınıf tahmini yapılarak, modelin tahmin sonucunun performans ölçüm değeri elde edilir. Elde edilen sonuçlar üzerinde bütün algoritmalar için elde edilen karışıklık matrisi, doğruluk ve F1 skor doğrultusunda performans değerlendirmesi yapılmıştır. Söz konusu algoritmalara ait karışıklık matrisi değerlerinin hesaplanma yöntemi ifade edilmiş ve bu algoritmaların doğruluk değerleri karşılaştırılmıştır.

2. Yöntem

Bu çalışmada, literatürde yaygın olarak kullanılan makine öğrenmesi yöntemleri kullanılarak sınıflandırma çalışması yapılmıştır. Makine öğrenmesi uygulamalarında oluşturulan modelin doğruluğunun test edilmesi büyük önem arz etmektedir. Modelin performansını sadece doğruluk metriği ile ölçmek çoğu zaman yanılgıya yol açabilmektedir. Bundan dolayı oluşturulan modelin performansının değerlendirilmesinde sadece doğruluk metriği değil buna ek olarak kullanılacak metriklere ihtiyaç duyulmaktadır. Makine öğrenmesi yöntemlerinden olan sınıflandırmada bu değerlendirme için karışıklık matrisi olarak isimlendirilen bir tablodan faydalanılmaktadır. Ayrıca sınıflandırma raporunda F1 skor, geri çağırma, duyarlılık gibi bilgilerin özetleri sunulmuştur. Bu çalışmada kullanılan yöntemlere ait bilgiler verilmiştir.

2.1.Veri Madenciliği Yöntemleri: Verilerin toplanıp önişlemden geçirilmesi, bu işlemden geçirilen verilerden anlamlı bilgilerin çıkartılması ve sonuçların doğrulandıktan sonra sunulmasına "bilgi keşfi" denir. Veri madenciliği, veriden gizli, anlamlı ve potansiyel olarak değerli örüntüleri ortaya çıkaran bir dizi yöntemlerdir. Veri madenciliği yöntemleri; sınıflandırma, kümeleme ve birliktelik kuralları analizi olmak üzere üç ana başlık altında toplanmaktadır. Sınıflandırma, sonuçları bilinen veriler ile bir sınıflandırma modeli oluşturulması ve bu model kullanılarak yeni gelen örneklem verinin daha önceden belirlenmiş sınıflardan birine dahil edilmesi işlemidir. Kümeleme, bir veri seti içerisindeki birbirine benzer nesneleri gruplayan bir tekniktir. Birliktelik kuralları analizi ise, verilerde yer alan öğeler arasındaki ilginç bağıntıların ve ilişkilerin kurallar halinde bulunması işlemidir.[9] Öğrenme problemleri genel olarak denetimli ve denetimsiz olarak ikiye ayrılır. Ana araştırma alanlarından biri, bilgisayar programlarının karmaşık kalıpları otomatik olarak tanımayı öğrenmesi ve verilere dayalı akıllı kararlar vermesidir. Burada, veri madenciliği ile oldukça ilgili olan makine öğrenimindeki denetimli öğrenme temelde sınıflandırma ile eşanlamlıdır. Öğrenmedeki denetim, eğitim veri setindeki etiketli örneklerden gelir. Denetimli öğrenme, girdi ölçülerinin sayısını temel alarak çıktı ölçüsünün değerini tahmin eder; denetimsiz öğrenmede ise çıktı ölçüsü yoktur ve hedef girdi ölçüleri kümesi arasındaki birliktelik ve örüntüleri betimler.[2] Bu çalışmada Sınıflandırma yöntemi kullanılmıştır.

2.2.Veri Önişleme: Sınıflandırma algoritmalarını kullanmak için verinin hazırlanması gerekir. Bunun için veri önişlemleri dediğimiz bir takım işlemler uygulanır.

- ✓ Veri temizleme (gürültüyü ve tutarsız verileri ortadan kaldırmak için)
- ✓ Veri entegrasyonu (birden çok veri kaynağının birleştirilebildiği durumlarda)
- ✓ Veri seçimi

✓ Veri dönüştürme (verilerin dönüştürüldüğü ve formlara birleştirildiği yer özet için uygun)

Veri kalitesi, doğruluk, tamlık, tutarlılık, zamanlılık, inanılırlık ve yorumlanma açısından tanımlanır. Bu nitelikler, verilerin kullanım amacına göre değerlendirilir. Veri temizleme rutinleri, eksik değerleri doldurmaya, aykırı değerleri belirlerken gürültüyü düzeltmeye ve verilerdeki tutarsızlıkları düzeltmeye çalışır. Veri temizleme, genellikle tutarsızlık tespiti ve veri dönüştürmeden oluşan yinelemeli iki aşamalı bir süreç olarak gerçekleştirilir. Veri entegrasyonu, tutarlı bir veri deposu oluşturmak için birden çok kaynaktan gelen verileri birleştirir. Veri azaltma teknikleri, bilgi içeriği kaybını en aza indirirken verilerin daha az temsil edilmesini sağlar. Bunlar, boyut azaltma, sayısal azaltma ve veri sıkıştırma yöntemlerini içerir. Boyutsal küçülme ele alınan rastgele değişkenlerin veya niteliklerin sayısını azaltır. Yöntemler, dalgacık dönüşümlerini, temel bileşenler analizini, öznitelik alt kümesi seçimini ve öznitelik oluşturmayı içerir. Sayısal azaltma yöntemleri, orijinal verilerin daha küçük temsillerini elde etmek için parametrik veya parametrik olmayan metrik modeller kullanır. Parametrik modeller, gerçek veriler yerine yalnızca model parametrelerini depolar. Örnekler, regresyon ve log-lineer modelleri içerir. Parametrik olmayan yöntemler arasında histogramlar, kümeleme, örnekleme ve veri küpü toplama yer alır. Veri sıkıştırma yöntemleri, orijinal verilerin azaltılmış veya "sıkıştırılmış" bir temsilini elde etmek için dönüşümler uygular. Orijinal veriler sıkıştırılmış verilerden herhangi bir bilgi kaybı olmadan yeniden oluşturulabiliyorsa, veri azaltma kayıpsızdır; aksi halde kayıptır. Veri dönüştürme rutinleri, verileri madencilik için uygun biçimlere dönüştürür. Örneğin, normalleştirmede öznitelik verileri, 0.0 ila 1.0 gibi küçük bir aralığa düşecek şekilde ölçeklenir. Diğer örnekler, veri ayrıklaştırması ve kavram hiyerarşisi oluşturmadır. Veri ayrıklaştırma, değerleri aralık veya kavram etiketlerine eşleyerek sayısal verileri dönüştürür. Bu tür yöntemler, birden fazla ayrıntı düzeyinde madenciliğe izin veren veriler için otomatik olarak kavram hiyerarşileri oluşturmak için kullanılabilir. Nominal veriler için, şema tanımlarına ve her bir veri için ayrı değerlerin sayısına dayalı olarak kavram hiyerarşileri oluşturulabilir. Çok sayıda veri ön işleme yöntemi geliştirilmiş olmasına rağmen, büyük miktarda tutarsız veya kirli veri ve sorunun karmaşıklığı nedeniyle veri ön işleme, aktif bir araştırma alanı olmaya devam etmektedir. [10]

2.3. Çalışmada Kullanılan Veriler

Öncelikle veri setleri incelendi. Bu çalışmada, UCI-Machine Learning Repository [11] sitesinden 3 farklı veri seti kullanıldı. Kullanılan veri setlerine ait öznitelik sayısı ve bilgileri Tablo 1'de verilmektedir.

Tablo 1. Çalışmada Kullanılan Veri Setleri

Veri seti	**Öznitellikler**	**Açıklaması**	**Veri sayısı**
1-Breast-cancer Wisconsin	1. Sample code number: id number	1.kimlik numarası: Algoritma sonucunu etkilemeyeceği için silindi	698
	2. Clump Thickness:	2. Küme Kalınlığı	
	3. Uniformity of Cell Size:	3. Hücre Boyutunun Tekdüzeliği	
	4. Uniformity of Cell Shape:	4. Hücre Şeklinin Tekdüzeliği:	
	5. Marginal Adhesion:	5. Marjinal Yapışma:	
	6. Single Epithelial Cell Size:	6. Tek Epitel Hücre Boyutu	
	7. Bare Nuclei:	7. Çıplak Çekirdekler	
	8. Bland Chromatin:	8. Mülayim Kromatin:	
	9. Normal Nucleoli:	9. Normal Nükleol	
	10. Mitoses:	10. Mitoz:	
	11.Class:(2 for benign, 4 for malignant)	11. Sınıf: (2 iyi huylu, 4 kötü huylu)	
2- Kalp Hastalık ları	1. (yaş)	1.Yıl olarak	
	2. (cinsiyet)	2. kadın , erkek	
	3. (cp)	3. cp:(chest pain type) göğüs ağrısı tipi -- Değer1: tipik angina -- Değer2: atipik angına -- Değer3:anjinal olmayan ağrı -- Değer4: asemptomatik	
	4. (trestb/sn)	4.Dinlenme kan basıncı (hastaneye kabulde mm Hg olarak)	
	5. chol:	5. (kol) mg/dl cinsinden serum kolesterolü ve/veya ST yükselmesi veya > 0.05 mV çökmesi) -- Değer 2: Estes kriterlerine göre olası veya kesin sol ventrikül hipertrofisini gösteriyor	

	6. thalach:	6.ulaşılan maksimum kalp atış hızı	303
	7. fbs:	7.(açlık kan şekeri > 120 mg/dl) (1 = doğru; 0 = yanlış)	
	8. restecg:	8.dinlenme elektrokardiyografik sonuçları -- Değer 0: normal -- Değer 1: ST-T dalga anormalliğine sahip olmak (T dalgası inversiyonları	
	9. (exang)	9.exang: egzersize bağlı göğüs ağrısı (1 = evet; 0 = hayır)	
	10. oldpeak	10.dinlenmeye göre egzersizin neden olduğu ST depresyonu	
	11. slope(eğim) :	11.Zirve egzersiz ST segmentinin eğimi -- Değer 1: eğimli -- Değer 2: düz -- Değer 3: aşağı eğimli	
	12. ca:	12.Floroskopi ile renklendirilen büyük kapların (0-3) sayısı	
	13. thal:	13.3 = normal; 6 = sabit kusur; 7 = tersine çevrilebilir kusur	
	14. Riskli:	14.kalp hastalığı teşhisi (anjiyografik hastalık durumu) -- Değer 0: < %50 çap daralması -- Değer 1: > %50 çap daralması	
3-Diabet	1.HastaNo	1.Hasta numarası: Sonuca etki etmediği için silindi	768
	2.Gebelik Haftası:	2.Hamilelikteki hafta	
	3.Glikoz:	3.Plazma glikoz konsantrasyonu,	
	4.Tansiyon	4.(kan basıncı) Diyastolik kan basıncı (mm Hg)	
	5.Cilt Kalınlığı:	5.Triceps deri kıvrım kalınlığı (mm)	
	6.İnsülin:	6.İki saatlik serum insülini (mu U/ml)	
	7.BMI:	7.Vücut kitle indeksi (kg cinsinden ağırlık/(m cinsinden boy)^2)	
	8.SoydaOlanHastalikFonksiyonu:	8.Diyabet soyağacı işlevi	
	9.Yaş:	9.Yaş (yıl)	
	10.Hasta :	10.Sonuç sınıf değişkeni (0 :hasta değil veya 1:hasta)	

Bu çalışmada eksik değerleri doldurma, sayısal azaltma, veri dönüştürme ve standardizasyon yöntemleri uygulandı.

Eğitim ve doğrulama için verilerin hangi bölümünün kullanılacağını seçmek, açık bir sorundur. Genelleme hatasının yüksek bir varyans tahmincisi olduğu bilinmektedir ve aşırı iyimser sonuçlar bir örneğin çıkarılmasından oluşur. Eğitim seti, tahmin ediciyi yalnızca kalan eğitim verileri temelinde yapılandırır, ardından kaldırılan örnek üzerinde test eder. Bu şekilde, eğitim verilerinin ve ortalamalarının tüm örnekleri test edilir.[12] Bu yöntemlerle yapılan tahminlerin doğruluğu araştırıldı. Bunlar ise doğrulama yöntemlerinde ele alınacaktır. Yani veriler sınıflandırma algoritmalarında kullanılmak üzere eğitim ve test verisi olarak ikiye ayrıldı.

2.4. Sınıflandırma: İki temel sınıflandırma türü vardır: İkili ve çok sınıflı problem. Bu bölümde bazı sınıflandırma algoritmaları açıklanacak ve bunların değerlendirmeleri incelenecektir.

Naive Bayes: Naive Bayes sınıflandırıcısı muhtemelen bugün endüstride en temel ve yaygın olarak kullanılan yöntemlerdir. Basittir, hızlıdır, kolayca güncellenir ve birçok teorik ve hatta teknik tuzaklara rağmen pratikte oldukça etkilidir. Bayes teoreminden faydalanılarak oluşturulan bir algoritmadır. Bu yöntem kullanılarak bir verinin hedeflenen niteliğin sınıf değerine ait olma olasılığı bulunabilmektedir. Diğer bir ifade ile elde var olan, sınıflandırılmış veriler kullanılarak yeni gelen verinin mevcut bulunan sınıf etiketlerinden birisi olma ihtimalini hesaplayan bir yöntemdir. [13] Tahmin işlemi bağımsız değişkenin, bağımlı değişkenler üzerindeki etkilerini bir araya getirerek yeni bir durumu sınıflandırarak yapar.

$$X = [x_1, x_2, x_3, \dots x_n]$$
$$C = [c_1, c_2, c_3, \dots c_n]$$
$$P(C_i/x) = \frac{p(x / C_i)\, P(C_i)}{P(x)} \qquad (1)$$

"Denklem (1)" durumların olasılık değerlerinin hesaplandığı formüldür.

$$\arg max_{ci}\{ P(x|c_i)\, P(c_i) \} \qquad (2)$$

Karşılaşılan yeni durum ise "(2)" nolu denklemde yer alan formül kullanılarak hesaplanan en yüksek olasılık değerinin sınıfına atanır. [14]

K-En Yakın Komşu: Komşu tabanlı sınıflandırma, bir örnek tabanlı öğrenme veya genelleştirmeyen öğrenme türüdür: Genel bir dahili model oluşturmaya çalışmaz, yalnızca eğitim verilerinin örneklerini depolar. Sınıflandırma, her noktanın en yakın komşularının basit çoğunluk oyu ile hesaplanır: bir sorgu noktasına, noktanın en yakın komşuları içinde en fazla temsilciye sahip olan veri sınıfı atanır. [15] En yakın komşu gibi mesafe tabanlı algoritmaları uygulamak için tüm özellikleri aynı ölçeğe

getirmek her zaman tavsiye edilir. Daha geniş menzilli özelliğin daha küçük özelliği tamamen nasıl gölgede bırakacağı veya küçülteceği ortadadır ve bu, daha yüksek büyüklüğe sahip değişkenlere daha yüksek ağırlık vereceğinden, tüm mesafe tabanlı modellerin performansını etkiler. Modelin eğitildiği aynı noktalarla test etmek yerine, verileri test etmek için bilinmeyen veri noktalarına sahip olmak gerekir. Bu, model performansının çok daha iyi yakalanmasına yardımcı olur.[16]

Karar Ağaçları: Temel düzeyde, makine öğrenimi, geçmişi temel alarak geleceği tahmin etmekle ilgilidir. Bu, o nesnenin gözlenen özelliklerine dayanarak, bazı nesnelerin gözlemlenmeyen bazı özellikleri hakkında bilinçli tahminler yapmak anlamına gelir. Öğrenen makineler geliştirmek için, öğrenmenin gerçekte ne anlama geldiğini ve başarıyı (veya başarısızlığı) nasıl belirleyeceğimizi bilmeliyiz. Ezberleme ile genelleme arasındaki farkı açıklamak gerekir. Genelleme, makine öğrenimindeki belki de en merkezi kavramdır. Geçmişe dayalı geleceği tahmin etmeye yönelik bu genel kurulum, çoğu makine öğreniminin merkezinde yer alır. Algoritmanın öğrenmesi beklenen eğitim verileri verilir. Bu eğitim verilerine dayanarak, öğrenme algoritması, yeni bir örneğe karşılık gelen bir tahminle eşleştirecek olan bir f fonksiyonunu indükler. Algoritmanın çok sayıda tahmin yapabilmesi gerekir, bu yüzden algoritmanın değerlendireceği örnekler koleksiyonuna test seti olarak atıf yapılır. Test seti yakından korunan bir sırdır: Algoritma ona önceden bakarsa, hile yapar ve olması gerekenden daha iyisini yapar. Endüktif makine öğreniminin amacı, bazı eğitim verilerini almak ve bir f fonksiyonunu indüklemek için kullanmaktır. Bu f fonksiyonu test verileri üzerinde değerlendirilecektir. Test verilerindeki performansı yüksekse makine öğrenme algoritması başarılı olmuştur. Karar Ağacı Öğrenme Modeli, klasik ve doğal bir öğrenme modelidir. Temel bilgisayar bilimi kavramı olan "böl ve yönet" kavramıyla yakından ilişkilidir. Karar ağaçları birçok öğrenme problemine uygulanabilse de, en basit durumla başlanır: ikili sınıflandırma. Amaç bilinmeyen bir hastalığın bilinmeyen bir kişide olup olmayacağını tahmin etmek olduğu varsayılırsa sadece "evet" veya "hayır" diye cevap verilir. Bir tahminde bulunmak için, söz konusu hastalık/kişi hakkında ikili sorular sorulabilir. Öğrenmedeki amaç, hangi soruların sorulacağını, hangi sırayla sorulacağını ve yeterli soru sorulduğunda hangi cevabı tahmin edeceğini bulmaktır. Karar ağacı, soru ve tahmin dizisini bir ağaç biçiminde yazabildiği için bu adla anılır. Bu şekilde, sorular iç ağaç düğümlerine ve tahminler yapraklara yazılmıştır.[17]

Rasgele Orman: RO, farklı ön yüklenmiş örnekler ve alt örneklenmiş özellikler üzerine kurulmuş çok sayıda karar ağacı modeli kullanan bir sınıflandırma (doğal olarak çok sınıflı) ve regresyon algoritmasıdır. Algoritma kullanımı kolay ve anlaşılırdır. (temeli karar ağacıdır) Basitliği nedeniyle RO, herkesin makine öğrenimini başarıyla uygulamasına izin verebilir. Bir ölçü (entropi veya gini indeksi) optimize ederek ağaçlar oluşturulabilir; Ağaç tamamlandığında ölçüyü en çok geliştiren özelliği seçer. Algoritma birkaç tekrarlanan adımla çalışır:

1.Eğitim setini birden çok kez ön yüklenir. Algoritma, her ön yükleme sırasında toplulukta tek bir ağaç oluşturmak için kullanmak üzere yeni bir küme elde eder.

2.Bir ağaçta her en iyi bölünmüş değişkeni bulmak için kullanılacak eğitim setinde rastgele bir kısmi özellik seçilir.

3.Ön yükleme örneklerini kullanarak eksiksiz bir ağaç oluşturur. Her bölmede yeni alt örneklenmiş özellikleri değerlendirir.

4.Önyükleme aşamasında seçmediğiniz örnekleri kullanarak her ağacın performansı hesaplanır.

5.Özellik önem istatistikleri üretilir ve örneklerin ağacın terminal düğümlerinde nasıl ilişkilendirildiği hesaplanır.

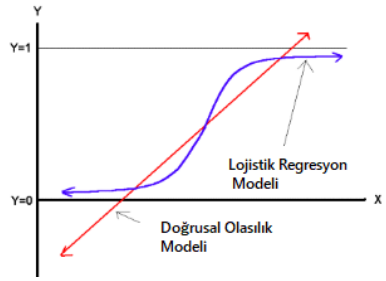
6.Topluluktaki tüm ağaçları tamamladığınızda yeni örnekler için bir ortalama hesaplanır. Tahmin olarak her biri için ortalama tahmini sınıf bildirilir.

Tüm bu adımlar, çözüm yanlılığını sınırladığı için nihai çözümün hem sapmasını hem de varyansını azaltır. Çözüm, her ağacı mümkün olan maksimum uzantısına göre oluşturur, böylece her ağacın diğerlerinden farklı olduğu anlamına gelen karmaşık hedef işlevlerin bile hassas bir şekilde yerleştirilmesine izin verir. Bir ağaç tarafından alınan her bölme, güçlü bir şekilde rastgeledir. Çözüm, yalnızca rastgele bir özellik seçimini dikkate alır. Sonuç olarak, önemli bir özellik tahmin gücü açısından diğerlerine hakim olsa bile, bir ağacın seçimi içermediği zamanlar, ağacın dallarını ve uç

yapraklarını geliştirmenin farklı yollarını bulmasını sağlar. Torbalama ile temel fark, ağaç dallarını bölerken göz önünde bulundurulması gereken özelliklerin sayısını sınırlamak için bir fırsattır. Seçilen öznitelik sayısı küçükse, tam ağaç diğerlerinden farklı olacak ve bu nedenle topluluğa ilişkisiz ağaçlar eklenecektir. Öte yandan, seçim küçükse, ağacın uydurma gücü sınırlı olduğu için yanlılık artar. Her zaman olduğu gibi, bölme için dikkate alınacak doğru sayıda özelliğin belirlenmesi, çapraz doğrulama sonuçlarını kullanmanızı gerektirir. Toplulukta çok sayıda ağaç yetiştirmede herhangi bir sorun ortaya çıkmaz. Hesaplama maliyeti göz önünde bulundurulmalıdır (büyük bir topluluğu tamamlamak uzun zaman alır). Farklı sayıda ağaçtan oluşan topluluklar üzerinde yapılan testler öğrenme eğrilerine benzer.[18]

Lojistik Regresyon: Regresyonda olduğu gibi girdi-çıktı ilişkileriyle ilgili ancak çıktı değişkeninin sürekli olmaktan çok ayrık olduğu birçok durum vardır. Özellikle ikili sonuçların olduğu birçok durum vardır (kişi kalp hastalığına yakalanacak ya da yakalanmayacak). İkili sonuca ek olarak, sürekli olabilen veya olmayabilen bazı girdi değişkenleri vardır. Bu tür veriler nasıl modellenebilir ve analiz edilebilir? Buna sınıflandırma denir ve istatistik ve makine öğreniminde önemli bir konudur. Bununla birlikte, "evet" veya "hayır" tahmininde bulunmak oldukça geneldir. Gürültüyü hesaba katan ve sadece ikili bir cevap vermeyen bir şey genellikle faydalı olacaktır. Lojistik Regresyonu özetlemek gerekirse: Bir ikili çıktı değişkeni var ve koşullu olasılık $Pr(Y = 1|X = x)$'i x'in bir fonksiyonu olarak modellenir; fonksiyondaki herhangi bir bilinmeyen parametre maksimum olabilirlik ile tahmin edilecektir. 1 Yanlış sınıflandırma oranını en aza indirmek için, $p \geq 0,5$ olduğunda $Y = 1$ ve $p < 0,5$ olduğunda $Y = 0$ tahmin etmeliyiz. Bu, x negatif olmadığında 1, aksi halde 0 tahmin edilmesi anlamına gelir. Lojistik regresyon bize lineer bir sınıflandırıcı verir. Öngörülen iki sınıfı ayıran karar sınırı, x tek boyutluysa bir nokta, iki boyutluysa bir doğrunun çözümüdür. Lojistik regresyon sadece sınıflar arasındaki sınırın nerede olduğunu söylemekle kalmaz, aynı zamanda sınıf olasılıklarının belirli bir şekilde sınırdan uzaklığa bağlı olduğunu ve bunların uç noktalara (0 ve 1) doğru gittiğini söyler. Lojistik regresyonu bir sınıflandırıcıdan daha fazlası yapan, olasılıklarla ilgili ifadelerdir. Daha güçlü, daha ayrıntılı tahminler yapar ve farklı bir şekle sığabilir; ancak bu güçlü tahminler yanlış olabilir. Sınıf olasılıklarını tahmin etmek için lojistik regresyon kullanmak, tıpkı lineer regresyon ile nicel değişkenleri tahmin etmek için bir modelleme seçimi olduğu gibi bir modelleme seçimidir.[19] Şekil 1'de Lojistik Regresyon Modeli gösterilmiştir.

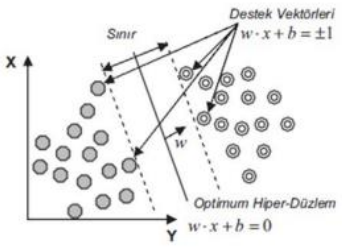
Şekil 1. Lojistik Regresyon Modeli



Destek Vektör Makineleri:

Yüksek boyutta doğrusal sınıflandırma yapabilir. Bir düzlemde bulunan iki grubu ayırmak için bu iki grup üyelerine en uzak çizilen sınırın nasıl çizileceğini belirler. Sınıfları ayırmak için en büyük uzaklığı olan doğrusal fonksiyonu arar. Doğrusal olarak ayrılamıyorsa daha yüksek boyutlu üst uzaya taşıyarak sınıflandırma yapar.[20] Destek vektör makineleri, sınıflandırma işlemi yaparken yüksek düzeyde başarım sağlamak için yüksek boyut özelliklerine sahip çekirdek fonksiyonları kullanmaktadır. DVM ile sınıflandırma yapılacak iki grup arasında sınır çizilerek ayırmak mümkündür. Ayırım işlemi yapılması için iki gruba da yakın ve birbirine paralel iki sınır çizgisi çizilir ve sınır çizgileri birbirine yaklaştırılarak ortak sınır çizgisi oluşturur. Sınıf çizgileri arasında oluşan aralık tolerans olarak adlandırılır.[21] Şekil 2'de destek vektörleri gösterilmiştir.

Şekil 2. Destek Vektörleri



Model Performans Değerlendirme Metrikleri: Sınıflandırma algoritmaları ile oluşturulan modellerin

değerlendirilmesi, hangi sınıflandırma modelinin daha doğru sonuçlar ürettiğinin belirlenmesinde bazı değerlendirme metrikleri kullanılmaktadır.

2.5 Karışıklık Matrisi: Yanlış tahminler yapan modelin sebep olduğu karışıklığı özetlemeye yardımcı olur. Kısacası, karışıklık matrisi, bir sınıflandırma algoritmasının performansını özetlemek için kullanılan bir tekniktir, yani ikili çıktılara sahiptir. Bir karışıklık matrisi, sınıflandırma modelimizin neyi doğru tahmin ettiği ve ne tür hatalar yaptığı konusunda bize daha iyi bir fikir verir. Doğru ve yanlış tahmin edilen değerlerin sayısı, sayılan değerlerle özetlenir ve aşağıda gösterildiği gibi gerçek değerlere karşı tablo 2'de saklanır: [22]

Tablo 2. Karışıklık Matrisi

Karışıklık Matrisi		Gerçek (Actual) Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahminlenen (Predicted) Sonuçlar	Pozitif (1)	DP [1,1] Doğru Pozizif	YP [1.0] Yanlış Pozitif
	Negatif (0)	YN [0, 1] Yanlış Negatif	DN [0, 0] Doğru Negatif

Karışıklık matrisinde geçen terimler ve anlamları aşağıda listelenmektedir;

- Doğru pozitif (DP); sınıflandırıcı tarafından pozitif sınıfa ait verilerden kaç tanesinin doğru şekilde sınıflandırıldığını temsil etmektedir.

- Doğru negatif (DN); sınıflandırıcı tarafından negatif sınıfa ait verilerden kaç tanesinin doğru şekilde sınıflandırıldığını temsil etmektedir.

- Yanlış Negatif (YN); gerçekte pozitif sınıfa ait olan bir verinin sınıflandırma sonucunda negatif sınıf olarak etiketlenmesidir.

- Yanlış pozitif (YP); gerçekte negatif sınıfa ait olan bir verinin sınıflandırma sonucunda pozitif sınıf olarak etiketlenmesidir.

Karışıklık matrisi yardımı ile sınıflandırma performansını belirlemek için kullanılan metrikler ise açıklamaları ile beraber aşağıda verilmektedir; -

Doğruluk; eğitim kümesi kullanılarak oluşturulan modelin test kümesindeki verileri doğru sınıflandırma oranıdır [23]. Denklem 3'de gösterildiği şekilde hesaplanmaktadır.

$$\text{Doğruluk(Accuracy)} = \frac{DP + DN}{DP + DN + YP + YN} \quad (3)$$

Doğruluk skoru, anlaşılması ve yorumlanması en basit ölçütlerden birisidir. Makine öğrenmesi sınıflandırma algoritmalarının testlerinde sıklıkla kullanılır. Doğruluk skoru 0 ve 1 arasında olup 1'e yaklaşan skorlarda model başarılı kabul edilir.

- Duyarlılık ya da doğru pozitif oran (DPO); sınıflandırıcının gerçekte pozitif sınıfa ait olan verileri doğru olarak tahmin etme oranını vermektedir[1]. Denklem 4'de gösterildiği şekilde hesaplanmaktadır.

$$\text{Duyarlılık (Sensitivity)} = \text{Geri çağırma(Recall)} = \frac{DP}{DP + YN} \quad (4)$$

- Özgüllük ya da doğru negatif oran (DNO); sınıflandırıcının gerçekte negatif sınıfa ait olan verileri doğru olarak tahmin etme oranını vermektedir[1]. Denklem 5'de gösterildiği şekilde hesaplanmaktadır.

$$\text{Özgüllük (Specificity)} = \frac{DN}{DN + YP} \quad (5)$$

- Kesinlik ya da Pozitif tahmin değeri (PTD); sınıflandırma sonucunda pozitif olarak tahmin edilenlerin ne oranda doğru olarak tahmin edildiğini vermektedir[1]. Denklem 6'de gösterildiği şekilde hesaplanmaktadır.

$$\text{Kesinlik (Precision)} = \frac{DP}{DP + YP} \quad (6)$$

-Negatif tahmin değeri (NTD); sınıflandırma sonucunda negatif olarak tahmin edilenlerin gerçekte ne kadarının negatif sınıfta olduğunu göstermektedir[1]. Denklem 7'da gösterildiği şekilde hesaplanmaktadır.

$$\text{NTD} = \frac{DN}{DN + YN} \quad (7)$$

F-ölçütü: İkili sınıflandırmanın istatistiksel analizinde, F puanı veya F ölçüsü, bir testin doğruluğunun bir ölçüsüdür. Kesinlik, doğru olarak tanımlanmayanlar da dahil olmak üzere tüm pozitif sonuçların sayısına bölünen gerçek pozitif sonuçların sayısına bölündüğü ve testin kesinliği ve geri çağrılmasından hesaplanır ve geri çağırma, gerçek pozitif sonuçların sayısına bölünür. Kesinlik, pozitif tahmin değeri olarak da bilinir ve geri çağırma, tanısal ikili sınıflandırmada duyarlılık olarak da bilinir. F1 puanı, kesinlik ve geri çağırmanın harmonik ortalamasıdır. F1-puanının mümkün olan en yüksek değeri 1.0'dır, bu da mükemmel kesinliği ve geri çağırmayı gösterir ve kesinlik veya geri çağırma sıfır ise mümkün olan en düşük değer 0'dır. Denklem 8'de gösterildiği şekilde hesaplanmaktadır.[24]

$$f_\beta - \text{ölçütü} = \frac{(1+\beta^2)x(\text{kesinlik} * \text{geri çağırma})}{\beta^2 * \text{kesinlik+ geri çağırma}} \quad (8)$$

ROC eğrisi: Bu, bir sınıflandırıcının tüm olası eşikler üzerindeki performansını özetleyen, yaygın olarak kullanılan bir grafiktir. Belirli bir sınıfa gözlem atama eşiğini değiştirirken, Gerçek Pozitif Oranı (y ekseni) Yanlış Pozitif Orana (x ekseni) karşı çizerek oluşturulur. [23] Alıcı işlem karakteristikleri eğrileri altında kalan alan(AUC); sınıflandırıcıların performansını test etmek için genel olarak kullanılmaktadır. AUC, 0 ile 1 arasında değerler üretmektedir. AUC değerinin bire yaklaşması sınıflandırıcının doğru tahminlerde bulunduğunu, 0.5 değerini alması sınıflandırıcının rastgele tahminlerde bulunduğunu ve bu değerin altında aldığı değerlerde ise sınıflandırıcının doğru çalışmadığını göstermektedir [25].

3. Deneysel Çalışmalar ve Sonuçları:

İlk olarak bu çalışmada kullanılan 1-Bereast Canser Wisconsin, 2-Diabetes ve 3- Heart Table veri setinde her biri için Bölüm 2'de yer alan sınıflandırma algoritmaları uygulanmıştır. Veri setini modeli oluşturmak için eğitim ve test parçalarına bölme işlemi yapılmıştır. Test ve eğitim seti yüzdeleri her zaman farklı değerler verilerek sonuçlar elde edilmiştir. Eğitim verisi ile model oluşturulurken, test verisi ile de bu modelin doğruluğunun testi yapılmaktadır. Bu çalışmada eğitim ve test verisi bölme işlemi aşağıdaki şekil 4'te gösterildiği gibi %75 eğitim ve %25 test verisi olacak şekilde yapılmıştır.

Modelin performansını belirlemek için bu işlem her parça için ayrı ayrı belirlenen sayıda yapılarak her aşamada bir başarım değeri elde edilmektedir. Tablo 3'te en yüksek doğruluk değerlerinin hangi sınıflandırma algoritmaları ile elde edildiği verilmektedir.

Tablo 3. Çalışmada Kullanılan Veri Setleri İçin Sınıflandırma Sonuçları.

Veri seti	Sınıflandırma Algoritmaları	Doğruluk		F1-skor		Yorum
		Karışıklık Matrisi	Doğruluk skoru	sınıf	F1	
1-Bereast canser wisconsin	Naive Bayes	[[106 6] [2 61]]	0,9542	2	0,96	Bu veri setinde en yüksek doğruluk değeri rastgele orman algoritmasında bulunmuştur. Lojistik regresyon ve naive bayes aynı sonuçla takip etmiştir. En düşük sonuç ise karar ağaçlarında elde edilmiştir.
				4	0,94	
	K-En Yakın Komşu	[[107 5] [6 57]]	0,9371	2	0,95	
				4	0,91	
	Karar Ağaçları	[[105 7] [6 57]]	0,9257	2	0,94	
				4	0,90	
	Rasgele Orman	[[109 3] [2 61]]	**0,9714**	2	0,98	
				4	0,96	
	Lojistik Regresyon	[[109 3] [5 58]]	0,9542	2	0,96	
				4	0,94	
	Destek Vektör Makineleri	[[106 6] [5 58]]	0,9371	2	0,95	
				4	0,91	
2-Diabetes	Naive Bayes	[[92 15] [18 29]]	0,7857	0	0,85	Bu veri setinde ise en yüksek doğruluk değeri rastgele orman algoritmasında bulunmuştur. Ancak Lojistik regresyon ile
				1	0,64	
	K-En Yakın Komşu	[[88 19] [17 30]]	0,7662	0	0,83	
				1	0,62	
	Karar Ağaçları	[[83 24] [22 25]]	0,7012	0	0,78	
				1	0,52	

	Rasgele Orman	[[94 13] [16 31]]	**0,8181**	0	0,87	çok yakın sonuç alınmıştır. En düşük sonuç ise yine karar ağaçlarında elde edilmiştir.
				1	0,70	
	Lojistik Regresyon	[[97 10] [19 28]]	0,8116	0	0,87	
				1	0,66	
	Destek Vektör Makineleri	[[87 20] [19 28]]	0,7467	0	0,82	
				1	0,59	
4- HeartTable	**Naive Bayes**	[[21 1] [8 31]]	0,8524	0	0,82	Bu veri setinde en yüksek doğruluk değeri K-En yakın Komşu algoritmasında bulunmuştur. Lojistik regresyon ve naive bayes ve sonra Rastgele orman algoritması onu takip etmektedir. En düşük sonuç yine karar ağaçlarında elde edilmiştir.
				1	0,87	
	K-En Yakın Komşu	[[20 2] [6 33]]	0,8688	0	0,83	
				1	0,89	
	Karar Ağaçları	[[17 5] [7 32]]	0,8032	0	0,74	
				1	0,84	
	Rasgele Orman	[[18 4] [7 32]]	**0,8196**	0	0,77	
				1	0,85	
	Lojistik Regresyon	[[20 2] [7 32]]	0,8524	0	0,82	
				1	0,88	
	Destek Vektör Makineleri	[[18 4] [6 33]]	0,8360	0	0,78	
				1	0,87	

Tablo 3'te gözlemlendiği gibi F1 skor ve doğruluk değerlerine bakıldığında en iyi sonuç Rastgele Orman Algoritması ile elde edilirken en düşük sonuçlara Karar Ağaçlarında rastlanmıştır. Lojistik Regresyonda yine sonuç olarak iyi başarım göstermiştir. Tablo 4'de uygulamada kullanılan algoritmaların karışıklık matrisleri incelenmektedir.

Tablo 4. Bereast Canser Wisconsin Veri Seti İçin Oluşturulan Karışıklık Matrisleri.

Tahmin Edilen Sınıf	Gerçek Sınıf: Kanser	Gerçek Sınıf: Sağlıklı
Kanser	DP [1,1] 109	YP [1.0] 3
Sağlıklı	YN [0,1] 2	DN [0,0] 61

a) Rastgele orman（Entropy）

Tahmin Edilen Sınıf	Gerçek Sınıf: Kanser	Gerçek Sınıf: Sağlıklı
Kanser	DP [1,1] 109	YP [1.0] 3
Sağlıklı	YN [0,1] 3	DN [0,0] 60

b) Rastgele orman（Gini）

Tahmin Edilen Sınıf	Gerçek Sınıf: Kanser	Gerçek Sınıf: Sağlıklı
Kanser	DP [1,1] 109	YP [1.0] 3
Sağlıklı	YN [0,1] 5	DN [0,0] 58

c) Lojistik Regresyon (C=0.1, max_iter=1000)

Tahmin Edilen Sınıf	Gerçek Sınıf: Kanser	Gerçek Sınıf: Sağlıklı
Kanser	DP [1,1] 107	YP [1.0] 5
Sağlıklı	YN [0,1] 6	DN [0,0] 57

d) K-en yakın komşu (n_neighbors=14, metric='minkowski')

Tahmin Edilen Sınıf	Gerçek Sınıf: Kanser	Gerçek Sınıf: Sağlıklı
Kanser	DP [1,1] 106	YP [1.0] 6
Sağlıklı	YN [0,1] 2	DN [0,0] 61

e) Naive Bayes

Tahmin Edilen Sınıf	Gerçek Sınıf: Kanser	Gerçek Sınıf: Sağlıklı
Kanser	DP [1,1] 106	YP [1.0] 6
Sağlıklı	YN [0,1] 5	DN [0,0] 58

f) SVC(kernel= 'sigmoid', degree=4)

Tahmin Edilen Sınıf	Gerçek Sınıf: Kanser	Gerçek Sınıf: Sağlıklı
Kanser	DP [1,1] 105	YP [1.0] 7
Sağlıklı	YN [0,1] 6	DN [0,0] 58

g) Karar ağaçları (criterion='entropy')

Tahmin Edilen Sınıf	Gerçek Sınıf: Kanser	Gerçek Sınıf: Sağlıklı
Kanser	DP [1,1] 108	YP [1.0] 4
Sağlıklı	YN [0,1] 8	DN [0,0] 55

h) Karar ağaçları (criterion=' gini ')

Tablo 4 (a)'da yer alan karışıklık matrisi, rasgele orman sınıflandırıcısı entropi yöntemi ile oluşturulan modeli temsil etmekte ve Tablo 4 (b)'deki de yine rasgele orman ve gini yöntemi ile oluşturulan modeli temsil etmektedir. Tablo 4 (c)'da yer alan karışıklık matrisi, lojistik regresyon sınıflandırıcı yöntemi ile oluşturulan modeli temsil etmekte ve Tablo 4 (d)'deki de k-en yakın komşu yöntemi ile oluşturulan modeli temsil etmektedir. Tablo 4 (e)'de yer alan karışıklık matrisi naive bayes sınıflandırıcı yöntemi ile oluşturulan modeli temsil etmekte ve Tablo 4 (f)'deki de svm yöntemi ile oluşturulan modeli temsil etmektedir. Tablo 4 (g)'de yer alan karışıklık karar ağacı sınıflandırıcı ve entropi yöntemi ile oluşturulan modeli temsil etmekte ve Tablo 4 (h)'deki de yine karar ağacı sınıflandırıcı ve gini yöntemi ile oluşturulan modeli temsil etmektedir. Rastgele ormanda entropi yöntemi ile oluşturulan modelin başarım değeri daha yüksektir. Çünkü yanlış negatif ve doru negatifte daha iyi tahmin yapmıştır. Karar ağaçlarında ise doğru pozitif ve yanlış pozitiflerde gini daha iyi sonuç üretmiştir. Oluşturulan modeller incelendiğinde dengesiz veri setlerinde sadece doğruluk metriğinin yeterli olmadığı tespit edilmiştir. Bu metriğe ek olarak F ölçütü ve AUC değerlerinin de verilmesi modelin seçiminde etkili olmaktadır. Son olarak Şekil 3'te Sınıflandırma raporu örneği verilmiştir.

Şekil 3. Sınıflandırma Raporu

	precision	recall	f1-score	support
2	0.96	0.96	0.96	112
4	0.94	0.92	0.93	63
accuracy			0.95	175
macro avg	0.95	0.94	0.94	175
weighted avg	0.95	0.95	0.95	175

4. Sonuçlar ve Tartışma:

Bu makalede Veri Madenciliği konusunda bir altyapı oluşturmak ve sağlık profesyonellerine sağlık sektöründe Veri Madenciliğinin kullanımı ile ilgili örnekler sunarak karar verme süreçleri açısından yeni bir bakış açısı kazandırmak amaçlanmıştır. Veri Madenciliği, sağlık profesyonellerinin en doğru ve güncel bilgiye ulaşmasını, en objektif ve optimum çözümleri kullanmasını sağlayacak bir karar destek aracıdır. Veri Madenciliğinin konunun uzmanı kişiler tarafından sağlık sektöründe kullanımı, sağlık hizmetlerinin daha etkin sunumu, kaynakların daha verimli kullanımı ve bilimsel, karşılaştırılabilir, şeffaf bilgi erişimi açısından önerilmektedir. Bu makalede aynı zamanda, sınıflandırma çalışmalarında karşılaşılan özel durumlar ve model seçimi yapılırken referans alınan metrikler dikkate alınarak sınıflandırma başarımının doğru bir şekilde değerlendirilebilmesi yönünde deneysel çalışmalar yapılmıştır. Doğruluk metriğinin yanı sıra, F ölçütü ve AUC gibi metriklerinde kullanılması gerektiğinin önemine dikkat çekilmiştir. Sınıflandırma modeli oluşturulurken kullanılan test tekniğinin önemi yapılan çalışmada ön plana çıkmaktadır. Sadece doğruluk metriğinin kullanılması yanlış model seçimine sebep olabilmektedir. Dolayısıyla oluşturulan modellerde sadece doğruluk metriğinden faydalanılması rasgele tahminlerde bulunan modellerin daha iyi sonuç verdiği sonucunu yansıtabilmektedir. Bu durumun önüne geçmek için F ölçütü metriğinden yararlanılması faydalı olacaktır. Üretilen ve paylaşılan veri miktarındaki artış, yüksek hacme sahip veri içerisinden bilgi keşfini zorlaştırmakta, etkin araç, yöntem ve stratejilerin geliştirilmesini zorunlu kılmaktadır. Bu sebeple, son yıllarda veri madenciliği tekniklerinin büyük veri setleri üzerinde uygulanabilmesi için bulut bilişim teknolojisinden faydalanıldığı ve oldukça başarılı sonuçlar elde edildiği görülmektedir. Bulut bilişim teknolojisi kullanıcıya yüksek boyutta veri saklama alanı, paralel işlem yeteneği, her yerden ulaşılabilirlik gibi birçok özelliği uygun maliyetle sunmaktadır. Bulut platformundaki veri madenciliği uygulama alanlarını, algoritmalarda yapılan iyileştirmeleri ve elde edilen sonuçları Tıp alanında da uygulanması için çalışmalar yapılmasının iyi olacağı sonucuna ulaşmıştır. Sonuç olarak gerçekleştirilen bu çalışmalar göstermiştir ki, algoritmaların büyük veriler üzerinde bulut bilişim teknolojisi ile, yük dağılımının yapılması ve böylece çalışma süreleri kısaltılarak performansın büyük ölçüde arttırılması mümkündür.

Kaynaklar

[1] A. ALAN, ve M. KARABATAK, "Veri Seti - Sınıflandırma İlişkisinde Performansa Etki Eden Faktörlerin Değerlendirilmesi" , Fırat Üniversitesi Müh. Bil. Dergisi, 32(2), 531-540, 2020

[2] A.S. Koyuncugil , ve N. Özgülbaş, "Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları", Bilişim Teknolojileri Dergisi, Cilt: 2, Sayı: 2, Mayıs 2009

[3] A. HALTAŞ ,ve A. ALKAN, "Medline Veritabanı Üzerinde Bulunan Tıbbi Dokümanların Kanser Türlerine Göre Otomatik Sınıflandırılması", Bilişim Teknolojileri Dergisi, Cilt: 9, Sayı: 2, Mayıs 2016

[4]A. Çınar, "Veri Madenciliğinde Sınıflandırma Algoritmalarının Performans Değerlendirmesi Ve R Dili İle Bir Uygulama", Marmara Üniversitesi Öneri Dergisi, Cilt 14, Sayı 51, Ocak 2019

[5] Ş. Hacıefendioğlu, "Makine öğrenmesi yöntemleri ile glokom hastalığının teşhisi", Yüksek Lisans Tezi, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Konya, 2012.

[6] E. Kartal, "Sınıflandırmaya dayalı makine öğrenmesi teknikleri ve kardiyolojik risk değerlendirmesine ilişkin bir uygulama", Doktora Tezi, İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, 2015.

[7] B. DAĞ ,ve A.VAROL, "2D:4D Sayısal Parmak Oranına Göre Bireylerin Kişilik Durumlarının Sınıflandırılması", International Symposium on Digital Forensics and Security (ISDFS'13), 20-21, 2013.

[8] F. DOĞAN, ve İ.TÜRKOĞLU, "Derin Öğrenme Algoritmalarının Yaprak Sınıflandırma Başarımlarının Karşılaştırılması", Sakarya Unıversıty Journal Of Computer And Informatıon Scıences Vol. 1, Id. Saucıs-1-2018, 2018

[9] P. YILDIRIM, ve D. BİRANT, "Bulut bilişimde veri madenciliği tekniklerinin uygulanması: Bir literatür taraması", Pamukkale Univ Muh Bilim Derg., 24(2), 336-343, 2018

[10] J. Han, M. Kamber, ve J. Pei, Data Mining Concepts and Techniques, Morgan Kaufmann, USA, 2012

[11] Machine Learning Repository, URL: https://archive.ics.uci.edu/ml/datasets/ (Erişim Tarihi: 01.09.2021)

[12] I.Guyon, A.Elisseeff,(L.P. Kaelbling), "An Introduction to Variable and Feature Selection", Journal of Machine Learning Research 3 (2003) 1157-1182

[13] Bozkır AS, Sezer E, Gök B. Öğrenci seçme sınavında (öss) öğrenci başarımını etkileyen faktörlerin veri madenciliği yöntemleriyle tespiti. 5. Uluslararası İleri Teknolojiler Sempozyumu, 2009, 13-15 Mayıs, Karabük, s:1-7.

[14] M. Aydoğan, ve A.Karci, "Apache Spark ile Naive Bayes Yöntemi Kullanarak Spam Mail Tespiti" , International Conference on Artificial Intelligence and Data Processing (IDAP) ,2018,

[15] Scikit Learn, URL: https://scikit-learn.org/stable/modules/neighbors.html#classification (Erişim zamanı: 05.09.2021)

[16] Kaggle, URL: https://www.kaggle.com/shrutimechlearn/step-by-step-diabetes-classification-knn-detailed (Erişim Tarihi;05.09.2021)

[17] Tom M. Mitchell, Machine Learning, McGraw Hill, 1997

[18]J.P. Mueller,ve L. Massaron, Machine Learning For Dummies, John Wiley & Sons, Kanada, 2016

[19] Zhao, Linqiao, Advanced Data Analysis from an Elementary Point of View, Cosma Rohilla Shalizi, 2017

[20] Vapnik, V.N.,"The Nature of Statistical Learning Theory", Springer-Verlag, New York, 1995.

[21] A. HALTAŞ ,ve A. ALKAN, "Medline Veritabanı Üzerinde Bulunan Tıbbi Dokümanların Kanser Türlerine Göre Otomatik Sınıflandırılması", Bilişim Teknolojileri Dergisi, Cilt: 9, Sayı: 2, Mayıs 2016

[22] Data scholl, URL: https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/ (Erişim Tarihi: 06.09.2021)

[23] Japkowicz N. Performance evaluation for learning algorithms, Cambridge University Press, Cambridge 2011.

[24] Wikipedia, URL: https://en.wikipedia.org/wiki/F-score, (Erişim Tarihi: 06.09.2021)

[25] Kılıç S. Klinik karar vermede ROC analizi. Journal of Mood Disorders 3 (3): 135-40;2013