TÜRKİYE CUMHURİYETİ
ANKARA ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİNDE HİBRİT MODEL YAKLAŞIMI

Batuhan BAKIRARAR

BİYOİSTATİSTİK ANABİLİM DALI
DOKTORA TEZİ

DANIŞMAN
Prof. Dr. Atilla Halil ELHAN

ANKARA
2021

ETİK BEYAN

Ankara Üniversitesi

Sağlık Bilimleri Enstitüsü Müdürlüğü'ne,

Doktora tezi olarak hazırlayıp sunduğum "Veri Madenciliğinde Hibrit Model Yaklaşımı" başlıklı tez; bilimsel ahlak ve değerlere uygun olarak tarafımdan yazılmıştır. Tezimin fikir/hipotezi tümüyle tez danışmanım ve bana aittir. Tezde yer alan araştırma tarafımdan yapılmış olup, tüm cümleler, yorumlar bana aittir.

Yukarıda belirtilen hususların doğruluğunu beyan ederim.

Öğrencinin Adı Soyadı: Batuhan BAKIRARAR

Tarih: 17.05.2021

İmza:

KABUL VE ONAY

Ankara Üniversitesi Sağlık Bilimleri Enstitüsü
Biyoistatistik Anabilim Dalında
Batuhan BAKIRARAR tarafından hazırlanan
"Veri Madenciliğinde Hibrit Model Yaklaşımı" adlı tez çalışması
aşağıdaki jüri tarafından DOKTORA TEZİ olarak OY BİRLİĞİ ile kabul edilmiştir.

Tez Savunma Tarihi:07.06.2021

Prof. Dr. Erdem KARABULUT

Hacettepe Üniversitesi

Jüri Başkanı

Prof. Dr. Atilla Halil ELHAN

Ankara Üniversitesi

Doç. Dr. S. Kenan KÖSE

Ankara Üniversitesi

Doç. Dr. Derya GÖKMEN

Ankara Üniversitesi

Dr. Öğr. Üyesi Osman DAĞ

Hacettepe Üniversitesi

Tez hakkında alınan jüri kararı, Ankara Üniversitesi Sağlık Bilimleri Enstitüsü Yönetim Kurulu tarafından onaylanmıştır.

Prof. Dr. Fügen AKTAN

Sağlık Bilimleri Enstitüsü Müdürü

İÇİNDEKİLER

ÖNSÖZ

Hibrit model, her bir veri madenciliği yönteminin gücünü kullanmak ve zayıflıklarını gidermek için bu yöntemlerin etkili bir kombinasyonu olarak tanımlanmaktadır. Özellikle sağlık alanında ihtiyaç duyulan ve sınıflama performansını artıran bu yaklaşım son zamanlarda oldukça popüler bir araştırma konusudur. Bu çalışmada çeşitli varyasyonlar sonucu üretilen simüle veriler ve gerçek sağlık verileri üzerinde klasik veri madenciliği yaklaşımları ve hibrit model yaklaşımı uygulanmıştır. Sonuç olarak hibrit model ile performans ölçütlerinin arttığı ve bu yaklaşımın kullanımının önemi vurgulanmaya çalışılmıştır.

Tez çalışmam ve doktora öğrenimim süresince bilgi ve tecrübeleri ile yaptığı yardımlardan ötürü danışmanım Sayın Prof. Dr. Atilla Halil ELHAN'a ve akademik anlamda desteğini esirgemeyen Anabilim Dalımız öğretim üyesi hocalarıma teşekkürlerimi ve saygılarımı sunarım.

Eğitim hayatım boyunca maddi ve manevi desteklerini bir an olsun esirgemeyen aileme sonsuz teşekkürü borç bilirim.

SİMGELER VE KISALTMALAR

AIC	Akaike Bilgi Kriterleri
AİK	Alıcı İşlem Karakteristiği
BD	Bilgi Değeri
ÇKA	Çok Katmanlı Algılayıcı
DÖ	Değişken Önemi
EAKA	Eğri Altında Kalan Alan
FURIA	Bulanık Sırasız Kural İndüksiyon Algoritması
IBk	Örnek Tabanlı
MKK	Matthews Korelasyon Katsayısı
PRC	Precision-Recall Eğrisi

ŞEKİLLER

ÇİZELGELER

1. GİRİŞ

Bilgisayarların kullanımının artmasıyla, bilgisayar tabanlı sistemler tarafından büyük miktarda veri üretilmektedir. Devlet kurumları, bilimsel kurumlar ve işletmeler, veri toplamak ve depolamak için muazzam kaynaklara sahiptir. Gerçekte, bu verilerin büyük bir kısmı hiç kullanılmayacaktır, çünkü çoğu durumda veri hacimleri yönetilemeyecek kadar büyüktür veya veri yapıları etkili bir şekilde analiz edilemeyecek kadar karmaşıktır (Kantardzic, 2011, p:2).

Büyük, karmaşık, bilgi açısından zengin veri kümelerini anlama ihtiyacı neredeyse tüm iş, bilim ve mühendislik alanlarında yaygındır. Bu verilerde saklı olan yararlı bilgileri çıkarma ve bu bilgilere göre hareket etme yeteneği günümüzün rekabetçi dünyasında giderek daha önemli hale gelmektedir. Verilerden bilgi çıkarmak için yeni teknikler de dahil olmak üzere bilgisayar tabanlı bir yöntem uygulama sürecine veri madenciliği denir (Kantardzic, 2011, p:2).

Veri tabanlarında bilginin çıkarılması yoluyla, büyük veri tabanları, bilginin alınması ve doğrulanması için zengin, güvenilir kaynaklar görevi görür ve keşfedilen bilgi, bilgi yönetimi, karar verme, süreç kontrolü ve diğer birçok uygulamaya uygulanabilir. Bu nedenle, veri madenciliği en önemli ve zorlu araştırma alanlarından biri olarak kabul edilmiştir. Veri tabanı sistemleri, bilgi tabanı sistemleri, yapay zeka, makine öğrenimi, bilgi edinimi, istatistik, veri tabanları ve veri görselleştirme dahil olmak üzere birçok farklı alandaki araştırmacılar veri madenciliğine büyük ilgi göstermiştir (Kantardzic, 2011, p:2).

Sağlık sektörü, büyük veriye sahip kaynaklar arasında yer almaktadır. Tıbbi bilgi, hasta bilgileri ve diğer hasta/hastane verileri günlük olarak büyümeye devam etmektedir. Örneğin, bir hastanenin yılda beş terabayt veri üretebileceği tahmin edilmektedir. Bu verileri kaliteli sağlık hizmetleri adına yararlı bilgiler elde etmek için kullanabilmek çok önemlidir (Kantardzic, 2011, p:4).

Ayrıca, bilgisayar kaynaklı bilgi alımı, kaliteli karar almayı desteklemeye ve insan hatalarını önlemeye yardımcı olabilir. İnsanların karar vermesi genellikle optimal olmasına rağmen, sınıflandırılacak büyük miktarda veri olduğunda zayıftır. Ayrıca, insanlar strese ve yoğun bir çalışmaya sokulduğunda kararların etkinliği ve doğruluğu azalacaktır. Örneğin, belli bir zaman diliminde 5 hasta kaydını incelemesi gereken bir doktor, yoğunluk sebebiyle aynı zaman diliminde 50 hasta kaydı incelemek zorunda kalırsa; doktorun sonuçları iletme doğruluğunun, analiz edilecek sadece beş kaydı olduğunda elde edilenler kadar yüksek olmayacağı neredeyse kesindir. Bu tür sebepler bizi bilgi keşfi, analizi için veri madenciliğinin en iyi çözüm olduğu sonucuna götürmektedir (Kantardzic, 2011, p:4).

1.1. Veri Madenciliği Nedir?

Veri madenciliği, büyük hacimli verilerde yeni, değerli ve keşfedilmemiş bilgi arayışıdır. Bu, insanların ve bilgisayarların ortak bir çabasıdır. Makine öğrenimi, veri tabanı teknolojisi, istatistik, yüksek performanslı bilgi işlem, görselleştirme ve matematik gibi mevcut teknolojilerin sinerjisinden doğan hızlı büyüyen bir teknolojidir. Daha önce bilinmeyen, potansiyel olarak yararlı ve ilginç bilgileri büyük ve sıklıkla farklı, geçmiş veri kaynaklarından ayıklama yarı otomatik süreci olarak da tanımlanır. Veri Madenciliği; verilerin büyük, karmaşık ve farklı olması nedeniyle ortaya çıkan sayısız zorluğun yanı sıra veri ön işleme, bilgi son işleme ve alan uzmanlığının dahil edilmesi gibi kullanıcıların işlevsellik gereksinimlerini ele almaktır (Wu ve ark., 2006, p:13).

Veri madenciliği ile ilgili popüler bir efsane, bir veri madenciliği motorunun (genellikle veri madencisi olarak adlandırılır) bir veri tabanından her türlü bilgiyi bağımsız olarak çıkarmasını ve bunları insan talimatları veya müdahalesi olmadan kullanıcılara sunmasını beklemektir. Ancak, büyük bir bilgi kümesi, veri tabanındaki veri kümelerinin birçok farklı kombinasyonundan üretilebilir. Veri tabanından üretilen tüm bilgi kümesi, bayt cinsinden ölçülürse, veri tabanının boyutundan çok daha büyük olabilir. Bu nedenle, veri tabanından keşfedilebilen bu tür bir bilgi

kümesi oluşturmak, saklamak veya sunmak gerçekçi değildir. Nispeten gerçekçi bir amaç, bir alan uzmanının ve bir veri madenciliği uzmanının birlikte çalışmasıdır (Wu ve ark., 2006, p:14).

Uygulamada, veri madenciliğinin iki temel amacı tahmin ve açıklama olma eğilimindedir. Tahmin, ilgili değişkenin/değişkenlerin bilinmeyen değerlerini tahmin etmek için veri kümesindeki bazı değişkenlerin kullanılmasını içerir. Tanımlama ise, insanlar tarafından yorumlanabilecek verileri tanımlayan kalıplar bulmaya odaklanır (Kantardzic, 2011, p:2).

Spektrumun tahmin edici ucunda, veri madenciliğinin amacı, sınıflandırma, tahmin veya diğer benzer görevleri gerçekleştirmek için kullanılabilen yürütülebilir bir kod olarak ifade edilen bir model üretmektir. Spektrumun tanımlayıcı ucunda ise amaç, büyük veri kümelerindeki örüntüleri ve ilişkileri ortaya çıkararak analiz edilen sistemi anlamaktır. Belirli veri madenciliği uygulamaları için tahmin ve tanımlamanın göreceli önemi büyük ölçüde değişebilir. Tahmin ve tanımlama hedeflerine, aşağıdaki görevleri açıklanan veri madenciliği teknikleri kullanılarak ulaşılmıştır (Kantardzic, 2011, p:3):

1. Sınıflandırma: Bir veri öğesini önceden belirlenmiş birkaç sınıftan birinde sınıflandıran tahmin edici bir öğrenme işlevidir.

2. Regresyon: Bir veri öğesini gerçek değere ait tahmin değişkeniyle eşleştiren öğrenme işlevidir.

3. Kümeleme: Verileri tanımlamak için sınırlı kategoriler veya kümeler tanımlamak isteyen ortak tanımlayıcı bir görevdir.

4. Özetleme: Bir veri kümesi (veya alt kümesi) için kompakt bir açıklama bulma yöntemlerini içeren ek bir tanımlayıcı görevdir.

5. Bağımlılık Modellemesi: Değişkenler arasında veya veri kümesindeki veya veri kümesinin bir bölümündeki bir özelliğin değerleri arasındaki önemli bağımlılıkları açıklayan yerel bir model bulma işidir.

6. Değişim ve Sapma Tespiti: Veri kümesindeki en önemli değişiklikleri keşfetmektir (Kantardzic, 2011, p:3).

1.2. Veri Madenciliği Yöntemleri

Veri madenciliği yöntemleri iki başlık altında toplanabilir. Bunlar denetimli ve denetimsiz öğrenme yöntemleridir (Dangeti, 2017, p:8).

1.2.1. Denetimli Öğrenme

Denetimli öğrenmede, her eğitim algoritması için doğru sonuçları (sınıf değişkenini) içeren bir eğitim seti vardır. Eğitim örneği tüm doğru sonuçları içerir ve eğitim algoritmasının görevi doğru sonuçları çoğaltmaktır (Dangeti, 2017, p:9).

Denetimli öğrenme, öngörülemeyen bir girdi örneği göz önüne alındığında, tahmin yapmak için inşa edilmiş bir öğrenme modelidir. Denetimli bir öğrenme algoritması, regresyon / sınıflandırma modelini öğrenmek için bilinen bir girdi veri kümesi ve bu kümeye ait bilinen yanıtları (çıktıları) alır. Daha sonra bir öğrenme algoritması, yeni girilecek verilere veya test veri kümesine yanıt için bir tahmin oluşturmak üzere modeli eğitir. Denetimli öğrenme, tahmin modelleri geliştirmek için sınıflandırma algoritmalarını ve regresyon tekniklerini kullanır. Sık kullanılan algoritmalar, Karar Ağaçları, Destek Vektör Makinesi, Random Forest, Naive Bayes ve K-en yakın komşu'nun yanı sıra Lojistik Regresyon ve Yapay Sinir Ağları'dır (Dangeti, 2017, p:9).

Bu tez kapsamında denetimli öğrenme yöntemlerinden Adaboost, Naive Bayes, Lojistik Regresyon, Çok Katmanlı Algılayıcı, Destek Vektör Makinası, J48, Random Forest, Decision Table, Hoeffding Tree, Örnek Tabanlı (IBk) Öğrenme Algoritması ve Bagging; denetimsiz öğrenme yöntemlerinden ise K-ortalama kullanılmıştır.

1.2.1.1. Random Forest

Random Forest, birçok karar ağacından oluşan ve tek tek ağaçların ürettiği sınıfların modu (çoğunluk ağacın seçtiği) olan sınıfı çıkaran bir sınıflandırma algoritmasıdır. Bu yöntemde kullanılan ağaç sayısı ise kullanılacak algoritmaya veri setine göre değişiklik gösterebilir. Eğitim verilerini değil ağaçları rastgele seçer ve seçtiği rastgeleleştirme yöntemi algoritmaya göre değişebilir. Genellikle diğer karar ağaçlarının performanslarını iyileştirir. Her bir düğümü ayırmak için rastgele bir özellik seçimi kullanmak, Adaboost'a göre daha uygun olan hata oranları sağlar. (Biau ve Scornet, 2016 ve Breiman, 2001).

Yüksek sapma veya yüksek yanlılık içeren diğer karar ağaçlarının aksine, iki uç arasında doğal bir denge bulmak için ortalamaları kullanır. Random forest, veri kümesinin çeşitli alt örneklerine bir dizi klasik karar ağacına uyan ve tahmin doğruluğunu geliştirmek ve aşırı uyumu kontrol etmek için ortalama alma kullanan bir meta tahmincidir. Bu yöntemde kullanılan her karar ağacı, eğitim verilerinin rastgele bir alt kümesi kullanılarak oluşturulur (Biau ve Scornet, 2016).

1.2.1.2. Hoeffding Ağacı

Bir Hoeffding Ağacı, dağıtım üreten örneklerin zamanla değişmediğini varsayarak, büyük veri akışlarından öğrenme yeteneğine sahip, artımlı, bir karar ağacı algoritmasıdır. Hoeffding ağaçları, küçük bir örneğin çoğu zaman en uygun bölme niteliğini seçmek için yeterli olabileceği gerçeğinden yararlanır. Bu fikir, bazı istatistikleri öngörülen bir kesinlikte tahmin etmek için gereken gözlem sayısını niceleyen Hoeffding sınırı tarafından matematiksel olarak desteklenmektedir.

Hoeffding Ağaçları'nın teorik olarak güçlü bir özelliği, çoğu durumda iyi performans gösteriyor olmasıdır. Hoeffding Ağacı tarafından öğrenilen model, eğer eğitim örneklerinin sayısı yeterince büyükse, artımlı olmayan bir veri madenciliği algoritması tarafından oluşturulan modelle asimptotik olarak hemen hemen aynıdır (Bifet ve ark., 2010).

1.2.1.3. J48

J48 (C4.5), Ross Quinlan tarafından geliştirilen ve bilgi teorisine dayanan karar ağacı üreten bir sınıflandırma algoritmasıdır. Quinlan'ın önceden geliştirdiği ID3 algoritmasının bir uzantısıdır. J48 tarafından üretilen karar ağaçları sınıflandırma için kullanılabilir ve bu nedenle J48'e genellikle istatistiksel bir sınıflandırıcı denir. 2008 yılında Springer LNCS tarafından yayınlanan Veri Madenciliğinde En İyi 10 Algoritma içinde 1. sırada yer aldıktan sonra oldukça popüler olmuştur. 2011 yılında, Weka makine öğrenimi yazılımının yazarları J48 algoritmasını "pratikte bugüne kadar en sık kullanılan makine öğrenimi algoritması ve önemli bir karar ağacı programı" olarak tanımlamışlardır (Kantardzic, 2011, p:442 ve Quinlan, 1986, p:88).

1.2.1.4. Adaboost

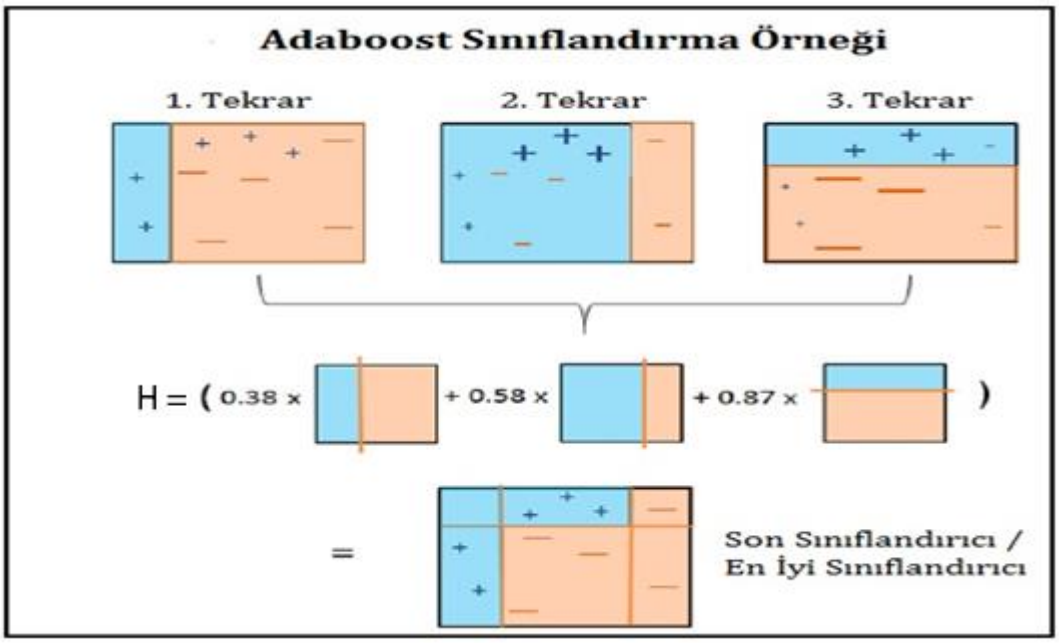
"Adaptive Boosting" in kısaltması olan AdaBoost, 1996 yılında Freund ve Schapire tarafından önerilen ilk pratik güçlendirme algoritmasıdır. Sınıflandırma problemlerine odaklanır ve bir dizi zayıf sınıflandırıcıyı güçlü bir biçime dönüştürmeyi amaçlar. Bu yöntemde başlangıçta, verilere basit bir sınıflandırıcı atanır ve doğru olarak sınıflandırılan örneklere bir sonraki yinelemede daha düşük ağırlık ve yanlış olarak sınıflandırılan örneklere ise daha yüksek ağırlık atanır. Bir sonraki yineleme için ağırlıklar tekrar hesaplanır ve aynı yöntem tekrarlanır. Yinelemeler tamamlandıktan sonra, sınıfları yüksek doğrulukla tahmin eden güçlü bir sınıflandırıcı bulmak için hesaplanan ağırlıklar (her bir sınıflandırıcıda her bir yinelemede hata oranına göre otomatik olarak hesaplanan ağırlıklar) birleştirilir (Şekil 1) (Dangeti, 2017, p:158).

$$H(x) = \sum_{t} p_t h_t(x)$$

t: tekrar sayısı

p: hesaplanan ağırlık

h: sınıflandırma



Şekil 1. AdaBoost Sınıflandırmasına ait Örnek Yaklaşım.

1.2.1.5. Bagging

Bagging, istatistiksel sınıflandırma ve regresyonda kullanılan makine öğrenme algoritmalarının kararlılığını ve doğruluğunu geliştirmek için tasarlanmış bir makina öğrenme topluluğu meta algoritmasıdır. Ayrıca varyansı azaltır ve aşırı öğrenmeyi önlemeye yardımcı olur. Genellikle karar ağacı yöntemlerine uygulansa da her türlü yöntemle kullanılabilir. Model ortalama yaklaşımının özel bir durumudur. Weka

programı tarafından bu özellikleri kullanılarak geliştirilmiş bir veri madenciliği sınıflama algoritması olarak da kullanılmaktadır (Breiman, 1996).

1.2.1.6. Naive Bayes

Bayes algoritması kavramı oldukça eskidir ve Thomas Bayes tarafından 18. yüzyılda bulunmuştur. Thomas, bilinen olaylardan bilinmeyen olayların olasılığını belirlemek için temel matematiksel ilkeleri geliştirmiştir. Örneğin, tüm elmalar kırmızı renkteyse ve ortalama çapları yaklaşık 10 cm ise, rastgele bir meyve sepetinden kırmızı renk ve 9,3 cm çapında bir meyve seçilirse, bu meyvenin bir elma olması olasılığı nedir? Naive terimi, bir sınıftaki belirli özelliklerin diğerlerinden bağımsızlığını varsayar. Bu durumda, renk ve çap arasında bir bağımlılık yoktur. Bu bağımsızlık varsayımı, gerçekçi olmasa bile, yüksek kelime hazinelerinin bulunduğu kelimelere dayalı e-posta sınıflandırması gibi belirli görevler için Naive Bayes sınıflandırıcısını hesaplama kolaylığı açısından kullanışlı hale getirir. Naive Bayes sınıflandırıcısı, pratik uygulamalarda şaşırtıcı derecede iyi performans sergilemektedir (Albon, 2018, p:279 ve Dangeti, 2017, p:203).

Bayes sınıflandırıcılar en iyi sonuçların olasılığını tahmin etmek için çok sayıda özellikten gelen bilgilerin aynı anda dikkate alınması gereken çalışmalarda uygulanır. Bayesci yöntemler tahmin için dikkate alınması gereken tüm kanıtları (düşük etkiye sahip özellikleri de) kullanır. Bu bilgi, göreceli olarak küçük etkileri olan çok sayıda özelliğin bir arada ele alınmasının güçlü sınıflandırıcılar oluşturacağı varsayımına dayanmaktadır (Albon, 2018, p:279 ve Dangeti, 2017, p:203).

1.2.1.7. Lojistik Regresyon

Lojistik regresyon, nitel bağımlı değişkeni, bağımsız değişkenlere göre bir logit değişkenine dönüştürdükten sonra maksimum olabilirlik tahmini uygular. Bu şekilde, lojistik regresyon belirli bir olayın gerçekleşme olasılığını tahmin eder. Aşağıdaki

denklem, açıklayıcı değişkenlerin bir fonksiyonu olarak doğrusal olarak değişir (Albon, 2018, p:259 ve Dangeti, 2017, p:85).

$$\log(odds) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \cdots + \beta_n * X_n$$

β_0: Sabit

$\beta_{1,2,\ldots,n}$: Değişkenlere ait regresyon katsayıları

1.2.1.7.1. Bilgi değeri (BD): Bu yöntem, değişkenlerin modele dahil edilmeden önce ön analizinde (filtrelemesinde) çok yararlıdır (Çizelge 1). BD temel olarak, modelin oluşturulmasından önceki ilk adımda önemsiz değişkenleri veriden çıkarmak için kullanılır, çünkü son modelde mevcut olan değişken sayısı yaklaşık 10 olmalıdır. Bu nedenle, değişken sayısını azaltmak için bu işlem gereklidir (Dangeti, 2017, p:87).

$$Bilgi\ Değeri = \ln\left(\frac{\%iyi}{\%kötü}\right) * (\%iyi - \%kötü)$$

$$Değişken\ Önemi\ (DÖ) = \ln(\frac{\%iyi}{\%kötü})$$

Çizelge 1. Bilgi Değeri Sınıflamasına göre Öngörü Gücü.

Bilgi Değeri	Öngörü Gücü
<0,02	Tahmin için değersiz
0,02-0,1	Zayıf tahmin edici
0,1-0,3	Orta düzeyde tahmin edici
0,3-0,5	Güçlü tahmin edici
>0,5	Şüpheli ya da çok iyi tahmin edici

Örnek: Aşağıdaki çizelgede, sürekli değişken olan fiyat 10 kutuda sayılan olayların ve olay olmayanların sayısına göre aralıklara ayrılmıştır ve tüm aralık değerleri için bilgi değeri hesaplanmıştır. Çizelge 2'de toplam BD 0,0356 olarak bulunmuş, yani bu model olayları sınıflandırmak için zayıf bir yordayıcıdır yorumuna varılmıştır (Dangeti, 2017, p:88).

Çizelge 2. Bilgi Değeri Hesaplama Örneği.

Fiyat Aralığı	**No**	**Olay**	**Olay Olmayan**	**Olay Oranı [O]**	**Olay Olmayan Oranı [OO]**	**[O]-[OO]**	**DÖ= ln(O/OO)**	**BD= DÖ*(O-OO)**
0-50	1	40	394	%5	%5	0,003	0,062	0,00002
51-100	2	68	900	%9	%12	-0,024	-0,234	0,0056
101-150	3	78	984	%10	%13	-0,021	-0,186	0,0040
151-200	4	102	1194	%14	%15	-0,016	-0,111	0,0018
201-250	5	108	1218	%15	%16	-0,011	-0,074	0,0008
251-300	6	110	1164	%15	%15	-0,001	-0,010	0,0000
301-350	7	82	772	%11	%10	0,011	0,107	0,0012
351-400	8	46	330	%6	%4	0,019	0,379	0,0074
401-450	9	42	368	%6	%5	0,009	0,179	0,0017
>451	10	68	470	%9	%6	0,031	0,416	0,0129
Toplam		744	7794					0,0356

1.2.1.7.2. Akaike bilgi kriterleri (AIC): Bu, belirli bir veri kümesi için istatistiksel bir modelin göreli kalitesini ölçer. Kurulan iki model arasındaki bir karşılaştırma sırasında, daha düşük AIC'ye sahip olan model daha yüksek değere sahip olana göre tercih edilir (Dangeti, 2017, p:88).

$$AIC = -2 * \ln(L) + 2 * k$$

L = Maksimum Olabilirlik değeri

k = Modeldeki değişken sayısı

1.2.1.7.3. Alıcı işlem karakteristiği eğrisi: Ayırıcı eşiği değiştiği için ikili bir sınıflandırıcının performansını gösteren grafiksel bir çizimdir. Eğri, çeşitli eşik değerlerinde yanlış pozitif orana karşı doğru pozitif oran çizilerek oluşturulur (Dangeti, 2017, p:89).

1.2.1.8. Çok Katmanlı Algılayıcı (ÇKA)

Çok katmanlı ileri beslemeli ağlar, gerçek dünya uygulamalarında en önemli ve en popüler yapay sinir ağları sınıflarından biridir. Tipik olarak ağ, ağın giriş katmanını, bir veya daha fazla ara katmanını ve son olarak hesaplama düğümlerinin bir çıkış katmanını oluşturan bir dizi girdi içerir. İşlem katman katman ileri yönde yapılır. Bu tip Yapay Sinir Ağları genel olarak, tek katmanlı bir ağ olan basit algılayıcının genelleştirilmesini temsil eden ÇKA olarak adlandırılır (Kantardzic, 2011, p:213).

ÇKA algoritmasının üç ayırt edici özelliği vardır:

1. Ağdaki her nöron modeli genellikle sigmoidal veya hiperbolik olan doğrusal olmayan bir aktivasyon fonksiyonu içerir.

2. Ağ, ağın giriş veya çıkışının bir parçası olmayan bir veya daha fazla ara nöron katmanı içerir. Bu ara katmanlar, giriş modellerinden giderek daha anlamlı özellikler çıkararak, ağın karmaşık ve oldukça doğrusal olmayan görevleri öğrenmesini sağlar.

3. Ağ, bir katmandan diğerine yüksek derecede bağlantı gösterir (Kantardzic, 2011, p:213).

ÇKA'lar, ağın hata geri yayılım algoritması olarak bilinen oldukça popüler bir algoritma ile denetimli bir şekilde eğitilmesiyle bazı farklı ve çeşitli sorunları çözmek için başarıyla uygulanır. Bu algoritma hata düzeltme öğrenme kuralına dayanır ve genelleştirilmiş hali olarak görülebilir. Temel olarak, hata geri yayılım öğrenimi ağın farklı katmanları üzerinden gerçekleştirilen iki aşamadan oluşur. Bunlar hatayı ileri ve geriye doğru yayma algoritmalarıdır (Kantardzic, 2011, p:214).

İleri yayma algoritmasında, ağın giriş düğümlerine bir eğitim örneği verilir ve etkisi ağ katmanı boyunca katmanlara yayılır. Son olarak, ağın gerçek yanıtı olarak bir dizi çıkış üretilir. İleri yayma sırasında, ağın sinaptik ağırlıkları sabittir. Geriye doğru yayma algoritmasında ise ağırlıkların tümü bir hata düzeltme kuralına göre ayarlanır. Özellikle, ağın gerçek yanıtı, bir hata sinyali üretmek için eğitim verisinin bir parçası olan arzu edilen (hedef) bir yanıttan çıkarılır. Bu hata sinyali daha sonra sinaptik bağlantıların yönüne karşı ağ üzerinden geriye doğru yayılır. Sinaptik ağırlıklar, ağın gerçek yanıtını istenen yanıta daha yakın yapacak şekilde ayarlanır (Kantardzic, 2011, p:215).

1.2.1.9. Destek Vektör Makinesi

Destek vektör makinaları, bir dizi etiketli eğitim verisinden girdi-çıktı haritalama işlevleri üreten denetimli öğrenme yöntemleridir. Destek vektör makinaları, genelleştirilmiş doğrusal modeller ailesine aittir. Ayrıca "çekirdek yöntemleri" ne ait oldukları da söylenir. İstatistiksel öğrenme teorisindeki sağlam matematiksel temeline ek olarak, Destek vektör makinaları bilgi keşfi ve veri madenciliği için popüler, modern araçlardır ve tıbbi tanı koyma, biyoinformatik, yüz tanıma, görüntü işleme ve metin madenciliği gibi çok sayıda gerçek dünya uygulamalarında da oldukça iyi bir performans sergilemektedir. Yapay sinir ağlarına benzer şekilde, Destek vektör makinaları, karmaşık sistemlerin ve süreçlerin modellenmesinde oldukça iyidirler (Dangeti, 2017, p:220 ve Huang ve ark., 2006, p:209).

1.2.1.10. Karar Tablosu (Decision Table)

Karar tabloları, verilen koşullara bağlı olarak hangi eylemlerin gerçekleştirileceğini belirlemek için kısa ve görsel bir sunumdur. Çıktıları bir dizi eylem olan algoritmalardır. Karar tablolarında ifade edilen bilgiler, karar ağaçları olarak veya programlama dilinde, if-then-else ve switch-case ifadeleri dizisi olarak da temsil edilebilir (Kohavi, 1995).

Her karar, olası değerleri koşul alternatifleri arasında listelenen bir değişkene veya ilişkiye karşılık gelir. Her eylem, gerçekleştirilecek bir yordam veya işlemdir ve girdiler, girdiye karşılık gelen koşul alternatifleri kümesi için eylemin (hangi sırayla) gerçekleştirileceğini belirtir. Bir karar tablosu, girdi değişkenlerinin olası her kombinasyonunu içeriyorsa dengeli veya eksiksiz kabul edilir. Başka bir deyişle, dengeli karar tabloları girdi değişkenlerinin sağlandığı her durumda bir eylem öngörmektedir (Kohavi, 1995).

1.2.1.11. Örnek Tabanlı (IBk) Öğrenme Algoritması

Örnek Tabanlı (IBk) öğrenme algoritması, öğrenilen bilgileri basitçe eğitim vakalarının veya örneklerinin bir koleksiyonu olarak temsil eder. Örneklerden denetlenen bir öğrenme biçimidir; eğitim olaylarının tam bir hafızasını tutar ve yeni vakaları en benzer eğitim örneklerini kullanarak sınıflandırır. Daha sonra en yüksek benzerliğe sahip örneği bularak ve sınıfını tahmin olarak kullanarak yeni bir durum sınıflandırılır. Bu nedenle, IBk algoritması çok düşük bir eğitim çabasıyla sınıflandırma yapabilir. Öte yandan, bu, tüm eğitim vakalarını hafızada tutma ihtiyacından kaynaklanan yüksek depolama taleplerine yol açar. Ayrıca, yeni vakaları mevcut tüm örneklerle karşılaştırmak gerekir, bu da sınıflandırma için yüksek hesaplama maliyetiyle sonuçlanır. Ek olarak çapraz doğrulamaya göre uygun K değerini seçebilir ve mesafe ağırlıklandırması da yapabilir (Gelbukh ve Reyes-Garcia, 2006, p:447).

IBk, öznitelik aralıklarını normalleştirmesi, örnekleri aşamalı olarak işlemesi ve eksik değerleri tolere etmek için basit bir algoritmaya sahip olması dışında en yakın komşu algoritmasıyla aynıdır. Ayrıca, IBk yalnızca yanlış sınıflandırılmış örnekleri kaydeder ve sınıflandırma sırasında kaydedilen örneklerden hangisinin iyi performans göstermesinin beklendiğini belirlemek için bir "bekle ve gör" kanıt toplama yöntemi kullanır (Gelbukh ve Reyes-Garcia, 2006, p:448).

1.2.2. Denetimsiz Öğrenme

Denetimsiz öğrenmenin amacı, sınıflandırma veya regresyon yöntemlerini gerçekleştirmek için hiçbir sonuç (bağımlı) değişkeninin bulunmadığı verilerin gizli desenlerini (kalıplarını) veya yapılarını keşfetmektir. Denetimsiz öğrenme yöntemleri genellikle daha zordur, çünkü sonuçlar özneldir ve analiz için sınıfı veya sürekli değişkeni tahmin etmek gibi basit bir amaç yoktur. Bu yöntemler, açıklayıcı veri analizinin bir parçası olarak gerçekleştirilir. Üstelik, denetimsiz öğrenme yöntemlerinden elde edilen sonuçları değerlendirmek zor olabilir, çünkü sonuçların onaylanmasını gerçekleştirmek için evrensel olarak kabul edilmiş bir mekanizma yoktur. Bununla birlikte, denetimsiz öğrenme yöntemleri günümüzde bir trend konusu olarak çeşitli alanlarda artan bir öneme sahiptir ve birçok araştırmacı şu anda bu yeni ufku keşfetmek için bu konu üzerinde çalışmaktadır. Literatürde en sık kullanılan algoritma ise K-Ortalama'dır (Dangeti, 2017, p:304 ve Huang ve ark., 2006, p:175).

1.2.2.1. K-Ortalama

Kümeleme, gözlemleri aynı kümenin üyeleri birbirine daha yakın olacak ve farklı kümelerin üyeleri birbirinden çok farklı olacak şekilde gruplama görevidir (Dangeti, 2017, p:305).

Kümeleme, bir veri kümesini araştırarak veri içindeki temel örüntüleri tanımlamak veya bir grup özellik oluşturmak için yaygın olarak kullanılır. Sosyal ağlar söz konusu olduğunda, toplulukları tanımlamak ve insanlar arasında bağlantılar önermek üzere kümelenebilirler. Sık kullanıldığı alanlardan birkaç örnek vermek gerekirse (Dangeti, 2017, p:305):

- Kara para aklama ile mücadele tedbirlerinde, şüpheli faaliyetler ve bireyler anomali tespiti kullanılarak tanımlanabilir.

- Biyolojide, kümeleme benzer ekspresyon paternlerine sahip gen gruplarını bulmak için kullanılır.

- Pazarlama analizinde, kümeleme benzer müşterilerin segmentlerini bulmak için kullanılır, böylece farklı müşteri tiplerine farklı pazarlama stratejileri uygulanabilir (Dangeti, 2017, p:305).

K-ortalama kümeleme algoritması, kümeler veya sentroidlerin merkezlerini kurucu noktalarının ortalama konumuna hareket ettiren ve mümkün olan küme merkezlerinin sayısında önemli bir değişiklik oluncaya veya belirlenen yineleme sayısına ulaşıncaya kadar örnekleri en yakın kümelere yeniden atamak için kullanılan yinelemeli bir süreçtir (Dangeti, 2017, p:306).

K-ortalama algoritmasının maliyet fonksiyonu, gözlemler arasındaki Öklid uzaklığı ile belirlenir. Denklemi anlamanın sezgisel bir yolu, yalnızca bir küme (k = 1) varsa, tüm gözlemler arasındaki mesafelerin tek ortalamasıyla karşılaştırılmasıdır. Öte yandan, kümelerin sayısı 2'ye (k = 2) yükselirse, iki yol hesaplanır ve gözlemlerden bazıları küme 1'e atanır ve diğer gözlemler yakınlığa dayalı ikinci kümeye atanır. Daha sonra mesafeler, maliyet fonksiyonlarında aynı mesafe ölçüsü uygulanarak ancak küme merkezlerine göre ayrı ayrı hesaplanır (Dangeti, 2017, p:307)

1.2.3. Hibrit Veri Madenciliği

Son yıllarda, tıp disiplinlerinde toplanan veri miktarı giderek artmaktadır. Dijital teknolojideki gelişmeler, toplanan verilerin boyutu, karmaşıklığı ve miktarında, yani tıbbi raporlarda ve ilgili görüntülerde benzeri görülmemiş bir büyümeye yol açmıştır. Dünya çapında her yıl milyarlarca sağlık kaydı işlemi yapılmaya başlanmıştır (Wu ve ark., 2006, p:14).

Öte yandan, hasta merkezli tıbbi uygulamalar (örneğin, elektronik hasta kayıtları, kişisel sağlık kayıtları, elektronik tıbbi kayıtlar, vb.) pratik olma, veri büyümesini daha da artırma ve veri açısından zengin ancak bilgi açısından fakir bir sağlık sistemine yol açmanın eşiğindedir. Bu nedenle, veri madenciliği araştırmacılarının, gelecekteki tıbbi uygulamaların temeli olarak yararlı kanıtlar sağlamak için bu değerli verileri uygun şekilde kullanabilen yeni bir yaklaşımı araştırması ve önermesi çok önemli hale gelmiştir. En önemli etken, sağlık analisti ve yöneticilerin stratejik kararlar vermelerine ve gerçek sonuçları dikkate alarak gelecekteki sonuçları tahmin etmelerine yardımcı olacak bilgiler sunan bu verileri kullanmaktır (Wu ve ark., 2006, p:14).

Ek olarak, Dünya Sağlık Örgütü tıbbi veri havuzlarından bilginin keşfedilmesi için bazı olası ihtiyaçları tanımlamaktadır. Buna tıbbi tanı ve prognoz, hasta sağlığı planlama ve geliştirme, sağlık sistemi izleme ve değerlendirme, sağlık planlama ve kaynak tahsisi, hastane ve sağlık hizmetleri yönetimi, epidemiyolojik ve klinik araştırma ve hastalık önleme dahildir, ancak kapsam bunlarla sınırlı değildir (Wu ve ark., 2006, p:14).

Son zamanlarda, sağlık hizmeti verilerinin yığılımı, toplanan verilerden "yararlı" bilgiyi keşfetmek için çok sayıda çabaya neden olmuştur ve gerçekten de birçok araştırmacı tarafından ilginç sonuçlar bildirilmiştir. Bununla birlikte, veri madenciliği algoritması olarak bilinen bilgi keşif yönteminin etkinliğine rağmen, bugün sağlık pratisyenlerinin karşılaştığı zorluk, karar destek sistemlerine dayalı bilgileri keşfetmek için doğru verilerle "uygun" veri madenciliği algoritmalarının kullanılmasıdır (Wu ve ark., 2006, p:15).

Özellikle, sinir ağları, istatistiksel modelleme, evrimsel algoritmalar ve görselleştirme araçları gibi veri madenciliği algoritmalarındaki son gelişmeler, her türlü ham verinin üst düzey bilgiye dönüştürülmesini mümkün kılmıştır. Ancak asıl sorun, her yöntemin veri yapısı, şekli ve geçerliliği ile ilgili kendi yaklaşımına sahip olmasıdır. Bu sınırlama sınıflandırma sistemlerinin performansını etkiler. Sonuç olarak, hibrit bir veri madenciliği yaklaşımına duyulan ihtiyaç, veri madenciliği

topluluğu tarafından kabul görmektedir ve bu konuda son yıllarda yapılan çalışma sayısı gittikçe artmaktadır (Wu ve ark., 2006, p:15).

Hibrit Sistemler Yapay Zeka alanında yeni değildir. Goonatilake ve arkadaşları akıllı hibrit sistemi, iki veya daha fazla tekniğin bireysel tekniklerin sınırlamalarının üstesinden gelecek şekilde birleştirildiği bir sistem olarak tanımlar. Hibrit sistemleri üç tipte sınıflandırırlar: fonksiyon değiştirme sistemleri, inter komünikasyon sistemleri ve polimorfik sistemler. Fonksiyon değiştirme sistemleri, bir tekniğin eksikliklerinin, bilgi işlemenin bu yönünde özellikle güçlü olan başka bir teknik kullanılarak giderildiği sistemlerdir. Performans, güvenilirlik veya işlevsellik açısından eksiklikler olabilir. Örneğin, bulanık sistemler, karma bir nöro-bulanık sistem kullanılarak aşılabilecek öğrenme gücünden yoksundur. Inter komünikasyon sistemleri, alt görevlerin çokluğu açısından karmaşıklığı nedeniyle bir problemi çözmek için tek bir tekniğin uygulanamayacağı sistemlerdir. Tek bir mimari içinde birden fazla bilgi işleme yeteneği sergilemek üzere geliştirilmiş sistemler ise, polimorfik sistemler olarak bilinir. Veri Madenciliği içinde, çok sayıda hedefe yönelik, mevcut sistemlerin işlevselliğini arttıran, performansı arttıran ve önyargıları azaltan veya dengeleyen bir hibrit sistem kullanılabilir. Bu hedeflerin her biri fonksiyon değiştirme, inter komünikasyon sistemleri veya polimorfik sistemler kullanılarak yerine getirilebilir. Aslında, polimorfik ve fonksiyon değiştirme sistemleri arasındaki ayrım bulanıktır, çünkü polimorfizm normalde fonksiyon değiştirme ile elde edilir. Bu nedenle, veri madenciliği için hibrit sistemler sınıflandırması, çeşitli tekniklerin birbirleriyle ne ölçüde etkileşime girdiğine dayanmaktadır. İki sınıf hibrit sistem tanımlaması vardır: gevşek bağlanmış ve sıkı bağlanmış Hibrit Sistemler (Syed ve Syed, 2008, p:7).

Gevşek bağlanmış hibrit sistemler, uygulama görevlerinin çokluğu nedeniyle karmaşık bir problemi çözmek veya sistemin performansını artırmak için kullanılması gereken bir takım tekniklerden oluşur. Örneğin, çapraz satış sorunu, iki veri madenciliği görevi tanımlarak çözülebilir. Bu sorun için kullanılan veri madenciliği görevlerinden biri karakteristik bir kural keşif amacına, diğeri ise sapma saptama amacına sahiptir. Böylece, çözüme sadece karakteristik bir kural keşif

algoritması ve bir sapma saptama algoritması ile gevşek bağlı bir hibrit model kullanılarak ulaşılabilir. Gevşek bağlanmış bir hibritin bileşenleri paralel veya sıralı olarak çalışabilir ve bileşenler arası mesajlar aracılığıyla iletişim kurabilir. Bu hibrit sınıfı, Goonatilake ve arkadaşlarının inter komünikasyon olarak sınıfladığı hibritlere karşılık gelir (Syed ve Syed, 2008, p:8).

Sıkı birleştirilmiş hibrit sistemler, tek bir görevi çözmek için birden fazla tekniğin entegrasyonundan oluşur. Bu hibrit sistemin amacı şunlar olabilir: işlevsellik eklemek, yeni görevlerle başa çıkmak, performansı artırmak veya gerçekçi olmayan önyargıları azaltmak için mekanizmalar sağlamak. Örneğin En Yakın Komşu algoritmasını kullanan bir sistem, uzaklık metriğindeki sapmaları kaldırmak için istatistiksel teknikler, işlevsellik eklemek ve performansı arttırmak için genetik bir algoritma kullanabilir. Böylece yeni hibrit bir sistem oluşturulmuş olur (Syed ve Syed, 2008, p:9)

Bu amaçla, her bir tekniğin gücünü kullanmak ve birbirlerinin zayıflıklarını gidermek için çeşitli veri madenciliği tekniklerinin etkili bir kombinasyonu olan hibrit bir veri madenciliği yaklaşımına ihtiyaç olduğu savunulmaktadır (Syed ve Syed, 2008, p:9).

1.2.4. Veri Madenciliği Yöntemlerinin Karşılaştırılması

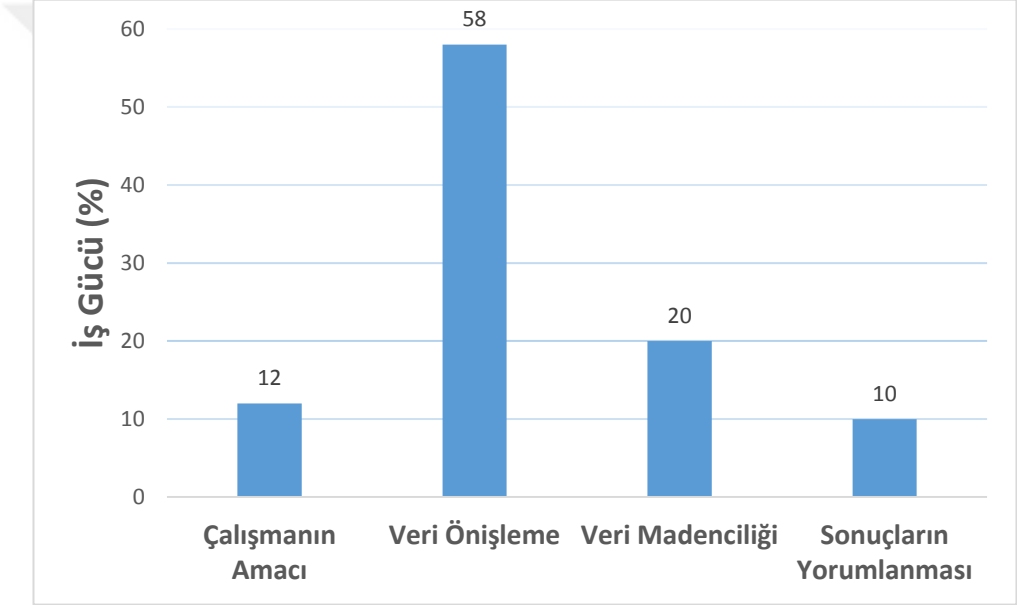
Çizelge 3'te veri madenciliği yöntemlerinin avantaj, dezavantaj ve varsayımları verilmiştir (Olson, 2018, p:24).

Çizelge 3. Veri Madenciliği Yöntemlerinin Avantaj, Dezavantaj ve Varsayımları.

Metod	Avantajları	Dezavantajları	Varsayımları
Kümeleme Analizi	Anlaşılabilir formül üretme	Veri seti boyutu arttıkça hesaplama süresinin artması	Verileri sayısal hale getirme ihtiyacı
	Otomatik olarak uygulanabilme	Seçimlere duyarlı sonuçlarla parametrelerin tanımlanması gerekliliği	
Diskriminant Analizi	Birden fazla değişkeni aynı anda dahil edebilme	Normallik ve bağımsızlık varsayımlarını ihlal etme	Gruplar arasında çok değişkenli normallik varsayımı
	Bağımsız değişkenleri birleştirebilen katsayılarına sahip	Boyutsallık sorunlarının azaltılması	Tüm gruplar arasında eşit grup kovaryansları varsayımı
	Yeni verilere uygulanabilme	Değişkenlerin göreceli önemine dair yorum çeşitliliği	Gruplar ayrık, örtüşmeyen ve tanımlanabilir olmalı
		Sınıflandırma algoritmasını belirlemede zorluk	
		Zaman serisi tahmin testlerini yorumlamada zorluk	
Regresyon	Anlaşılabilir formül üretme	Veri seti boyutu arttıkça hesaplama süresinin artması	Hataların normalliği
	Yaygın olarak anlaşılabilme	Doğrusal olmayan verilerde tahminlerin çok iyi olmaması	Hata otokorelasyonu, heteroskedastisite, çoklu bağlantı
	Güçlü teori yapısı		
Sinir Ağı Modelleri	Çeşitli problemlerle başa çıkabilir	Nitel veri girişi gerekli	Gruplar ayrık, örtüşmeyen ve tanımlanabilir olmalı
	Hem sürekli hem de kategorik değişkenlerle başa çıkabilme	Sonuçların anlaşılması zor	
	Birçok yazılım paketinde yer almakta	Atanan zamandan önce kötü bir tahmin üretebilme	
Karar Ağaçları	Anlaşılabilir kurallar oluşturabilme	Bazı algoritmaların yalnızca iki kategorili hedef sınıflara uygulanabilmesi	Gruplar ayrık, örtüşmeyen ve tanımlanabilir olmalı
	Minimum hesaplama ile sınıflandırma yapabilme	Çoğu algoritmanın aynı anda yalnızca tek bir alanı inceleyebilmesi	
	Kolay formüller kullanma	Hesaplama maliyeti fazla olabilmekte	
	Hem sürekli hem de kategorik değişkenlerle başa çıkabilme		
	Değişken önemini açık bir şekilde belirtme		

1.2.5. Veri Önişleme

Veri madenciliği çalışmalarında yapılan en büyük yanlış verileri daha detaylı incelemeden, veriler arası ilişkileri, aşırı/aykırı değerlerin varlığını göz ardı ederek analizlere başlanmasıdır. Ancak bu tür teknik ve araçları kullanmak için, analiz algoritmalarının uygulanmasından önce projenin anahtar parçalarından biri ve en uzun zaman gerektireni olan önişleme aşaması uygulamak gerekir. Şekil 2 veri madenciliğinin her aşamasında gerekli olan iş gücünü göstermektedir (Dangeti, 2017, p:11 ve Olson, 2018, p:30).



Şekil 2. Veri Madenciliğinin Her Aşaması için Gerekli İş Gücü.

1.2.5.1 Veri Seçimi

Seçim aşamasının hedefi, mevcut veri kaynaklarının tanımlanması ve analiz için gerekli verilerin oluşturulmasıdır. Verilerin seçimi çözülecek sorunun türüne ve izlenen hedefe bağlıdır. Verilerin toplanmasından sonra ilk görev verilerin boyutunu ve kalitesini kontrol etmektir. Sağlam modeller oluşturmak için büyük miktarda veri gerekir. Ancak büyük miktarda veriye sahip olmak yeterli değildir, hem yüksek kalitede hem de büyük miktarda veriye sahip olmak için veri türünü, dağılımını,

değişkenlerdeki maksimum ve minimum değerleri incelemeniz gerekir (Dangeti, 2017, p:11 ve Olson, 2018, p:29).

1.2.5.2. Veri Keşfi

Veri keşfinde amaç, seçilen verilerin kalitesini sağlamaktır. Verilerin temiz ve tutarsızlıklardan arınmış olması, başarılı bir analiz için ön koşuldur. Öte yandan, veriler ne kadar iyi olursa ve değişkenler ne kadar iyi bilinirse, modelleme aşamasında nereye bakılacağını bilmek o kadar kolay olacaktır. Yapılması gereken ilk görev, verilerin kalitesinin ilk ölçüsünü sağlayabilmek için verilerin yapısının denetlenmesidir. Bunun için genellikle veri görselleştirme araçları ve istatistiksel yöntemler kullanılır. Kategorik değişkenler için, değerlerin frekans dağılımlarına bakmak en iyi yoldur. Histogramlar veya pasta grafikler gibi basit araçlar, eksik ve aykırı değerleri farketmeye yardımcı olabilir. Nicel değişkenler için ortalama, standart sapma, varyans, ortanca, minimum ve maksimum değerleri kullanılır. Tüm bunlara bakarak, analize başlamadan önce değişkenin veri setinde kalıp kalmayacağını belirlemek mümkün olacaktır. Diğer faydalı araçlar, değişkenlerin dağılımını incelemek ve ilişkilerini analiz etmek için bir değişkenin başka bir değişkene göre dağılımını gösteren kutu grafikleri, histogramlar veya Q-Q grafikleridir (Kantardzic, 2011, p:55).

1.2.5.2.1. Değişken/Özellik Azaltma (Feature Reduction)

Görüntüler, metin ve multimedya gibi veriler doğası gereği yüksek boyutludur ve verilerin bu boyutsallığı veri madenciliği görevlerini zorlaştırır. Araştırmacılar, verilerin boyutsallığının azaltılmasının verinin doğruluğunu (özelliğini) korurken daha hızlı bir hesaplama ile sonuçlandığını bulmuşlardır. Birçok alakasız değişkenin varlığında, veri madenciliği algoritmaları modelde aşırı uyuma sebep olma eğilimindedir. Bu nedenle, veri madenciliğinde verinin özelliklerinde değişime neden olmadan birçok değişken veri setinden çıkarılabilir (Kantardzic, 2011, p:56).

Değişken/Özellik azaltma işlemi sadece gürültülü verilerle değil, aynı zamanda alakasız, ilişkişiz ve gereksiz verilerle ilgilidir. Temel olarak amaç, minimum veri ve veri işleme çabasıyla maksimum performans elde etmektir. Bir değişken/özellik azaltma süreci sonucunda (Kantardzic, 2011, p:57):

1. Daha az değişken/özellik ile veri madenciliği algoritması daha hızlı öğrenebilir.

2. Daha yüksek doğruluk ile modelin verilerden daha iyi genelleme yapabilmesi sağlanabilir.

3. Veri madenciliği sürecinin anlaşılmasının ve kullanılmasının daha kolay olması için modelin basit sonuçlar vermesi sağlanabilir.

4. Daha az değişken/özellik kullanılır, böylece veri toplama işleminde gereksiz veya alakasız özellikler kaldırılarak zamandan ve veri depolama maliyetinden tasarruf sağlanabilir (Kantardzic, 2011, p:57).

1.2.5.3. Veri Dönüşümü

Analiz aşamasında elde edilecek modellerin başarısı ve doğruluğu, veri analistinin değişkeni nasıl yapılandırdığı ve kullandığına bağlıdır. Öte yandan, bu aşamada verilerin kullanılacak veri madenciliği algoritmaları için uygun bir formata dönüştürülmesi gerekmektedir. Bu şekilde, kullanılacak algoritma nicel/nitel veri gerektiriyorsa ve seçilen veriler nitel/nicel ise, veriler uygun formatı dönüştürülerek analize uygun hale getirilir. Ayrıca, bu aşamada yeni değişkenlerin türetilmesi de çok yaygındır (Kantardzic, 2011, p:33).

1.2.5.4. Değişken Önemi

Değişken önemi aşaması, bazı kaynaklarda veri önişleme aşamasının son aşaması, bazı kaynaklarda da veri önişleme adımından sonra uygulanacak ilk adım olarak belirtilmektedir. Veride gerekli önişleme adımları uygulandıktan sonra değişkenlerin sonuç değişkenine göre önemine, veri setine kattığı değere bakılır. Değişken önemi için kullanılan birçok yöntem bulunmaktadır ancak en yaygın kullanılan yöntemler Chi-Squared Attribute Eval., Info Gain Attribute Eval. ve Gain Ratio Attribute Eval.'dir (Dangeti, 2017, p:113).

1.2.6. Veri Madenciliğinin Gereksinimleri

Veri madenciliği, daha iyi anlaşılmasına neden olabilecek verilerin toplanmasıyla birlikte bir sorunun tanımlanmasını ve istatistiksel araçları veya diğer analiz araçlarını sağlamak için bilgisayar modellerini gerektirir. Veri madenciliğinde çeşitli analitik bilgisayar modelleri kullanılmıştır. Verilere hızlı ve doğru erişim büyük önem arz etmektedir. Oldukça sık, büyük miktarda veriyi yönetmek için veri ambarları ve/veya veri depolayan sistemler kullanılır. Diğer veri madenciliği analizleri, çevrimiçi analitik işleme sistemlerinde düzenlenebilen daha küçük veri kümeleriyle yapılır (Olson, 2018, p:10).

Klasik istatistiksel yaklaşımlar veri madenciliği için temeldir. Otomatik yapay zeka yöntemleri de kullanılır. Bununla birlikte, klasik istatistiksel yöntemlerle sistematik bir araştırma hala veri madenciliğinin temelini oluşturmaktadır. İstatistiksel analiz alanı tarafından geliştirilen araçlardan bazıları, verilerle uğraşırken otomatik kontrol (insan rehberliği) gerektirir (Olson, 2018, p:10).

Veri madenciliği araçlarının çok yönlü, ölçeklenebilir olması, eylemler ve sonuçlar arasındaki yanıtları doğru bir şekilde tahmin edebilmesi ve otomatik uygunabilmesi gerekir. Çok yönlülük, aracın çok çeşitli modelleri uygulama yeteneğini ifade eder. Ölçeklenebilir araçlar, küçük bir veri kümesinde çalışıyorsa,

bu aracın daha büyük bir veri kümesinde de çalışması gerektiğini belirtir. Otomasyon faydalıdır, ancak uygulaması görecelidir. Bazı analitik işlevler genellikle otomatiktir, ancak prosedürleri uygulamadan önce analist/uzman görüşü gereklidir. Aslında, analist kararı veri madenciliğinin başarılı bir şekilde uygulanması için kritik öneme sahiptir. Analize dahil edilecek verilerin doğru seçilmesi çok önemlidir. Veri dönüşümü de sıklıkla gereklidir. Çok fazla değişken çok fazla çıktı üretir, ancak çok azı da verilerdeki önemli ilişkileri göz ardı edebilir (Olson, 2018, p:11).

1.2.7. Veri Madenciliği Yazılımları

Pahalı olmalarına rağmen birçok mükemmel ticari veri madenciliği yazılım ürünü vardır. Bunlar arasında en popülerleri SAS Enterprise Miner ve IBM Intelligent Miner'dır. Bazı yazılımlar ise ücretsizdir. En popüler yazılımlar ise Çizelge 4'te gösterilmiştir (Olson, 2018, p:56).

Çizelge 4. Popülerliklerine Göre Veri Madenciliği Yazılımları.

Sıra	**Ad**	**Tür**
1	R	Açık kaynak kodlu
2	Python	Açık kaynak kodlu
3	SAS	Ticari
4	SPSS	Ticari
5	WEKA	Ücretsiz

Bu çalışmanın amacı, son teknoloji veri madenciliği algoritmalarını ve uygulamalarını sunmak ve tıbbi verilerin kümelenmesi ve sınıflandırılması için yeni bir hibrit veri madenciliği yaklaşımı önermektir. Ayrıca çalışmada, denetimli ve denetimsiz öğrenme yöntemlerinin dengeli ve dengesiz veri setlerinde, farklı örneklem büyüklüklerinde ve farklı değişkenler arası ilişkiler olması durumunda performans ölçütlerinin hesaplanması ve bu ölçütlerin hibrit modelden elde edilen ölçütler ile karşılaştırılması amaçlanmıştır.

2. GEREÇ ve YÖNTEM

2.1. Hibrit Model Yapısı

Hibrit model için ilk önce kullanılan veri seti için en iyi performansa sahip beş veri madenciliği yöntemi oluşturulan java kaynak kodlu yazılım sayesinde seçilmektedir. İkinci aşama olarak seçilen yöntemler en iyi performansa sahip yöntemden en kötü performansa sahip yönteme doğru sıralanmaktadır. Sonraki aşamada en iyi performansa sahip yöntem hibrit model için ilk seçilen yöntem olacak şekilde, bu yöntemlerden sırasıyla ikili, üçlü ve dörtlü yöntemler grubu oluşturulmaktadır. Oluşturulan bu gruplara ait performans ölçütleri teker teker hesaplanmakta ve en iyi sonuç veren grup baz alınarak hibrit model oluşturulmaktadır.

2.2. Karışıklık Matrisi (Confusion Matrix)

Karışıklık matrisi, genellikle bir sınıflandırma modelinin (veya "sınıflandırıcı" nın) gerçek değerlerin bilindiği bir dizi test verisi üzerindeki performansını açıklamak için kullanılan bir tablodur. Bu tablo üzerinden veri madenciliği performans ölçütleri hesaplanabilmektedir (Chicco ve Jurman, 2020).

2.2.1. Matthews Korelasyon Katsayısı

Matthews Korelasyon Katsayısı (MKK), 1975'te biyokimyacı Brian W. Matthews tarafından getirilen ikili (iki sınıf) sınıflandırmaların kalitesinin bir ölçüsü olarak makine öğreniminde kullanılır. MKK, on yıllar önce geliştirilen Karl Pearson'un phi katsayısına eşdeğer olsa da, MKK terimi biyoinformatik alanında yaygın olarak kullanılmaktadır (Chicco ve Jurman, 2020).

Katsayı, doğru ve yanlış pozitifleri ve negatifleri dikkate alır ve genellikle sınıflar çok farklı boyutlarda olsa bile kullanılabilecek dengeli bir önlem olarak kabul edilir. MKK özünde, gözlemlenen ve tahmin edilen ikili sınıflandırmalar arasında bir korelasyon katsayısıdır; −1 ile +1 arasında bir değer döndürür. +1 katsayısı mükemmel bir tahmini temsil eder, 0 rastgele tahminden daha iyi değildir ve −1 tahmin ve gözlem arasındaki toplam anlaşmazlığı gösterir. İstatistikte, phi katsayısı olarak da bilinir. MKK, 2 × 2 çapraz tablosu için ki-kare istatistiği ile ilgilidir (Chicco ve Jurman, 2020):

$$|MKK| = \sqrt{\frac{X^2}{n}}$$

n: Toplam gözlem sayısı

Doğru ve yanlış pozitif ve negatiflerin karışıklık (confusion) matrisini tek bir sayı ile tanımlamanın mükemmel bir yolu olmasa da, Matthews korelasyon katsayısı kullanlabilecek en iyi ölçütlerden biri olarak kabul edilir. Doğru tahminlerin oranı (doğruluk) gibi diğer ölçütler, iki sınıf çok farklı boyutlarda olduğunda yararlı değildir. Örneğin, her nesneyi daha büyük kümeye atamak, yüksek oranda doğru tahmine yol açar, ancak küçük sınıfı gözardı ettiği için genellikle yararlı bir sınıflandırma yöntemi değildir (Chicco ve Jurman, 2020).

MKK, aşağıdaki formül kullanarak doğrudan karışıklık matrisinden hesaplanabilir (Chicco ve Jurman, 2020):

$$MKK = \frac{DP * DN - YP * YN}{\sqrt{(DP + YP) * (DP + YN) * (DN + YP) * (DN + YN)}}$$

Bu denklemde DP, doğru pozitiflerin sayısı, DN doğru negatiflerin sayısı, YP yanlış pozitiflerin sayısı ve YN yanlış negatiflerin sayısıdır. Paydadaki dört toplamdan herhangi biri sıfırsa, payda isteğe bağlı olarak bire ayarlanabilir; bu, doğru sınırlayıcı değer olarak gösterilebilecek sıfır MKK ile sonuçlanır (Chicco ve Jurman, 2020).

2.2.2. Alıcı İşlem Karakteristiği Eğrisi ve Precision-Recall Eğrisi (PRC)

Çoğu durumda biyokimyasal testlerin performansı değerlendirildiğinde, elde edilen veriler büyük ölçüde çarpık veya dengesizdir, yani test edilen deneklerin çoğu, test edilen hastalık / rahatsızlığa sahip olmayan gruba (sağlıklı) aittir. Tipik hastalık prevalansları yaklaşık %10 civarındadır. Bu, belirli bir hastalığı öneren semptomlarla başvuran hastaların sadece %10'unun bu hastalığa sahip olduğunun teşhis edileceği ve %90'ının ise bu hastalığa sahip olmadığı anlamına gelir (Saito ve Rehmsmeier, 2017).

Bir biyokimyasal testin klinik performansı değerlendirilirken, sıklıkla bir alıcı işlem karakteristiği (AİK) eğrisi kullanılır. AİK eğrisi, bir test veya bir testler grubu için mümkün olan her kesim noktası için duyarlılık ve seçicilik arasındaki bağıntıyı grafik şeklinde gösterir ve AİK eğrisi altındaki alan, söz konusu testin performansı hakkında fikir verir. Son zamanlarda ise AİK eğrisine ek olarak PRC'nin de kullanıldığı ve ikisinin birlikte kullanılmasının klinik olarak daha değerli olduğu vurgulanmaktadır. Özellikle dengesiz veri setlerine dayalı AİK eğrilerinin görsel yorumu ve karşılaştırmaları yanıltıcı olabilir. Bu yüzden dengesiz veri setleri için PRC daha iyi bir seçenek olabilir (Saito ve Rehmsmeier, 2017).

2.2.2.1. AİK Eğrisi

İki kategorili nitel değişkenin olasılığının yordanmasında kullanışlı olan araçlardan biri Alıcı İşlem Karakteristiği eğrisidir. 0 ile 1 arasındaki bir dizi farklı aday eşik değerleri için doğru pozitif orana (y ekseni) karşı yanlış pozitif oranın (x ekseni) bir grafiğidir. AİK eğrisi birkaç sebepten ötürü kullanışlı bir araçtır (Davis ve Goadrich, 2006):

- Farklı modellerin eğrileri genel olarak veya farklı eşik değerleri için doğrudan karşılaştırılabilir.

• Eğri altında kalan alan (EAKA) model performansının bir özeti olarak kullanılabilir.

Eğrinin şekli bir problem için en çok nelere dikkat etmemiz gerektiği, beklenen yanlış pozitif oran ve yanlış negatif oran gibi birçok bilgi içermektedir. Daha detaylı anlatmak gerekirse (Davis ve Goadrich, 2006):

• Grafiğin x eksenindeki daha küçük değerler, daha düşük yanlış pozitif ve daha yüksek doğru negatif değerleri gösterir.

• Grafiğin y eksenindeki daha büyük değerler, daha yüksek doğru pozitif ve daha düşük yanlış negatif değerleri gösterir (Davis ve Goadrich, 2006).

Genel olarak iyi bir model, grafiğin sol üstüne yay çizen eğrilerle temsil edilir. Kötü bir sınıflandırıcı ise, sınıflar arasında ayrım yapamaz ve tüm durumlarda rastgele bir sınıfı veya sabit bir sınıfı yordar. Kötü performansa sahip bir model, (0,5, 0,5) noktasında temsil edilir. Kötü bir model, grafiğin sol altından sağ üstüne çizilen köşegen bir doğruyla temsil edilir ve 0,5'lik bir eğri altında kalan alana sahiptir. İyi bir model ise, (0,1) noktasında temsil edilir. Yani grafiğin sol altından sol üstüne ve sonra üstten üst sağa giden bir doğruyla temsil edilir (Davis ve Goadrich, 2006).

2.2.2.2. Precision-Recall Eğrisi

Bir PRC, mümkün olan her kesim noktası için precision (pozitif prediktif değer) ile recall (duyarlılık) arasındaki ilişkiyi gösterir. PRC, şunlara sahip bir grafiktir (Saito ve Rehmsmeier, 2017):

• Recall (duyarlılık = DP / (DP + YN)) değerini gösteren x ekseni ve

• Precision (pozitif tahmini değer = DP / (DP + YP)) değerini gösteren y ekseni.

Dolayısıyla PRC'deki her nokta, seçilmiş bir kesim noktasını temsil eder. Seçilen kesim noktası da bize precision ve recall değerlerini verir. Verilerden bir PRC oluşturmak için, tüm sonuçlar sıralanır ve her bir değeri, tanıya (hastalık var veya yok) bağlayarak başlanır (Çizelge 5).

Çizelge 5. Teşhisle Birlikte Sıralanmış Veriler.

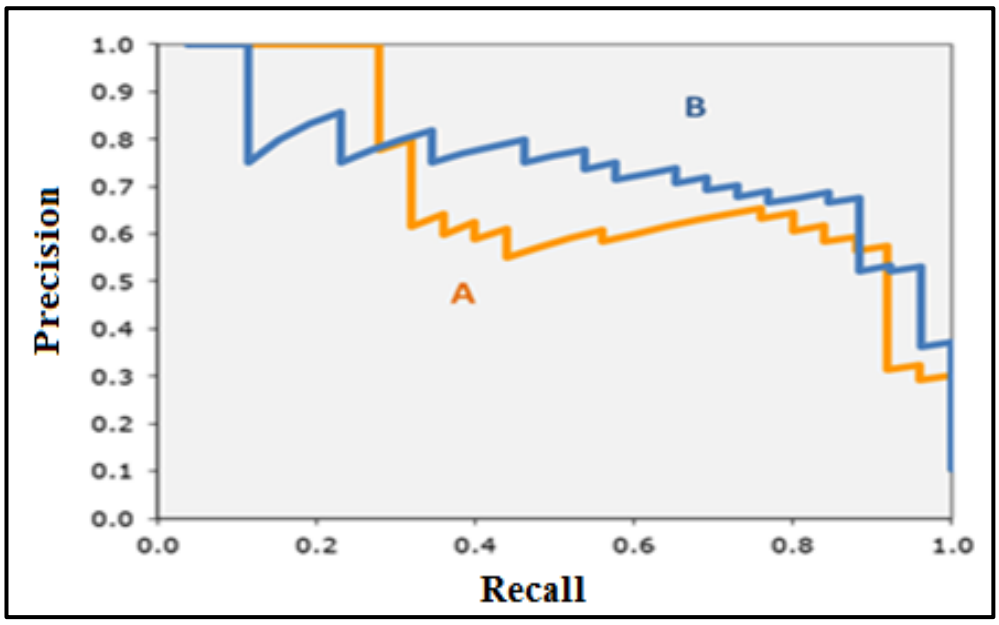
No	Parametre Değeri	Hastalık (Var/Yok)
1	33,63	Var
2	10,63	Var
3	9,90	Yok
4	6,87	Var
5	6,15	Var
6	6,15	Var
7	5,53	Var
8	5,08	Var
…	…	…
…	…	…
151	0,0041	Yok
152	0,0039	Var
153	0,0039	Yok
154	0,0039	Var
155	0,0038	Yok
156	0,0038	Var
157	0,0038	Yok
158	0,0038	Yok
159	0,0036	Yok
160	0,0036	Yok

Her bir parametre için, bu parametre ile aynı veya üzerinde bir sonuç pozitif kabul edildiği takdirde, precision (pozitif prediktif değer) ve recall (duyarlılık) değerlerinin ne olacağı hesaplanır (Çizelge 6).

Çizelge 6. Hesaplanan Precision ve Recall Değerleri ile Birlikte Sıralanmış Veriler.

No	Parametre	Hastalık (Var/Yok)	Var (Toplam)	Yok (Toplam)	Precision	Recall
1	33,63	Var	1	0	1,00	0,013
2	10,63	Var	2	0	1,00	0,025
3	9,90	Yok	2	1	0,67	0,025
4	6,87	Var	3	1	0,75	0,038
5	6,15	Var	4	1	0,80	0,050
6	6,15	Var	5	1	0,83	0,063
7	5,53	Var	6	1	0,86	0,075
8	5,08	Var	7	1	0,88	0,088
…	…	…	…	…	…	…
151	0,0041	Yok	77	74	0,51	0,96
152	0,0039	Var	78	74	0,51	0,98
153	0,0039	Yok	78	75	0,51	0,98
154	0,0039	Var	79	75	0,51	0,99
155	0,0038	Yok	79	76	0,51	0,99
156	0,0038	Var	80	76	0,51	1,00
157	0,0038	Yok	80	77	0,51	1,00
158	0,0038	Yok	80	78	0,51	1,00
159	0,0036	Yok	80	79	0,50	1,00
160	0,0036	Yok	80	80	0,50	1,00

Son olarak ise precision ve recall veri çiftleri kullanılarak eğri oluşturulur (Şekil 3).



Şekil 3. Precision-Recall Eğrisi Örneği.

PRC'ler genellikle zikzak eğrilerdir. Bu nedenle PRC'ler, AİK eğrilerinden daha sık birbirine geçme eğilimindedir. Bu, eğriler arasındaki karşılaştırmaları zorlaştırabilir. Bununla birlikte, mükemmel bir test için üst çizgiye yakın eğriler, taban çizgisine yakın olanlardan daha iyi bir performansa sahiptir. Başka bir deyişle, iki eğri karşılaştırılırken, diğer eğrinin üzerindeki bir eğri daha iyi bir performansa sahiptir (Saito ve Rehmsmeier, 2015).

2.2.2.3. AİK ve Precision-Recall Eğrileri Arasındaki Fark

AİK ve PRC arasındaki esas fark, bir PRC yapmak için doğru negatif sonuçların sayısının kullanılmamasıdır (Saito ve Rehmsmeier, 2015).

Saito ve Rehmsmeier (2015), çalışmalarında AİK eğrilerinin çok popüler olduğunu, popülerliklerinin son on yılda arttığını ve dengesiz verilerde en yaygın kullanılan değerlendirme yöntemi olduğunu tespit etmiştir.

Performans karşılaştırmasına yönelik çalışmalarda birçok araştırmacı, ana değerlendirme yöntemi olarak AİK yerine PRC kulllanmanın birçok çalışmayı etkileyebileceği sonucuna varmaktadır. Bunu da aşağıdaki gibi açıklamaktadırlar (Saito ve Rehmsmeier, 2015):

"Bir çalışmaya hastalığı olmayan ve düşük test sonuçları olan birçok bireyin eklenmesi, AİK eğrisini, değerlendirilen parametrenin duyarlılık veya pozitif prediktif değerinde bir iyileşme olmaksızın anlamlı olarak iyileştirebilmektedir. Precision-Recall eğrileri ise, hastalığı olmayan ve düşük test sonuçları olan hastaların eklenmesinden etkilenmemiştir."

Dolayısıyla, testlerin değerlendirilmesi ve karşılaştırılmasında resmin tamamını görmek amacıyla, rutin olarak kullanılan AİK eğrilerine ek olarak Precision-Recall eğrilerinin kullanılması önerilmektedir. Sonuç olarak AİK eğrileri ve Precision-Recall eğrilerinin kullanımı için öneri (Saito ve Rehmsmeier, 2015):

- AİK eğrileri, her bir sınıf için hemen hemen eşit sayıda gözlemin bulunduğu durumlarda kullanılmalıdır.

- PRC'ler, orta ile büyük sınıf dengesizliğinin bulunduğu durumlarda kullanılmalıdır.

2.3. Sınıf Dengesizliği

Eşit olmayan sınıf dağılımına sahip herhangi bir veri kümesi teknik olarak dengesizdir. Başka bir deyişle, sınıf dengesizliği, bir sınıfı temsil eden örneklem sayısı diğer sınıflarınkinden çok daha düşük/yüksek olduğunda ortaya çıkar. Sınıf dengesizliği literatürde genellikle iki sonuçlu sınıflandırma problemlerini içermektedir, burada bir sınıf diğerinden önemli ölçüde daha fazladır (bu nedenle yeterince temsil edilmez). Bununla birlikte, çarpık sınıf dağılımları ile ilgili çok sınıflı problemler de vardır. Bu problem, ham verilerin toplandığı gerçek dünya uygulamalarının sayısı nedeniyle araştırmacılar ve uygulayıcıların dikkatini çekmiştir. Örneğin, bu sorun anomali tespiti, tıbbi teşhis, yüz tanıma gibi birçok çalışma alanında karşımıza çıkmaktadır (Fernandez ve ark., 2018).

İki sınıflı problemlerde azınlık (yeterince temsil edilmeyen) sınıf genellikle pozitif sınıf olarak adlandırılırken, çoğunluk sınıfı negatif sınıf olarak kabul edilir. Literatürde bu sorunu çözmek için çeşitli yöntemler önerilmektedir. Bu yöntemlerden en sık kullanılanları ise Kümeleme yöntemleri kullanılarak veriyi kümelere bölmek ve analizleri bölünen kümelerde ayrı ayrı uygulayarak sonuçlarını birleştirmektir. Sık kullanılan diğer iki yöntem ise veriye göre under sampling yöntemi kullanarak çok örnekleme sahip veri sayısını azaltmak ya da over sampling yöntemi kullanarak az örnekleme sahip veri sayısını çoğaltmaktır (Fernandez ve ark., 2018).

2.4. Çalışmada Kullanılan Veri Setleri

Çalışmada kullanılan simüle veri setleri ilk aşamada 5 bağımsız nicel ve 1 bağımlı nicel değişken olacak şekilde ve değişkenler arasındaki korelasyonlar 0,25, 0,5 ve 0,75 olarak belirlenerek, çoklu normal dağılımdan üretilmiştir. Daha sonra bağımlı nitel değişken, bağımsız seçilen 5 nicel değişkenin Fuzzy CMeans algoritması kullanılarak kümelere ayrılması ile oluşturulmuştur. Fuzzy CMeans algoritmasında seçilen küme sayıları verileri 0,8-0,2, 0,7-0,3 ve 0,5-0,5 dağılımlara ayıracak şekilde seçilmiş ve bağımsız değişken bu küme sayılarına uygun olarak iki kategoriye ayrılmıştır. Simüle veri setlerinde kullanılacak örneklem büyüklükleri de sırasıyla 250, 500 ve 1000 seçilerek toplamda 27 senaryo üzerinde hibrit modelin performansı veri madenciliği yöntemlerinin performanslarıyla karşılaştırılmıştır. Çalışmada ayrıca UCI veri tabanında bulunan Hepatit ve Meme Kanseri veri setleri kullanılmıştır.

2.5. İstatistiksel Analiz

Verilerin analizinde R 4.0.2 programlama dili ve WEKA 3.7 programlarından faydalanılmıştır. Tanımlayıcı olarak nicel değişkenler için ortalama±standart sapma ve ortanca (minimum-maksimum), nitel değişkenler için ise hasta sayısı (yüzde) verilmiştir. Nicel değişken bakımından iki kategoriye sahip nitel değişkenin kategorileri arasında fark olup olmadığına; normal dağılım varsayımları sağlanmadığı için Mann-Whitney U testi kullanılarak bakılmıştır. İki nitel değişken arasındaki ilişki incelenmek ise istendiğinde Ki-kare ve Fisher-exact testleri kullanılmıştır. İstatistiksel anlamlılık düzeyi 0,05 olarak alınmıştır. Veri madenciliği yöntemlerinin hepsi denenerek her analiz için en iyi sonuç veren 5 yöntem kullanılmıştır. Verilerin analizleri 1000 tekrar üzerinden yapılmıştır ve genelleme yöntemi olarak 10-kat çapraz geçerlik kullanılmıştır. Performans ölçütü olarak doğru sınıflama oranı, F-ölçütü, kesinlik, MKK, PRC alanı ve EAKA kullanılmıştır. R programlama dilinde ise e1071, pROC, RSNNS, JWileymisc, ppclust, factoextra, pheatmap, dplyr ve fclust paketleri kullanılmıştır.

3. BULGULAR

Çalışmada oluşturulan simüle ve gerçek veri setlerine ait analiz sonuçları elde edilmiş, bu sonuçların değişik senaryolarda nasıl değiştiğine bakılmış ve hibrit model ile elde edilen sonuçlar ile karşılaştırmaları yapılmıştır.

3.1. 0,8-0,2 Dağılıma Sahip Simüle Veri Seti için Sonuçlar

Çizelge 7 incelendiğinde; hem veri setinin küçük olması hem de dengesiz olmasından dolayı 2. sınıfa ait performans ölçütlerinin daha düşük olduğu gözlenmektedir. Genel performans ölçütlerine bakıldığında ise doğru sınıflama oranı ve F-ölçütü değeri dengesiz dağılımdan etkilendiği için yüksek bulunmuştur. Bu yüzden dengesiz dağılımlarda MKK ölçütüne göre performans değerlendirmesinin daha objektif olduğu görülmektedir. MKK'ya bakıldığında en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Naive Bayes, AdaBoost, Lojistik Regresyon, Bagging ve Çok Katmanlı Algılayıcı izlemektedir.

Çizelge 7. 250 Hastadan Oluşan Değişkenler Arası 0,25 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,964	0,880	0,809	0,199	0,891	0,699
	2	0,151	0,235	0,533	0,199	0,389	0,699
	Genel	0,792	0,743	0,750	0,199	0,784	0,699
Çok Katmanlı Algılayıcı	**1**	0,904	0,852	0,805	0,118	0,844	0,605
	2	0,189	0,244	0,345	0,118	0,297	0,605
	Genel	0,752	0,723	0,708	0,118	0,728	0,605
Naive Bayes	**1**	0,873	0,856	0,839	0,266	0,890	0,705
	2	0,377	0,408	0,444	0,266	0,399	0,705
	Genel	0,768	0,761	0,755	0,266	0,786	0,705
AdaBoost	**1**	0,898	0,861	0,827	0,233	0,895	0,710
	2	0,302	0,360	0,444	0,233	0,365	0,710
	Genel	0,772	0,755	0,746	0,233	0,782	0,710
Bagging	**1**	0,970	0,878	0,803	0,158	0,860	0,625
	2	0,113	0,185	0,500	0,158	0,324	0,625
	Genel	0,788	0,731	0,738	0,158	0,746	0,625
Hibrit Model	**1**	0,970	0,918	0,872	0,547	0,964	0,721
	2	0,472	0,595	0,806	0,547	0,817	0,721
	Genel	0,864	0,850	0,858	0,547	0,933	0,721

Çizelge 8 incelendiğinde; veri setinin küçük olması nedeniyle 2. sınıfa ait performans ölçütlerinin daha düşük olduğu gözlenmektedir. Genel performans ölçütlerine bakıldığında ise doğru sınıflama oranı ve F-ölçütü değeri dengesiz dağılımdan etkilendiği için yüksek bulunmuştur. Bu yüzden dengesiz dağılımlarda MKK ölçütüne göre performans değerlendirmesinin daha objektif olduğu görülmektedir. MKK'ya bakıldığında en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Naive Bayes, Hoeffding Tree, Lojistik Regresyon, Çok Katmanlı Algılayıcı ve Bagging izlemektedir.

Çizelge 8. 250 Hastadan Oluşan Değişkenler Arası 0,5 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

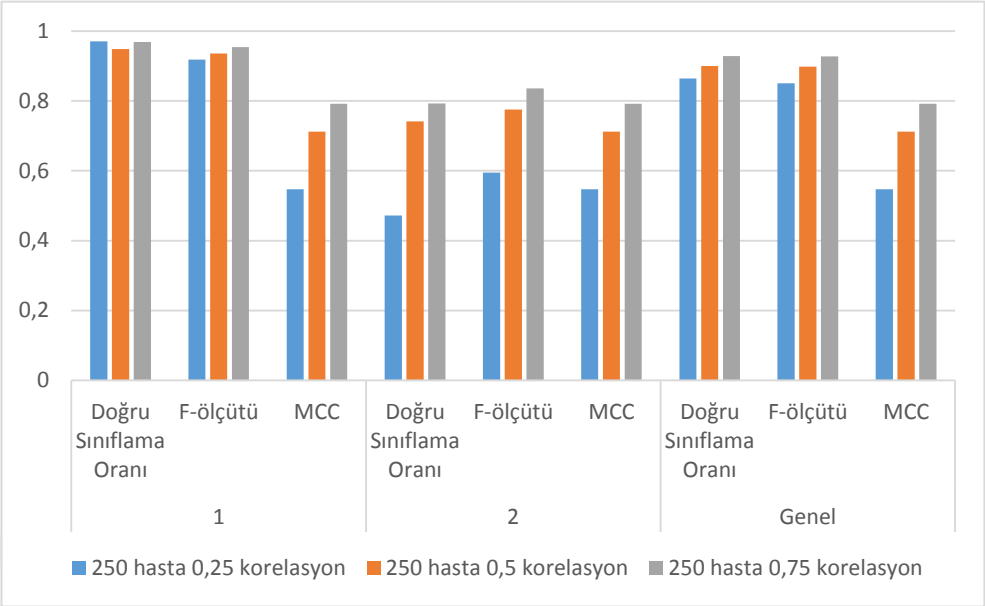
Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,917	0,876	0,838	0,380	0,924	0,802
	2	0,414	0,490	0,600	0,380	0,579	0,802
	Genel	0,800	0,786	0,783	0,380	0,844	0,802
Çok Katmanlı Algılayıcı	**1**	0,901	0,863	0,828	0,320	0,915	0,781
	2	0,379	0,444	0,537	0,320	0,507	0,781
	Genel	0,780	0,766	0,760	0,320	0,821	0,781
Naive Bayes	**1**	0,813	0,855	0,902	0,475	0,928	0,815
	2	0,707	0,607	0,532	0,475	0,611	0,815
	Genel	0,788	0,797	0,816	0,475	0,854	0,815
Bagging	**1**	0,896	0,856	0,819	0,277	0,909	0,781
	2	0,345	0,408	0,500	0,277	0,483	0,781
	Genel	0,768	0,752	0,745	0,277	0,811	0,781
Hoeffding Tree	**1**	0,839	0,852	0,866	0,394	0,907	0,766
	2	0,569	0,541	0,516	0,394	0,528	0,766
	Genel	0,776	0,780	0,784	0,394	0,819	0,766
Hibrit Model	**1**	0,948	0,936	0,924	0,712	0,988	0,845
	2	0,741	0,775	0,811	0,712	0,850	0,845
	Genel	0,900	0,898	0,898	0,712	0,956	0,845

Çizelge 9 incelendiğinde; değişkenler arası (bağımlı ve bağımsız değişkenler) korelasyon yüksek olduğu için 2. sınıfa ait performans ölçütlerinin de 0,25 ve 0,5 korelasyona göre yüksek olduğu gözlenmektedir. Bu yüzden dengesiz dağılımlarda MKK ölçütüne göre performans değerlendirmesinin daha objektif olduğu görülmektedir. MKK'ya bakıldığında en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Çok Katmanlı Algılayıcı, Lojistik Regresyon, Random Forest, Destek Vektör Makinası ve Bagging izlemektedir.

Çizelge 9. 250 Hastadan Oluşan Değişkenler Arası 0,75 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,943	0,928	0,914	0,676	0,975	0,925
	2	0,707	0,745	0,788	0,676	0,820	0,925
	Genel	0,888	0,886	0,885	0,676	0,939	0,925
Çok Katmanlı Algılayıcı	**1**	0,943	0,933	0,923	0,702	0,942	0,872
	2	0,741	0,768	0,796	0,702	0,748	0,872
	Genel	0,896	0,895	0,894	0,702	0,897	0,872
Destek Vektör Makinası	**1**	0,969	0,925	0,886	0,639	0,882	0,777
	2	0,586	0,694	0,850	0,639	0,594	0,777
	Genel	0,880	0,872	0,877	0,639	0,815	0,777
Bagging	**1**	0,943	0,919	0,896	0,622	0,971	0,913
	2	0,638	0,698	0,771	0,622	0,796	0,913
	Genel	0,872	0,868	0,867	0,622	0,930	0,913
Random Forest	**1**	0,943	0,923	0,905	0,649	0,970	0,649
	2	0,672	0,722	0,780	0,649	0,747	0,649
	Genel	0,880	0,877	0,876	0,649	0,918	0,649
Hibrit Model	**1**	0,969	0,954	0,939	0,792	0,998	0,881
	2	0,793	0,836	0,885	0,792	0,859	0,881
	Genel	0,928	0,927	0,927	0,792	0,966	0,881

Çizelge 7, 8 ve 9 birlikte değerlendirildiğinde ise hem kategori bazlı hem de genel performans ölçütlerinin değişkenler arası korelasyon arttıkça arttığı görülmektedir. 0,8-0,2 dağılıma sahip veri seti için hibrit modele ait performans ölçütlerinin karşılaştırılması ise Şekil 4'te gösterilmiştir. Değişkenler arası korelasyon arttıkça dengesiz sınıfa ait performans ölçütlerinin arttığı ve ayrıca tüm sınıflarda korelasyon ile birlikte MKK ölçütü değerlerinin de arttığı görülmektedir (Şekil 4). Çalışmada tüm şekillerde MKK yerine MCC kısaltması kullanılmıştır.



Şekil 4. 250 Hastadan Oluşan Simüle Veri seti için Hibrit Modele ait Korelasyonlara Bağlı Performans Ölçütleri.

Çizelge 10 incelendiğinde; hem korelasyonun düşük olması hem de veri setinin dengesiz olmasından dolayı 2. sınıfa ait performans ölçütlerinin daha düşük olduğu gözlenmektedir. Genel performans ölçütlerine bakıldığında ise doğru sınıflama oranı ve F-ölçütü değeri dengesiz dağılımdan etkilendiği yüksek bulunmuştur. Bu yüzden dengesiz dağılımlarda MKK ölçütüne göre performans değerlendirmesinin daha objektif olduğu görülmektedir. MKK'ya bakıldığında en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Naive Bayes, Random Forest, Bagging, Çok Katmanlı Algılayıcı ve Lojistik Regresyon izlemektedir.

Çizelge 10. 500 Hastadan Oluşan Değişkenler Arası 0,25 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,990	0,876	0,785	0,031	0,344	0,637
	2	0,019	0,035	0,333	0,031	0,852	0,637
	Genel	0,780	0,694	0,688	0,031	0,742	0,637
Çok Katmanlı Algılayıcı	**1**	0,959	0,867	0,792	0,080	0,819	0,584
	2	0,083	0,135	0,360	0,080	0,279	0,584
	Genel	0,770	0,709	0,698	0,080	0,702	0,584
Naive Bayes	**1**	0,936	0,871	0,814	0,219	0,845	0,644
	2	0,222	0,306	0,490	0,219	0,346	0,644
	Genel	0,782	0,749	0,744	0,219	0,737	0,644
Bagging	**1**	0,962	0,871	0,795	0,118	0,839	0,598
	2	0,102	0,164	0,423	0,118	0,298	0,598
	Genel	0,776	0,718	0,715	0,118	0,722	0,598
Random Forest	**1**	0,944	0,868	0,803	0,155	0,837	0,603
	2	0,157	0,231	0,436	0,155	0,336	0,603
	Genel	0,774	0,730	0,723	0,155	0,729	0,603
Hibrit Model	**1**	0,995	0,920	0,855	0,557	0,958	0,692
	2	0,389	0,553	0,955	0,557	0,815	0,692
	Genel	0,864	0,841	0,877	0,557	0,927	0,692

Çizelge 11 incelendiğinde, 2. sınıfa ait performans ölçütlerinin daha düşük olduğu gözlenmektedir. Genel performans ölçütlerine bakıldığında ise doğru sınıflama oranı ve F-ölçütü değeri dengesiz dağılımdan etkilendiği için yüksek bulunmuştur. Bu yüzden dengesiz dağılımlarda MKK ölçütüne göre performans değerlendirmesinin daha objektif olduğu görülmektedir. MKK'ya bakıldığında en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Naive Bayes, Lojistik Regresyon, AdaBoost, J48 ve Çok Katmanlı Algılayıcı izlemektedir.

Çizelge 11. 500 Hastadan Oluşan Değişkenler Arası 0,5 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

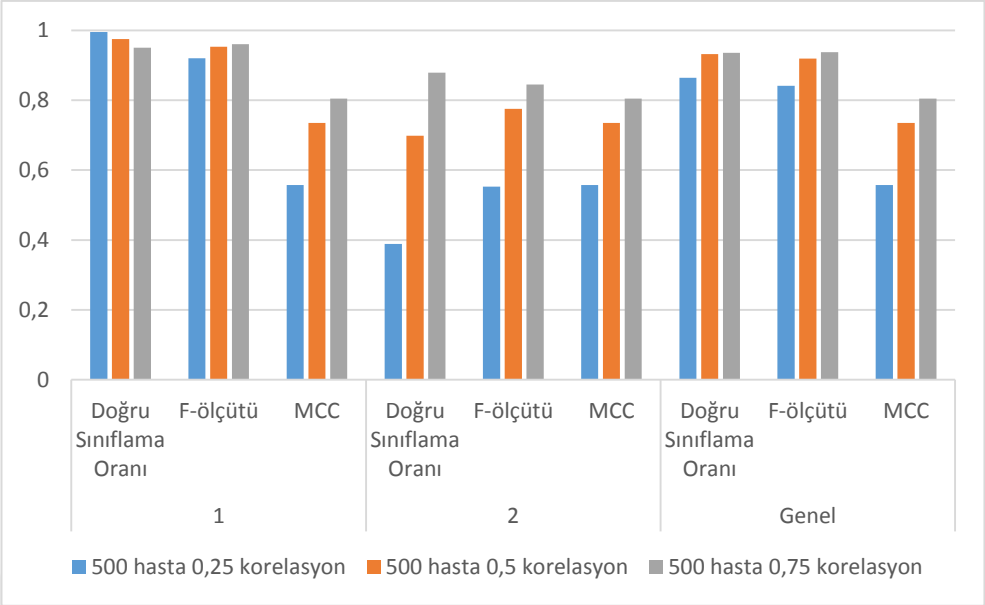
Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,955	0,905	0,860	0,389	0,949	0,824
	2	0,344	0,449	0,647	0,389	0,535	0,824
	Genel	0,838	0,817	0,819	0,389	0,870	0,824
Çok Katmanlı Algılayıcı	**1**	0,931	0,887	0,847	0,278	0,934	0,781
	2	0,292	0,368	0,500	0,278	0,425	0,781
	Genel	0,808	0,787	0,780	0,278	0,836	0,781
Naive Bayes	**1**	0,819	0,864	0,914	0,437	0,946	0,827
	2	0,677	0,556	0,471	0,437	0,557	0,827
	Genel	0,792	0,805	0,829	0,437	0,871	0,827
AdaBoost	**1**	0,911	0,884	0,858	0,311	0,926	0,778
	2	0,365	0,419	0,493	0,311	0,451	0,778
	Genel	0,806	0,794	0,788	0,311	0,834	0,778
J48	**1**	0,916	0,885	0,856	0,310	0,888	0,709
	2	0,354	0,415	0,500	0,310	0,395	0,709
	Genel	0,808	0,795	0,788	0,310	0,794	0,709
Hibrit Model	**1**	0,975	0,953	0,931	0,735	0,995	0,837
	2	0,698	0,775	0,870	0,735	0,828	0,837
	Genel	0,932	0,919	0,920	0,735	0,963	0,837

Çizelge 12 incelendiğinde; değişkenler arası korelasyon yüksek olduğu için 2. sınıfa ait performans ölçütlerinin de 0,25 ve 0,5 korelasyona göre yüksek olduğu gözlenmektedir. Bu yüzden dengesiz dağılımlarda MKK ölçütüne göre performans değerlendirmesinin daha objektif olduğu görülmektedir. MKK'ya bakıldığında en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Naive Bayes, Lojistik Regresyon, AdaBoost, J48 ve Çok Katmanlı Algılayıcı izlemektedir.

Çizelge 12. 500 Hastadan Oluşan Değişkenler Arası 0,75 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,950	0,925	0,901	0,581	0,977	0,915
	2	0,576	0,648	0,740	0,581	0,745	0,915
	Genel	0,876	0,870	0,869	0,581	0,931	0,915
Çok Katmanlı Algılayıcı	**1**	0,925	0,913	0,900	0,535	0,964	0,885
	2	0,586	0,620	0,659	0,535	0,670	0,885
	Genel	0,858	0,855	0,853	0,535	0,905	0,885
Naive Bayes	**1**	0,853	0,902	0,958	0,618	0,978	0,919
	2	0,848	0,694	0,587	0,618	0,756	0,919
	Genel	0,852	0,861	0,885	0,618	0,934	0,919
AdaBoost	**1**	0,933	0,920	0,908	0,574	0,963	0,875
	2	0,616	0,652	0,693	0,574	0,647	0,875
	Genel	0,870	0,867	0,865	0,574	0,900	0,875
J48	**1**	0,930	0,916	0,903	0,553	0,906	0,786
	2	0,596	0,634	0,678	0,553	0,481	0,786
	Genel	0,864	0,861	0,859	0,553	0,822	0,786
Hibrit Model	**1**	0,950	0,960	0,969	0,805	0,998	0,914
	2	0,879	0,845	0,813	0,805	0,906	0,914
	Genel	0,936	0,937	0,939	0,805	0,980	0,914

Çizelge 10, 11 ve 12 birlikte değerlendirildiğinde ise hem kategori bazlı hem de genel performans ölçütlerinin değişkenler arası korelasyon arttıkça arttığı görülmektedir. 0,8-0,2 dağılıma sahip veri seti için hibrit modele ait performans ölçütlerinin karşılaştırılması ise Şekil 5'te gösterilmiştir. Değişkenler arası korelasyon arttıkça dengesiz sınıfa ait performans ölçütlerinin arttığı ve ayrıca tüm sınıflarda korelasyon ile birlikte MKK ölçütü değerlerinin de arttığı görülmektedir (Şekil 5).



Şekil 5. 500 Hastadan Oluşan Simüle Veri seti için Hibrit Modele ait Korelasyonlara Bağlı Performans Ölçütleri.

Çizelge 13 incelendiğinde; hem korelasyonun düşük olması hem de veri setinin dengesiz olmasından dolayı 2. sınıfa ait performans ölçütlerinin daha düşük olduğu gözlenmektedir. Genel performans ölçütlerine bakıldığında ise doğru sınıflama oranı ve F-ölçütü değeri dengesiz dağılımdan etkilendiği yüksek bulunmuştur. Bu yüzden dengesiz dağılımlarda MKK ölçütüne göre performans değerlendirmesinin daha objektif olduğu görülmektedir. MKK'ya bakıldığında en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Naive Bayes, Çok Katmanlı Algılayıcı, Lojistik Regresyon, Bagging ve Random Forest izlemektedir.

Çizelge 13. 1000 Hastadan Oluşan Değişkenler Arası 0,25 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,972	0,871	0,790	0,169	0,876	0,698
	2	0,115	0,190	0,542	0,169	0,408	0,698
	Genel	0,778	0,717	0,734	0,169	0,770	0,698
Çok Katmanlı Algılayıcı	**1**	0,942	0,866	0,802	0,211	0,866	0,665
	2	0,204	0,290	0,505	0,211	0,378	0,665
	Genel	0,775	0,736	0,735	0,211	0,755	0,665
Naive Bayes	**1**	0,898	0,856	0,819	0,253	0,877	0,702
	2	0,319	0,382	0,477	0,253	0,420	0,702
	Genel	0,767	0,749	0,741	0,253	0,774	0,702
Bagging	**1**	0,946	0,860	0,789	0,127	0,844	0,629
	2	0,133	0,201	0,417	0,127	0,328	0,629
	Genel	0,762	0,711	0,705	0,127	0,728	0,629
Random Forest	**1**	0,942	0,858	0,788	0,118	0,822	0,603
	2	0,133	0,199	0,400	0,118	0,316	0,603
	Genel	0,759	0,709	0,700	0,118	0,708	0,603
Hibrit Model	**1**	0,990	0,922	0,863	0,597	0,961	0,725
	2	0,460	0,615	0,929	0,597	0,824	0,725
	Genel	0,870	0,853	0,878	0,597	0,930	0,725

Çizelge 14 incelendiğinde, 2. sınıfa ait performans ölçütlerinin daha düşük olduğu gözlenmektedir. Genel performans ölçütlerine bakıldığında ise doğru sınıflama oranı ve F-ölçütü değeri dengesiz dağılımdan etkilendiği için yüksek bulunmuştur. Bu yüzden dengesiz dağılımlarda MKK ölçütüne göre performans değerlendirmesinin daha objektif olduğu görülmektedir. MKK'ya bakıldığında en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Naive Bayes, Lojistik Regresyon, AdaBoost, Çok Katmanlı Algılayıcı ve J48 izlemektedir.

Çizelge 14. 1000 Hastadan Oluşan Değişkenler Arası 0,5 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

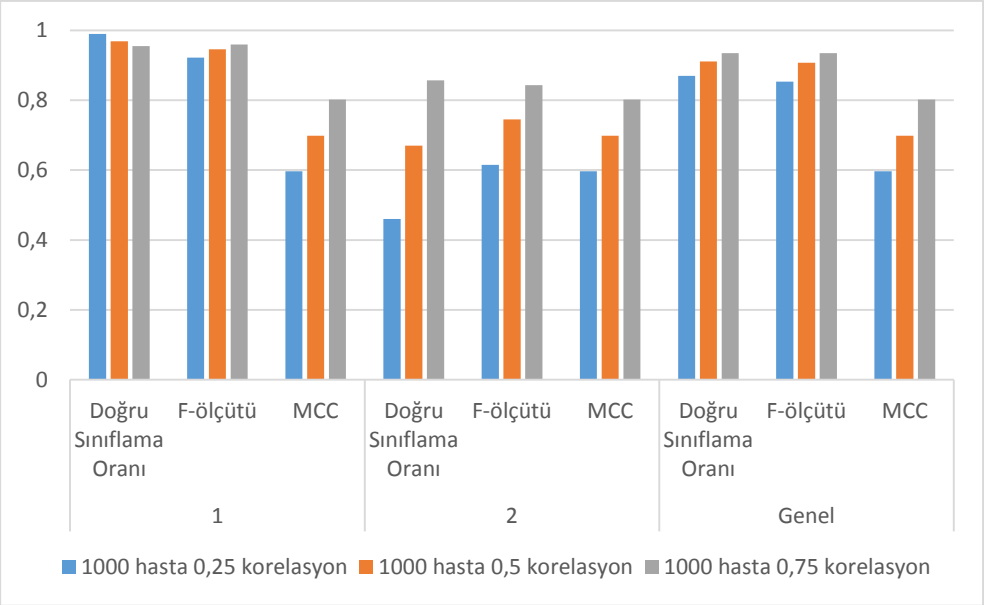
Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,952	0,904	0,861	0,396	0,953	0,834
	2	0,361	0,462	0,642	0,396	0,556	0,834
	Genel	0,837	0,818	0,818	0,396	0,876	0,834
Çok Katmanlı Algılayıcı	**1**	0,943	0,898	0,858	0,365	0,950	0,829
	2	0,351	0,442	0,596	0,365	0,522	0,829
	Genel	0,828	0,810	0,807	0,365	0,867	0,829
Naives Bayes	**1**	0,835	0,867	0,901	0,413	0,954	0,838
	2	0,619	0,537	0,474	0,413	0,567	0,838
	Genel	0,793	0,803	0,818	0,413	0,879	0,838
AdaBoost	**1**	0,919	0,895	0,872	0,396	0,941	0,801
	2	0,438	0,494	0,567	0,396	0,454	0,801
	Genel	0,826	0,817	0,813	0,396	0,847	0,801
J48	**1**	0,952	0,898	0,849	0,335	0,859	0,675
	2	0,299	0,399	0,598	0,335	0,425	0,675
	Genel	0,825	0,801	0,801	0,335	0,775	0,675
Hibrit Model	**1**	0,969	0,946	0,924	0,698	0,991	0,820
	2	0,670	0,745	0,839	0,698	0,826	0,820
	Genel	0,911	0,907	0,908	0,698	0,959	0,820

Çizelge 15 incelendiğinde; değişkenler arası korelasyon yüksek olduğu için 2. sınıfa ait performans ölçütlerinin de 0,25 ve 0,5 korelasyona göre yüksek olduğu gözlenmektedir. Bu yüzden dengesiz dağılımlarda MKK ölçütüne göre performans değerlendirmesinin daha objektif olduğu görülmektedir. MKK'ya bakıldığında en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Lojistik Regresyon, Destek Vektör Makinası, Random Forest, AdaBoost ve Hoeffding Tree izlemektedir.

Çizelge 15. 1000 Hastadan Oluşan Değişkenler Arası 0,75 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

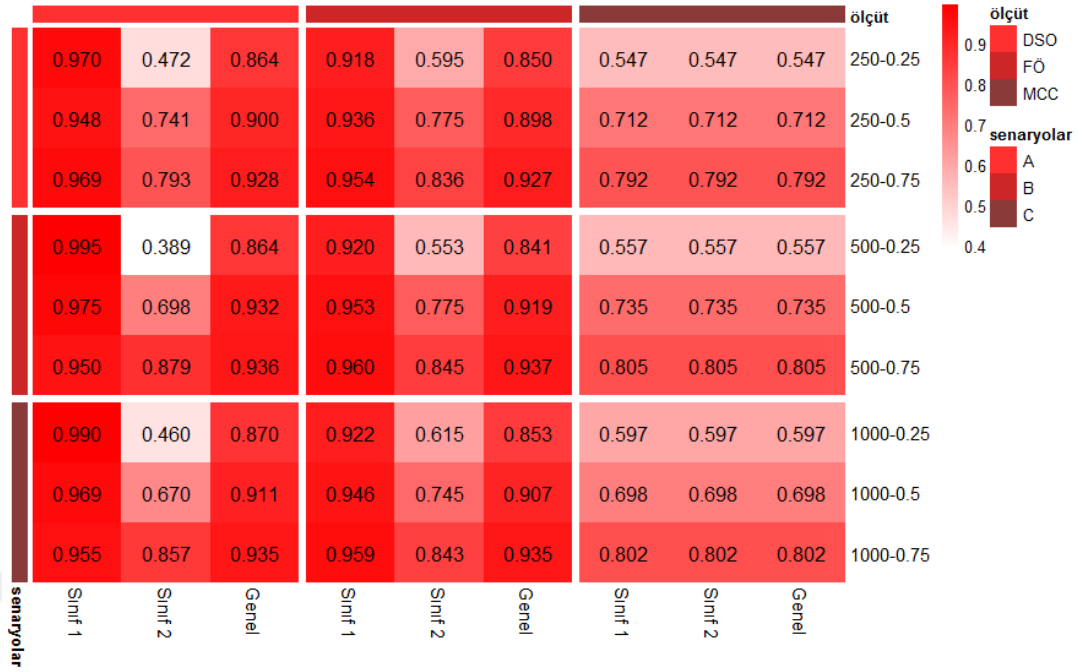
Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,945	0,932	0,919	0,647	0,983	0,936
	2	0,675	0,714	0,757	0,647	0,798	0,936
	Genel	0,890	0,888	0,886	0,647	0,945	0,936
Destek Vektör Makinası	**1**	0,954	0,932	0,911	0,637	0,906	0,795
	2	0,635	0,699	0,777	0,637	0,568	0,795
	Genel	0,889	0,885	0,884	0,637	0,837	0,795
AdaBoost	**1**	0,906	0,920	0,935	0,632	0,979	0,922
	2	0,754	0,710	0,671	0,632	0,733	0,922
	Genel	0,875	0,878	0,882	0,632	0,929	0,922
Hoeffding Tree	**1**	0,859	0,902	0,949	0,608	0,984	0,937
	2	0,818	0,690	0,597	0,608	0,801	0,937
	Genel	0,851	0,859	0,877	0,608	0,947	0,937
Random Forest	**1**	0,951	0,931	0,912	0,635	0,976	0,920
	2	0,640	0,699	0,769	0,635	0,761	0,920
	Genel	0,888	0,884	0,883	0,635	0,932	0,920
Hibrit Model	**1**	0,955	0,959	0,963	0,802	0,998	0,906
	2	0,857	0,843	0,829	0,802	0,895	0,906
	Genel	0,935	0,935	0,936	0,802	0,977	0,906

Çizelge 13, 14 ve 15 birlikte değerlendirildiğinde ise hem kategori bazlı hem de genel performans ölçütlerinin değişkenler arası korelasyon arttıkça arttığı görülmektedir. 0,8-0,2 dağılıma sahip veri seti için hibrit modele ait performans ölçütlerinin karşılaştırılması ise Şekil 6'da gösterilmiştir. Değişkenler arası korelasyon arttıkça dengesiz sınıfa ait performans ölçütlerinin arttığı ve ayrıca tüm sınıflarda korelasyon ile birlikte MKK ölçütü değerlerinin de arttığı görülmektedir (Şekil 6).



Şekil 6. 1000 Hastadan Oluşan Simüle Veri seti için Hibrit Modele ait Korelasyonlara Bağlı Performans Ölçütleri.

0,8-0,2 dağılıma göre üretilen 9 ayrı senaryoya ait performans ölçütlerinin ısı haritası üzerinde gösterimi ise Şekil 7'de verilmiştir.



Şekil 7. 0,8-0,2 Dağılım için Veri Setlerine ait Performans Ölçütlerinin Isı Haritası ile Gösterimi. DSO:Doğru Sınıflama Oranı, FÖ:F-ölçütü, A:250 hastadan oluşan veri setleri, B: 500 hastadan oluşan veri setleri, C:1000 hastadan oluşan veri setleri, 250-0.25:250 hastadan oluşan 0,25 korelasyona sahip veri seti, 250-0.5:250 hastadan oluşan 0,5 korelasyona sahip veri seti, 250-0.75:250 hastadan oluşan 0,75 korelasyona sahip veri seti, 500-0.25:500 hastadan oluşan 0,25 korelasyona sahip veri seti, 500-0.5:500 hastadan oluşan 0,5 korelasyona sahip veri seti, 500-0.75:500 hastadan oluşan 0,75 korelasyona sahip veri seti, 1000-0.25:1000 hastadan oluşan 0,25 korelasyona sahip veri seti, 1000-0.5:1000 hastadan oluşan 0,5 korelasyona sahip veri seti, 1000-0.75:1000 hastadan oluşan 0,75 korelasyona sahip veri seti.

3.2. 0,7-0,3 Dağılıma Sahip Simüle Veri Seti için Sonuçlar

Çizelge 16 incelendiğinde; hem veri setinin küçük olması hem de dengesiz olmasından dolayı 2. sınıfa ait performans ölçütlerinin daha düşük olduğu gözlenmektedir. Genel performans ölçütlerine bakıldığında ise doğru sınıflama oranı ve F-ölçütü değeri dengesiz dağılımdan etkilendiği yüksek bulunmuştur. Bu yüzden dengesiz dağılımlarda MKK ölçütüne göre performans değerlendirmesinin daha objektif olduğu görülmektedir. MKK'ya bakıldığında en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Bagging, Naive Bayes, Random Forest, Çok Katmanlı Algılayıcı ve Lojistik Regresyon izlemektedir.

Çizelge 16. 250 Hastadan Oluşan Değişkenler Arası 0,25 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,943	0,813	0,714	0,109	0,778	0,615
	2	0,120	0,191	0,474	0,109	0,402	0,615
	Genel	0,696	0,626	0,642	0,109	0,665	0,615
Çok Katmanlı Algılayıcı	**1**	0,914	0,804	0,717	0,110	0,720	0,557
	2	0,160	0,235	0,444	0,110	0,373	0,557
	Genel	0,688	0,633	0,636	0,110	0,616	0,557
Naive Bayes	**1**	0,823	0,783	0,746	0,185	0,786	0,629
	2	0,347	0,394	0,456	0,185	0,397	0,629
	Genel	0,680	0,666	0,659	0,185	0,669	0,629
Bagging	**1**	0,897	0,809	0,737	0,194	0,735	0,581
	2	0,253	0,339	0,514	0,194	0,387	0,581
	Genel	0,704	0,668	0,670	0,194	0,630	0,581
Random Forest	**1**	0,886	0,797	0,724	0,129	0,741	0,595
	2	0,213	0,288	0,444	0,129	0,398	0,595
	Genel	0,684	0,644	0,640	0,129	0,638	0,595
Hibrit Model	**1**	0,960	0,891	0,832	0,589	0,940	0,753
	2	0,547	0,667	0,854	0,589	0,839	0,753
	Genel	0,836	0,824	0,838	0,589	0,910	0,753

Çizelge 17 incelendiğinde; veri setinin küçük olması nedeniyle 2. sınıfa ait performans ölçütlerinin daha düşük olduğu gözlenmektedir. Genel performans ölçütlerine bakıldığında ise doğru sınıflama oranı ve F-ölçütü değeri dengesiz dağılımdan etkilendiği için yüksek bulunmuştur. Bu yüzden dengesiz dağılımlarda MKK ölçütüne göre performans değerlendirmesinin daha objektif olduğu görülmektedir. MKK'ya bakıldığında en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Naive Bayes, AdaBoost, Lojistik Regresyon, Random Forest ve Çok Katmanlı Algılayıcı izlemektedir.

Çizelge 17. 250 Hastadan Oluşan Değişkenler Arası 0,5 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

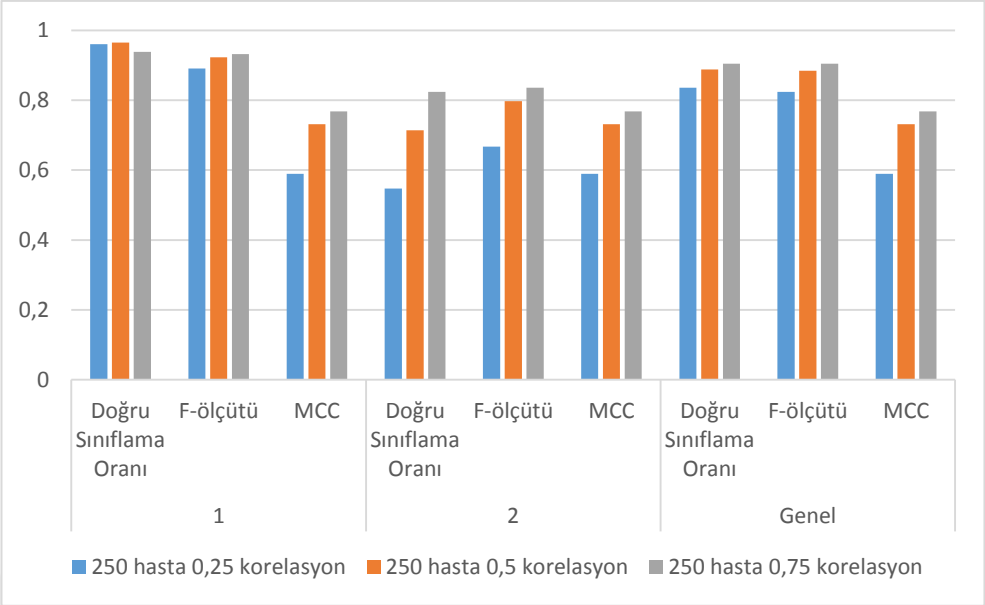
Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,890	0,832	0,782	0,375	0,884	0,779
	2	0,442	0,523	0,642	0,375	0,639	0,779
	Genel	0,752	0,737	0,739	0,375	0,809	0,779
Çok Katmanlı Algılayıcı	**1**	0,850	0,817	0,786	0,351	0,880	0,769
	2	0,481	0,529	0,587	0,351	0,615	0,769
	Genel	0,736	0,728	0,725	0,351	0,798	0,769
Naive Bayes	**1**	0,798	0,814	0,831	0,424	0,894	0,797
	2	0,636	0,609	0,583	0,424	0,658	0,797
	Genel	0,748	0,751	0,755	0,424	0,822	0,797
AdaBoost	**1**	0,873	0,837	0,803	0,419	0,855	0,757
	2	0,519	0,576	0,645	0,419	0,603	0,757
	Genel	0,764	0,756	0,755	0,419	0,778	0,757
Random Forest	**1**	0,873	0,825	0,782	0,360	0,850	0,764
	2	0,455	0,522	0,614	0,360	0,573	0,764
	Genel	0,744	0,732	0,731	0,360	0,765	0,764
Hibrit Model	**1**	0,965	0,923	0,884	0,731	0,968	0,840
	2	0,714	0,797	0,902	0,731	0,866	0,840
	Genel	0,888	0,884	0,889	0,731	0,936	0,840

Çizelge 18 incelendiğinde; değişkenler arası korelasyon yüksek olduğu için 2. sınıfa ait performans ölçütlerinin de 0,25 ve 0,5 korelasyona göre yüksek olduğu gözlenmektedir. Bu yüzden dengesiz dağılımlarda MKK ölçütüne göre performans değerlendirmesinin daha objektif olduğu görülmektedir. MKK'ya bakıldığında en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla AdaBoost, Lojistik Regresyon, Naive Bayes, Çok Katmanlı Algılayıcı ve Destek Vektör Makinası izlemektedir.

Çizelge 18. 250 Hastadan Oluşan Değişkenler Arası 0,75 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,903	0,893	0,883	0,630	0,958	0,904
	2	0,716	0,736	0,757	0,630	0,822	0,904
	Genel	0,848	0,847	0,846	0,630	0,918	0,904
Çok Katmanlı Algılayıcı	**1**	0,943	0,892	0,847	0,597	0,946	0,893
	2	0,595	0,688	0,815	0,597	0,822	0,893
	Genel	0,840	0,832	0,837	0,597	0,909	0,893
Destek Vektör Makinası	**1**	0,920	0,885	0,853	0,579	0,841	0,771
	2	0,622	0,687	0,767	0,579	0,589	0,771
	Genel	0,832	0,826	0,827	0,579	0,766	0,771
Naive Bayes	**1**	0,813	0,861	0,917	0,600	0,961	0,910
	2	0,824	0,726	0,649	0,600	0,831	0,910
	Genel	0,816	0,821	0,837	0,600	0,923	0,910
AdaBoost	**1**	0,841	0,878	0,919	0,634	0,951	0,900
	2	0,824	0,748	0,685	0,634	0,817	0,900
	Genel	0,836	0,840	0,850	0,634	0,911	0,900
Hibrit Model	**1**	0,938	0,932	0,927	0,768	0,986	0,881
	2	0,824	0,836	0,847	0,768	0,879	0,881
	Genel	0,904	0,904	0,903	0,768	0,954	0,881

Çizelge 16, 17 ve 18 birlikte değerlendirildiğinde ise hem kategori bazlı hem de genel performans ölçütlerinin değişkenler arası korelasyon arttıkça arttığı görülmektedir. 0,7-0,3 dağılıma sahip veri seti için hibrit modele ait performans ölçütlerinin karşılaştırılması ise Şekil 8'de gösterilmiştir. Değişkenler arası korelasyon arttıkça dengesiz sınıfa ait performans ölçütlerinin arttığı ve ayrıca tüm sınıflarda korelasyon ile birlikte MKK ölçütü değerlerinin de arttığı görülmektedir (Şekil 8).



Şekil 8. 250 Hastadan Oluşan Simüle Veri seti için Hibrit Modele ait Korelasyonlara Bağlı Performans Ölçütleri.

Çizelge 19 incelendiğinde; hem korelasyonun düşük olması hem de veri setinin dengesiz olmasından dolayı 2. sınıfa ait performans ölçütlerinin daha düşük olduğu gözlenmektedir. Genel performans ölçütlerine bakıldığında ise doğru sınıflama oranı ve F-ölçütü değeri dengesiz dağılımdan etkilendiği yüksek bulunmuştur. Bu yüzden dengesiz dağılımlarda MKK ölçütüne göre performans değerlendirmesinin daha objektif olduğu görülmektedir. MKK'ya bakıldığında en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Lojistik Regresyon, Naive Bayes, Random Forest, Bagging ve Çok Katmanlı Algılayıcı izlemektedir.

Çizelge 19. 500 Hastadan Oluşan Değişkenler Arası 0,25 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,928	0,822	0,737	0,239	0,823	0,699
	2	0,243	0,346	0,597	0,239	0,481	0,699
	Genel	0,720	0,677	0,695	0,239	0,719	0,699
Çok Katmanlı Algılayıcı	**1**	0,879	0,797	0,729	0,162	0,832	0,685
	2	0,250	0,328	0,475	0,162	0,464	0,685
	Genel	0,688	0,654	0,651	0,162	0,720	0,685
Naive Bayes	**1**	0,822	0,786	0,753	0,219	0,821	0,700
	2	0,382	0,426	0,483	0,219	0,480	0,700
	Genel	0,688	0,677	0,671	0,219	0,718	0,700
Bagging	**1**	0,885	0,802	0,733	0,186	0,800	0,675
	2	0,263	0,345	0,500	0,186	0,467	0,675
	Genel	0,696	0,663	0,662	0,186	0,699	0,675
Random Forest	**1**	0,862	0,796	0,739	0,194	0,816	0,663
	2	0,303	0,374	0,489	0,194	0,446	0,663
	Genel	0,692	0,668	0,663	0,194	0,704	0,663
Hibrit Model	**1**	0,945	0,905	0,869	0,666	0,958	0,812
	2	0,678	0,752	0,844	0,666	0,856	0,812
	Genel	0,863	0,858	0,861	0,666	0,927	0,812

Çizelge 20 incelendiğinde, 2. sınıfa ait performans ölçütlerinin daha düşük olduğu gözlenmektedir. Genel performans ölçütlerine bakıldığında ise doğru sınıflama oranı ve F-ölçütü değeri dengesiz dağılımdan etkilendiği için yüksek bulunmuştur. Bu yüzden dengesiz dağılımlarda MKK ölçütüne göre performans değerlendirmesinin daha objektif olduğu görülmektedir. MKK'ya bakıldığında en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Naive Bayes, Hoeffding Tree, AdaBoost, Lojistik Regresyon ve Destek Vektör Makinası izlemektedir.

Çizelge 20. 500 Hastadan Oluşan Değişkenler Arası 0,5 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

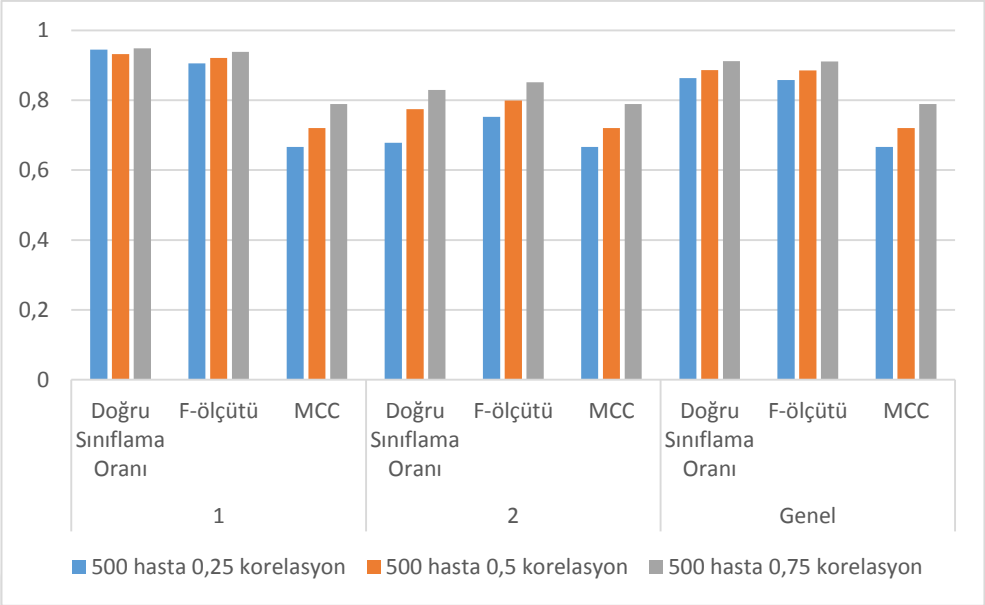
Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,895	0,844	0,798	0,391	0,912	0,812
	2	0,452	0,530	0,641	0,391	0,636	0,812
	Genel	0,766	0,752	0,752	0,391	0,832	0,812
Destek Vektör Makinası	**1**	0,924	0,846	0,780	0,362	0,775	0,647
	2	0,370	0,476	0,667	0,362	0,431	0,647
	Genel	0,762	0,738	0,747	0,362	0,674	0,647
Naive Bayes	**1**	0,805	0,831	0,858	0,465	0,916	0,818
	2	0,678	0,631	0,589	0,465	0,639	0,818
	Genel	0,768	0,772	0,780	0,465	0,835	0,818
AdaBoost	**1**	0,845	0,838	0,831	0,432	0,875	0,761
	2	0,582	0,594	0,607	0,432	0,546	0,761
	Genel	0,768	0,767	0,765	0,432	0,779	0,761
Hoeffding Tree	**1**	0,808	0,830	0,854	0,457	0,917	0,819
	2	0,664	0,624	0,588	0,457	0,645	0,819
	Genel	0,766	0,770	0,776	0,457	0,837	0,819
Hibrit Model	**1**	0,932	0,921	0,909	0,720	0,976	0,853
	2	0,774	0,799	0,825	0,720	0,868	0,853
	Genel	0,886	0,885	0,884	0,720	0,945	0,853

Çizelge 21 incelendiğinde; değişkenler arası korelasyon yüksek olduğu için 2. sınıfa ait performans ölçütlerinin de 0,25 ve 0,5 korelasyona göre yüksek olduğu gözlenmektedir. Bu yüzden dengesiz dağılımlarda MKK ölçütüne göre performans değerlendirmesinin daha objektif olduğu görülmektedir. MKK'ya bakıldığında en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Destek Vektör Makinası, Hoeffding Tree, Lojistik Regresyon, Naive Bayes ve Çok Katmanlı Algılayıcı izlemektedir.

Çizelge 21. 500 Hastadan Oluşan Değişkenler Arası 0,75 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,911	0,897	0,883	0,649	0,961	0,919
	2	0,724	0,751	0,780	0,649	0,860	0,919
	Genel	0,854	0,852	0,852	0,649	0,931	0,919
Çok Katmanlı Algılayıcı	**1**	0,922	0,898	0,875	0,645	0,948	0,903
	2	0,697	0,744	0,797	0,645	0,833	0,903
	Genel	0,854	0,851	0,851	0,645	0,913	0,903
Destek Vektör Makinası	**1**	0,940	0,902	0,867	0,652	0,857	0,805
	2	0,671	0,742	0,829	0,652	0,656	0,805
	Genel	0,858	0,853	0,856	0,652	0,796	0,805
Naive Bayes	**1**	0,845	0,880	0,919	0,646	0,962	0,920
	2	0,829	0,759	0,700	0,646	0,861	0,920
	Genel	0,840	0,843	0,852	0,646	0,931	0,920
Hoeffding Tree	**1**	0,845	0,882	0,922	0,651	0,962	0,920
	2	0,836	0,763	0,702	0,651	0,861	0,920
	Genel	0,842	0,845	0,855	0,651	0,931	0,920
Hibrit Model	**1**	0,948	0,938	0,927	0,789	0,987	0,889
	2	0,829	0,851	0,875	0,789	0,883	0,889
	Genel	0,912	0,911	0,911	0,789	0,956	0,889

Çizelge 19, 20 ve 21 birlikte değerlendirildiğinde ise hem kategori bazlı hem de genel performans ölçütlerinin değişkenler arası korelasyon arttıkça arttığı görülmektedir. 0,7-0,3 dağılıma sahip veri seti için hibrit modele ait performans ölçütlerinin karşılaştırılması ise Şekil 9'da gösterilmiştir. Değişkenler arası korelasyon arttıkça dengesiz sınıfa ait performans ölçütlerinin arttığı ve ayrıca tüm sınıflarda korelasyon ile birlikte MKK ölçütü değerlerinin de arttığı görülmektedir (Şekil 9).



Şekil 9. 500 Hastadan Oluşan Simüle Veri seti için Hibrit Modele ait Korelasyonlara Bağlı Performans Ölçütleri.

Çizelge 22 incelendiğinde; hem korelasyonun düşük olması hem de veri setinin dengesiz olmasından dolayı 2. sınıfa ait performans ölçütlerinin daha düşük olduğu gözlenmektedir. Genel performans ölçütlerine bakıldığında ise doğru sınıflama oranı ve F-ölçütü değeri dengesiz dağılımdan etkilendiği yüksek bulunmuştur. Bu yüzden dengesiz dağılımlarda MKK ölçütüne göre performans değerlendirmesinin daha objektif olduğu görülmektedir. MKK'ya bakıldığında en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Naive Bayes, Lojistik Regresyon, Random Forest, Bagging ve AdaBoost izlemektedir.

Çizelge 22. 1000 Hastadan Oluşan Değişkenler Arası 0,25 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,921	0,812	0,727	0,193	0,832	0,696
	2	0,216	0,309	0,545	0,193	0,499	0,696
	Genel	0,705	0,658	0,671	0,193	0,730	0,696
Naive Bayes	**1**	0,813	0,781	0,751	0,215	0,834	0,696
	2	0,389	0,429	0,478	0,215	0,497	0,696
	Genel	0,683	0,673	0,667	0,215	0,731	0,696
AdaBoost	**1**	0,866	0,792	0,730	0,170	0,807	0,657
	2	0,275	0,348	0,475	0,170	0,435	0,657
	Genel	0,685	0,656	0,652	0,170	0,693	0,657
Bagging	**1**	0,889	0,801	0,729	0,180	0,798	0,647
	2	0,252	0,335	0,500	0,180	0,433	0,647
	Genel	0,694	0,659	0,659	0,180	0,686	0,647
Random Forest	**1**	0,895	0,804	0,730	0,185	0,813	0,658
	2	0,248	0,334	0,510	0,185	0,439	0,658
	Genel	0,697	0,660	0,663	0,185	0,699	0,658
Hibrit Model	**1**	0,973	0,912	0,858	0,683	0,955	0,803
	2	0,634	0,748	0,911	0,683	0,855	0,803
	Genel	0,869	0,861	0,874	0,683	0,925	0,803

Çizelge 23 incelendiğinde, 2. sınıfa ait performans ölçütlerinin daha düşük olduğu gözlenmektedir. Genel performans ölçütlerine bakıldığında ise doğru sınıflama oranı ve F-ölçütü değeri dengesiz dağılımdan etkilendiği için yüksek bulunmuştur. Bu yüzden dengesiz dağılımlarda MKK ölçütüne göre performans değerlendirmesinin daha objektif olduğu görülmektedir. MKK'ya bakıldığında en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Hoeffding Tree, Destek Vektör Makinası, Lojistik Regresyon, AdaBoost ve Çok Katmanlı Algılayıcı izlemektedir.

Çizelge 23. 1000 Hastadan Oluşan Değişkenler Arası 0,5 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

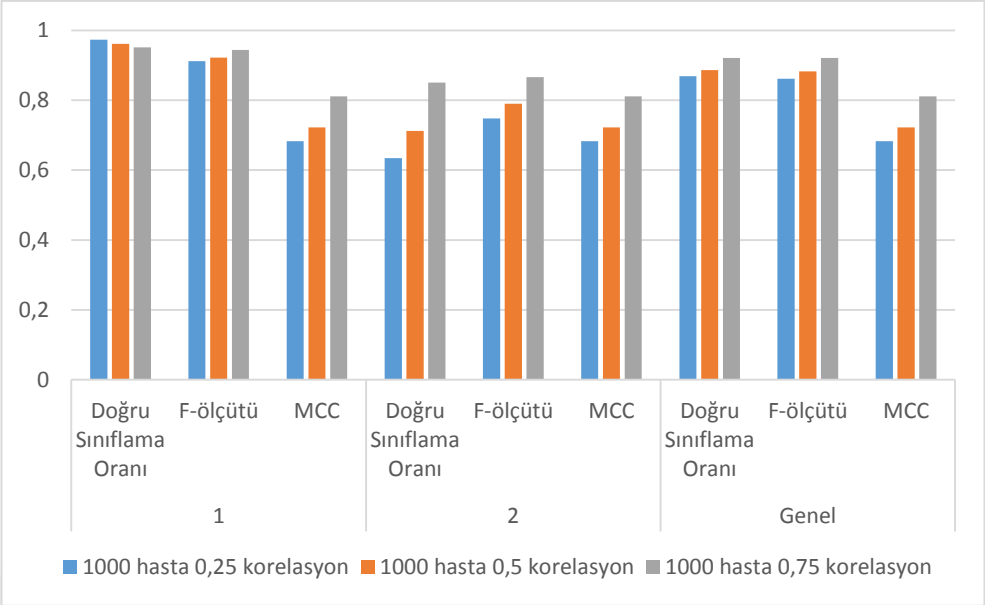
Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,894	0,835	0,784	0,370	0,897	0,797
	2	0,430	0,514	0,637	0,370	0,647	0,797
	Genel	0,754	0,738	0,740	0,370	0,822	0,797
Çok Katmanlı Algılayıcı	**1**	0,913	0,837	0,773	0,354	0,884	0,776
	2	0,381	0,481	0,653	0,354	0,598	0,776
	Genel	0,752	0,730	0,737	0,354	0,797	0,776
Destek Vektör Makinası	**1**	0,940	0,848	0,773	0,386	0,768	0,650
	2	0,361	0,481	0,722	0,386	0,454	0,650
	Genel	0,765	0,737	0,757	0,386	0,673	0,650
AdaBoost	**1**	0,874	0,830	0,790	0,369	0,876	0,772
	2	0,464	0,528	0,614	0,369	0,606	0,772
	Genel	0,750	0,739	0,737	0,369	0,794	0,772
Hoeffding Tree	**1**	0,787	0,806	0,827	0,394	0,897	0,800
	2	0,619	0,586	0,557	0,394	0,649	0,800
	Genel	0,736	0,740	0,745	0,394	0,822	0,800
Hibrit Model	**1**	0,961	0,922	0,885	0,722	0,968	0,837
	2	0,712	0,790	0,888	0,722	0,865	0,837
	Genel	0,886	0,882	0,886	0,722	0,937	0,837

Çizelge 24 incelendiğinde; değişkenler arası korelasyon yüksek olduğu için 2. sınıfa ait performans ölçütlerinin de 0,25 ve 0,5 korelasyona göre yüksek olduğu gözlenmektedir. Bu yüzden dengesiz dağılımlarda MKK ölçütüne göre performans değerlendirmesinin daha objektif olduğu görülmektedir. MKK'ya bakıldığında en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Hoeffding Tree, Lojistik Regresyon, Bagging, Destek Vektör Makinası ve Çok Katmanlı Algılayıcı izlemektedir.

Çizelge 24. 1000 Hastadan Oluşan Değişkenler Arası 0,75 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

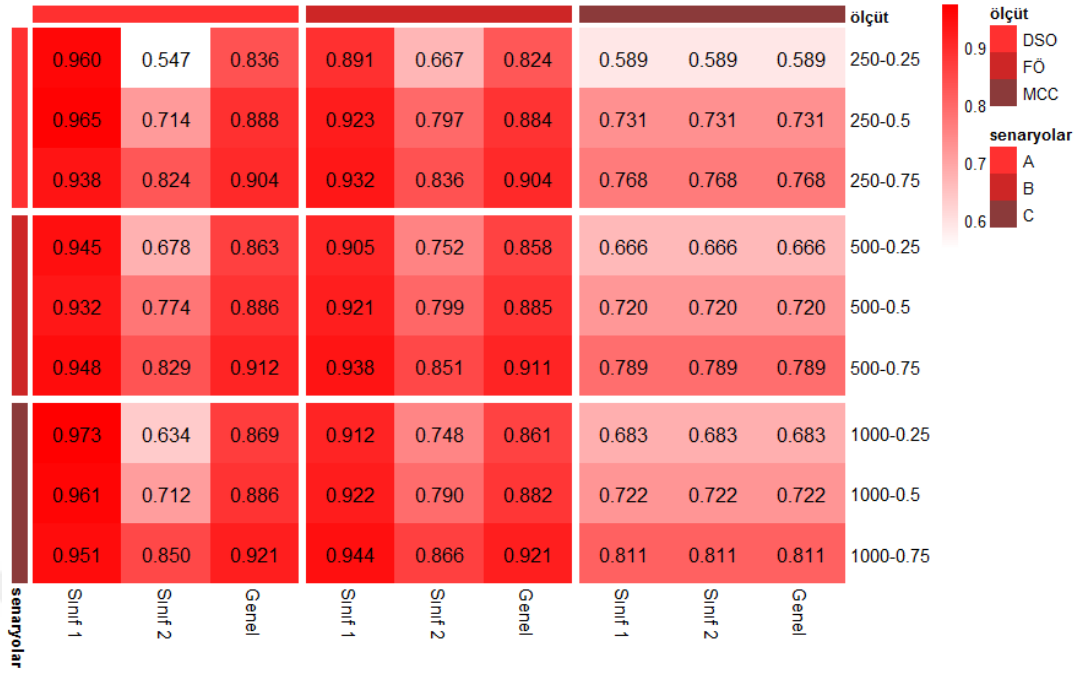
Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,917	0,890	0,864	0,610	0,954	0,904
	2	0,664	0,716	0,775	0,610	0,818	0,904
	Genel	0,841	0,837	0,837	0,610	0,913	0,904
Çok Katmanlı Algılayıcı	**1**	0,923	0,886	0,852	0,589	0,947	0,894
	2	0,628	0,695	0,778	0,589	0,795	0,894
	Genel	0,834	0,828	0,830	0,589	0,901	0,894
Destek Vektör Makinası	**1**	0,924	0,889	0,857	0,602	0,845	0,783
	2	0,641	0,706	0,785	0,602	0,611	0,783
	Genel	0,839	0,834	0,835	0,602	0,774	0,783
Bagging	**1**	0,918	0,889	0,861	0,604	0,946	0,890
	2	0,654	0,710	0,776	0,604	0,788	0,890
	Genel	0,839	0,835	0,835	0,604	0,898	0,890
Hoeffding Tree	**1**	0,843	0,878	0,916	0,635	0,956	0,906
	2	0,821	0,751	0,692	0,635	0,819	0,906
	Genel	0,836	0,840	0,849	0,635	0,914	0,906
Hibrit Model	**1**	0,951	0,944	0,937	0,811	0,993	0,901
	2	0,850	0,866	0,883	0,811	0,887	0,901
	Genel	0,921	0,921	0,920	0,811	0,961	0,901

Çizelge 22, 23 ve 24 birlikte değerlendirildiğinde ise hem kategori bazlı hem de genel performans ölçütlerinin değişkenler arası korelasyon arttıkça arttığı görülmektedir. 0,7-0,3 dağılıma sahip veri seti için hibrit modele ait performans ölçütlerinin karşılaştırılması ise Şekil 10'da gösterilmiştir. Değişkenler arası korelasyon arttıkça dengesiz sınıfa ait performans ölçütlerinin arttığı ve ayrıca tüm sınıflarda korelasyon ile birlikte MKK ölçütü değerlerinin de arttığı görülmektedir (Şekil 10).



Şekil 10. 1000 Hastadan Oluşan Simüle Veri seti için Hibrit Modele ait Korelasyonlara Bağlı Performans Ölçütleri.

0,7-0,3 dağılıma göre üretilen 9 ayrı senaryoya ait performans ölçütlerinin ısı haritası üzerinde gösterimi ise Şekil 11'de verilmiştir.



Şekil 11. 0,7-0,3 Dağılım için Veri Setlerine ait Performans Ölçütlerinin Isı Haritası ile Gösterimi. DSO:Doğru Sınıflama Oranı, FÖ:F-ölçütü, A:250 hastadan oluşan veri setleri, B: 500 hastadan oluşan veri setleri, C:1000 hastadan oluşan veri setleri, 250-0.25:250 hastadan oluşan 0,25 korelasyona sahip veri seti, 250-0.5:250 hastadan oluşan 0,5 korelasyona sahip veri seti, 250-0.75:250 hastadan oluşan 0,75 korelasyona sahip veri seti, 500-0.25:500 hastadan oluşan 0,25 korelasyona sahip veri seti, 500-0.5:500 hastadan oluşan 0,5 korelasyona sahip veri seti, 500-0.75:500 hastadan oluşan 0,75 korelasyona sahip veri seti, 1000-0.25:1000 hastadan oluşan 0,25 korelasyona sahip veri seti, 1000-0.5:1000 hastadan oluşan 0,5 korelasyona sahip veri seti, 1000-0.75:1000 hastadan oluşan 0,75 korelasyona sahip veri seti.

3.3. 0,5-0,5 Dağılıma Sahip Simüle Veri Seti için Sonuçlar

Çizelge 25 incelendiğinde; hem veri setinin küçük olması hem de değişkenler arası korelasyonun düşük olmasından dolayı algoritmaların düşük performansa sahip olduğu gözlenmektedir. Veri setinin dengeli olması nedeniyle doğru sınıflama oranı, F-ölçütü ve MKK değerleri daha geçerli ölçütler olduğu için veri madenciliği algoritmalarının performanslarını değerlendirmek için seçilmiştir. Ancak bu ölçütlerin benzer sonuçlar vermesi durumunda diğer performans ölçütleri de değerlendirmeye dahil edilecektir. Tüm ölçütler birlikte değerlendirildiğinde en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Lojistik Regresyon, Naive Bayes, Hoeffding Tree, AdaBoost ve Bagging izlemektedir.

Çizelge 25. 250 Hastadan Oluşan Değişkenler Arası 0,25 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,598	0,603	0,608	0,200	0,605	0,623
	2	0,602	0,597	0,592	0,200	0,589	0,623
	Genel	0,600	0,600	0,600	0,200	0,597	0,623
Naive Bayes	**1**	0,606	0,602	0,597	0,184	0,612	0,626
	2	0,577	0,582	0,587	0,184	0,589	0,626
	Genel	0,592	0,592	0,592	0,184	0,601	0,626
AdaBoost	**1**	0,622	0,608	0,594	0,183	0,554	0,582
	2	0,561	0,575	0,590	0,183	0,560	0,582
	Genel	0,592	0,592	0,592	0,183	0,557	0,582
Hoeffding Tree	**1**	0,606	0,602	0,597	0,184	0,612	0,625
	2	0,577	0,582	0,587	0,184	0,589	0,625
	Genel	0,592	0,592	0,592	0,184	0,601	0,625
Bagging	**1**	0,591	0,588	0,586	0,160	0,592	0,575
	2	0,569	0,571	0,574	0,160	0,544	0,575
	Genel	0,580	0,580	0,580	0,160	0,568	0,575
Hibrit Model	**1**	0,780	0,832	0,892	0,686	0,861	0,841
	2	0,902	0,847	0,799	0,686	0,921	0,841
	Genel	0,840	0,840	0,846	0,686	0,892	0,841

Çizelge 26 incelendiğinde; veri setinin küçük olmasından dolayı algoritmaların düşük performansa sahip olduğu gözlenmektedir. Veri setinin dengeli olması nedeniyle doğru sınıflama oranı, F-ölçütü ve MKK değerleri daha geçerli ölçütler olduğu için veri madenciliği algoritmalarının performanslarını değerlendirmek için seçilmiştir. Ancak bu ölçütlerin benzer sonuçlar vermesi durumunda diğer performans ölçütleri de değerlendirmeye dahil edilecektir. Tüm ölçütler birlikte değerlendirildiğinde en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Naive Bayes, Hoeffding Tree, Destek Vektör Makinası, Çok Katmanlı Algılayıcı ve Lojistik Regresyon izlemektedir.

Çizelge 26. 250 Hastadan Oluşan Değişkenler Arası 0,5 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

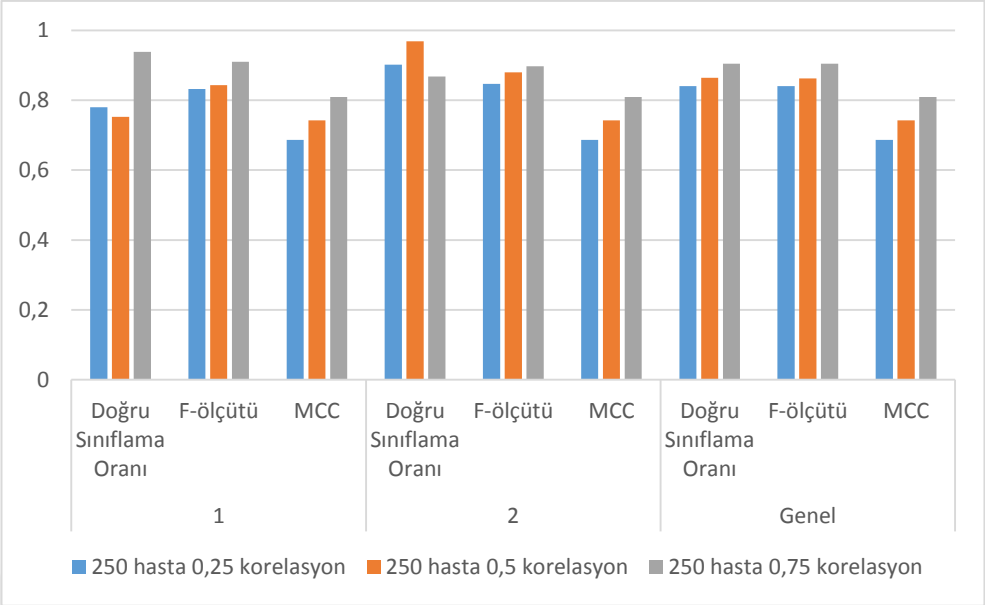
Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,702	0,685	0,669	0,377	0,785	0,780
	2	0,674	0,690	0,707	0,377	0,775	0,780
	Genel	0,688	0,688	0,689	0,377	0,780	0,780
Naive Bayes	**1**	0,702	0,702	0,702	0,423	0,795	0,791
	2	0,721	0,721	0,721	0,423	0,789	0,791
	Genel	0,712	0,712	0,712	0,423	0,792	0,791
Çok Katmanlı Algılayıcı	**1**	0,645	0,672	0,703	0,391	0,776	0,764
	2	0,744	0,716	0,691	0,391	0,765	0,764
	Genel	0,696	0,695	0,696	0,391	0,770	0,764
Hoeffding Tree	**1**	0,702	0,702	0,702	0,423	0,794	0,790
	2	0,721	0,721	0,721	0,423	0,788	0,790
	Genel	0,712	0,712	0,712	0,423	0,791	0,790
Destek Vektör Makinası	**1**	0,702	0,702	0,702	0,423	0,637	0,712
	2	0,721	0,721	0,721	0,423	0,664	0,712
	Genel	0,712	0,712	0,712	0,423	0,651	0,712
Hibrit Model	**1**	0,752	0,843	0,958	0,742	0,886	0,861
	2	0,969	0,880	0,806	0,742	0,945	0,861
	Genel	0,864	0,862	0,880	0,742	0,915	0,861

Çizelge 27 incelendiğinde; hem veri setinin dengeli ve yeterli büyüklükte olmasından hem de bağımsız değişkenler arası ilişkinin yeterli büyüklükte olmasından dolayı doğru sınıflama oranı, F-ölçütü ve MKK değerleri daha geçerli ölçütler olduğu için veri madenciliği algoritmalarının performanslarını değerlendirmek için seçilmiştir. Ancak bu ölçütlerin benzer sonuçlar vermesi durumunda diğer performans ölçütleri de değerlendirmeye dahil edilecektir. Tüm ölçütler birlikte değerlendirildiğinde en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Naive Bayes, Hoeffding Tree, Destek Vektör Makinası, Lojistik Regresyon ve Bagging izlemektedir.

Çizelge 27. 250 Hastadan Oluşan Değişkenler Arası 0,75 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,814	0,811	0,808	0,608	0,888	0,890
	2	0,793	0,797	0,800	0,608	0,902	0,890
	Genel	0,804	0,804	0,804	0,608	0,895	0,890
Naive Bayes	**1**	0,845	0,835	0,826	0,656	0,887	0,894
	2	0,810	0,820	0,831	0,656	0,910	0,894
	Genel	0,828	0,828	0,828	0,656	0,898	0,894
Bagging	**1**	0,814	0,808	0,802	0,599	0,837	0,861
	2	0,785	0,792	0,798	0,599	0,852	0,861
	Genel	0,808	0,800	0,800	0,599	0,844	0,861
Hoeffding Tree	**1**	0,845	0,835	0,826	0,656	0,887	0,894
	2	0,810	0,820	0,831	0,656	0,910	0,894
	Genel	0,828	0,828	0,828	0,656	0,898	0,894
Destek Vektör Makinası	**1**	0,853	0,830	0,809	0,640	0,766	0,819
	2	0,785	0,809	0,833	0,640	0,758	0,819
	Genel	0,820	0,820	0,821	0,640	0,762	0,819
Hibrit Model	**1**	0,938	0,910	0,883	0,809	0,957	0,903
	2	0,868	0,897	0,929	0,809	0,893	0,903
	Genel	0,904	0,904	0,905	0,809	0,927	0,903

Çizelge 25, 26 ve 27 birlikte değerlendirildiğinde ise hem kategori bazlı hem de genel performans ölçütlerinin değişkenler arası korelasyon arttıkça arttığı görülmektedir. 0,5-0,5 dağılıma sahip veri seti için hibrit modele ait performans ölçütlerinin karşılaştırılması ise Şekil 12'de gösterilmiştir. Değişkenler arası korelasyon arttıkça dengesiz sınıfa ait performans ölçütlerinin arttığı ve ayrıca tüm sınıflarda korelasyon ile birlikte MKK ölçütü değerlerinin de arttığı görülmektedir (Şekil 12).



Şekil 12. 250 Hastadan Oluşan Simüle Veri seti için Hibrit Modele ait Korelasyonlara Bağlı Performans Ölçütleri.

Çizelge 28 incelendiğinde; bağımsız değişkenler arası korelasyonun düşük olmasından dolayı performans ölçütlerinin daha düşük olduğu gözlenmektedir. Veri setinin dengeli olması nedeniyle doğru sınıflama oranı, F-ölçütü ve MKK değerleri daha geçerli ölçütler olduğu için veri madenciliği algoritmalarının performanslarını değerlendirmek için seçilmiştir. Ancak bu ölçütlerin benzer sonuçlar vermesi durumunda diğer performans ölçütleri de değerlendirmeye dahil edilecektir. Tüm ölçütler birlikte değerlendirildiğinde en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Naive Bayes, Hoeffding Tree, Lojistik Regresyon, Destek Vektör Makinası ve Çok Katmanlı Algılayıcı izlemektedir.

Çizelge 28. 500 Hastadan Oluşan Değişkenler Arası 0,25 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,671	0,657	0,645	0,304	0,672	0,693
	2	0,633	0,646	0,660	0,304	0,683	0,693
	Genel	0,652	0,652	0,652	0,304	0,678	0,693
Naive Bayes	**1**	0,703	0,672	0,643	0,318	0,670	0,693
	2	0,614	0,643	0,675	0,318	0,685	0,693
	Genel	0,658	0,657	0,659	0,318	0,678	0,693
Çok Katmanlı Algılayıcı	**1**	0,651	0,635	0,621	0,256	0,653	0,668
	2	0,606	0,620	0,636	0,256	0,650	0,668
	Genel	0,628	0,628	0,628	0,256	0,652	0,668
Hoeffding Tree	**1**	0,703	0,672	0,643	0,318	0,670	0,691
	2	0,614	0,643	0,675	0,318	0,684	0,691
	Genel	0,658	0,657	0,659	0,318	0,677	0,691
Destek Vektör Makinası	**1**	0,679	0,653	0,628	0,281	0,586	0,640
	2	0,602	0,627	0,654	0,281	0,593	0,640
	Genel	0,640	0,639	0,641	0,281	0,590	0,640
Hibrit Model	**1**	0,859	0,844	0,829	0,684	0,909	0,842
	2	0,825	0,840	0,855	0,684	0,853	0,842
	Genel	0,842	0,842	0,842	0,684	0,879	0,842

Çizelge 29 incelendiğinde; hem veri setinin dengeli ve yeterli büyüklükte olmasından hem de bağımsız değişkenler arası ilişkinin yeterli büyüklükte olmasından dolayı doğru sınıflama oranı, F-ölçütü ve MKK değerleri daha geçerli ölçütler olduğu için veri madenciliği algoritmalarının performanslarını değerlendirmek için seçilmiştir. Ancak bu ölçütlerin benzer sonuçlar vermesi durumunda diğer performans ölçütleri de değerlendirmeye dahil edilecektir. Tüm ölçütler birlikte değerlendirildiğinde en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Çok Katmanlı Algılayıcı, Naive Bayes, Hoeffding Tree, Destek Vektör Makinası ve Lojistik Regresyon izlemektedir.

Çizelge 29. 500 Hastadan Oluşan Değişkenler Arası 0,5 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

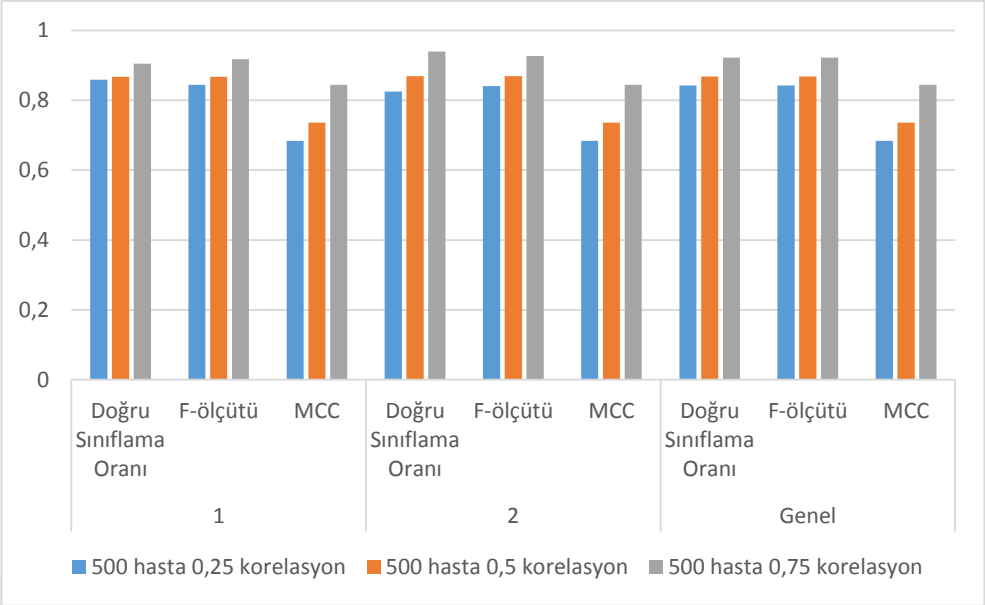
Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,715	0,719	0,724	0,444	0,785	0,801
	2	0,729	0,725	0,720	0,444	0,812	0,801
	Genel	0,722	0,722	0,722	0,444	0,799	0,801
Naive Bayes	**1**	0,727	0,731	0,736	0,468	0,788	0,803
	2	0,741	0,737	0,732	0,468	0,812	0,803
	Genel	0,734	0,734	0,734	0,468	0,800	0,803
Çok Katmanlı Algılayıcı	**1**	0,719	0,735	0,752	0,484	0,738	0,780
	2	0,765	0,749	0,733	0,484	0,768	0,780
	Genel	0,742	0,742	0,742	0,484	0,753	0,780
Hoeffding Tree	**1**	0,727	0,730	0,733	0,464	0,788	0,803
	2	0,737	0,734	0,731	0,464	0,812	0,803
	Genel	0,732	0,732	0,732	0,464	0,800	0,803
Destek Vektör Makinası	**1**	0,747	0,729	0,713	0,449	0,658	0,724
	2	0,701	0,718	0,736	0,449	0,666	0,724
	Genel	0,724	0,724	0,725	0,449	0,662	0,724
Hibrit Model	**1**	0,867	0,867	0,867	0,736	0,898	0,868
	2	0,869	0,869	0,869	0,736	0,902	0,868
	Genel	0,868	0,868	0,868	0,736	0,900	0,868

Çizelge 30 incelendiğinde; hem veri setinin dengeli ve yeterli büyüklükte olmasından hem de bağımsız değişkenler arası ilişkinin yeterli büyüklükte olmasından dolayı doğru sınıflama oranı, F-ölçütü ve MKK değerleri daha geçerli ölçütler olduğu için veri madenciliği algoritmalarının performanslarını değerlendirmek için seçilmiştir. Ancak bu ölçütlerin benzer sonuçlar vermesi durumunda diğer performans ölçütleri de değerlendirmeye dahil edilecektir. Tüm ölçütler birlikte değerlendirildiğinde en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Hoeffding Tree, Lojistik Regresyon, Çok Katmanlı Algılayıcı, AdaBoost ve Bagging izlemektedir.

Çizelge 30. 500 Hastadan Oluşan Değişkenler Arası 0,75 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,799	0,809	0,820	0,639	0,904	0,907
	2	0,839	0,830	0,820	0,639	0,917	0,907
	Genel	0,820	0,820	0,820	0,639	0,911	0,907
Çok Katmanlı Algılayıcı	**1**	0,812	0,807	0,802	0,628	0,888	0,896
	2	0,816	0,821	0,826	0,628	0,907	0,896
	Genel	0,814	0,814	0,814	0,628	0,898	0,896
AdaBoost	**1**	0,787	0,800	0,814	0,623	0,879	0,890
	2	0,835	0,823	0,810	0,623	0,892	0,890
	Genel	0,812	0,812	0,812	0,623	0,886	0,890
Bagging	**1**	0,770	0,786	0,803	0,599	0,866	0,879
	2	0,828	0,812	0,797	0,599	0,886	0,879
	Genel	0,800	0,800	0,800	0,599	0,876	0,879
Hoeffding Tree	**1**	0,816	0,823	0,830	0,663	0,909	0,912
	2	0,847	0,840	0,834	0,663	0,921	0,912
	Genel	0,832	0,832	0,832	0,663	0,915	0,912
Hibrit Model	**1**	0,904	0,917	0,931	0,844	0,926	0,921
	2	0,939	0,926	0,914	0,844	0,956	0,921
	Genel	0,922	0,922	0,922	0,844	0,940	0,921

Çizelge 28, 29 ve 30 birlikte değerlendirildiğinde ise hem kategori bazlı hem de genel performans ölçütlerinin değişkenler arası korelasyon arttıkça arttığı görülmektedir. 0,5-0,5 dağılıma sahip veri seti için hibrit modele ait performans ölçütlerinin karşılaştırılması ise Şekil 13'te gösterilmiştir. Değişkenler arası korelasyon arttıkça dengesiz sınıfa ait performans ölçütlerinin arttığı ve ayrıca tüm sınıflarda korelasyon ile birlikte MKK ölçütü değerlerinin de arttığı görülmektedir (Şekil 13).



Şekil 13. 500 Hastadan Oluşan Simüle Veri seti için Hibrit Modele ait Korelasyonlara Bağlı Performans Ölçütleri.

Çizelge 31 incelendiğinde; bağımsız değişkenler arası korelasyonun düşük olmasından dolayı performans ölçütlerinin daha düşük olduğu gözlenmektedir. Veri setinin dengeli olması nedeniyle doğru sınıflama oranı, F-ölçütü ve MKK değerleri daha geçerli ölçütler olduğu için veri madenciliği algoritmalarının performanslarını değerlendirmek için seçilmiştir. Ancak bu ölçütlerin benzer sonuçlar vermesi durumunda diğer performans ölçütleri de değerlendirmeye dahil edilecektir. Tüm ölçütler birlikte değerlendirildiğinde en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Hoeffding Tree, Lojistik Regresyon, Destek Vektör Makinası Çok Katmanlı Algılayıcı ve J48 izlemektedir.

Çizelge 31. 1000 Hastadan Oluşan Değişkenler Arası 0,25 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,630	0,629	0,628	0,248	0,665	0,667
	2	0,617	0,619	0,620	0,248	0,645	0,667
	Genel	0,624	0,624	0,624	0,248	0,655	0,667
Çok Katmanlı Algılayıcı	**1**	0,658	0,632	0,608	0,224	0,620	0,639
	2	0,565	0,590	0,617	0,224	0,619	0,639
	Genel	0,612	0,611	0,612	0,224	0,619	0,639
Destek Vektör Makinası	**1**	0,628	0,627	0,626	0,244	0,581	0,622
	2	0,615	0,617	0,618	0,244	0,570	0,622
	Genel	0,622	0,622	0,622	0,244	0,576	0,622
Hoeffding Tree	**1**	0,638	0,634	0,630	0,254	0,665	0,669
	2	0,615	0,620	0,624	0,254	0,648	0,669
	Genel	0,627	0,627	0,627	0,254	0,656	0,669
J48	**1**	0,676	0,634	0,597	0,211	0,592	0,616
	2	0,532	0,571	0,616	0,211	0,575	0,616
	Genel	0,605	0,603	0,606	0,211	0,584	0,616
Hibrit Model	**1**	0,913	0,858	0,809	0,699	0,897	0,846
	2	0,779	0,834	0,897	0,699	0,869	0,846
	Genel	0,847	0,846	0,853	0,699	0,883	0,846

Çizelge 32 incelendiğinde; hem veri setinin dengeli ve yeterli büyüklükte olmasından hem de bağımsız değişkenler arası ilişkinin yeterli büyüklükte olmasından dolayı doğru sınıflama oranı, F-ölçütü ve MKK değerleri daha geçerli ölçütler olduğu için veri madenciliği algoritmalarının performanslarını değerlendirmek için seçilmiştir. Ancak bu ölçütlerin benzer sonuçlar vermesi durumunda diğer performans ölçütleri de değerlendirmeye dahil edilecektir. Tüm ölçütler birlikte değerlendirildiğinde en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Naive Bayes, Hoeffding Tree, Lojistik Regresyon, Destek Vektör Makinası ve Çok Katmanlı Algılayıcı izlemektedir.

Çizelge 32. 1000 Hastadan Oluşan Değişkenler Arası 0,5 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

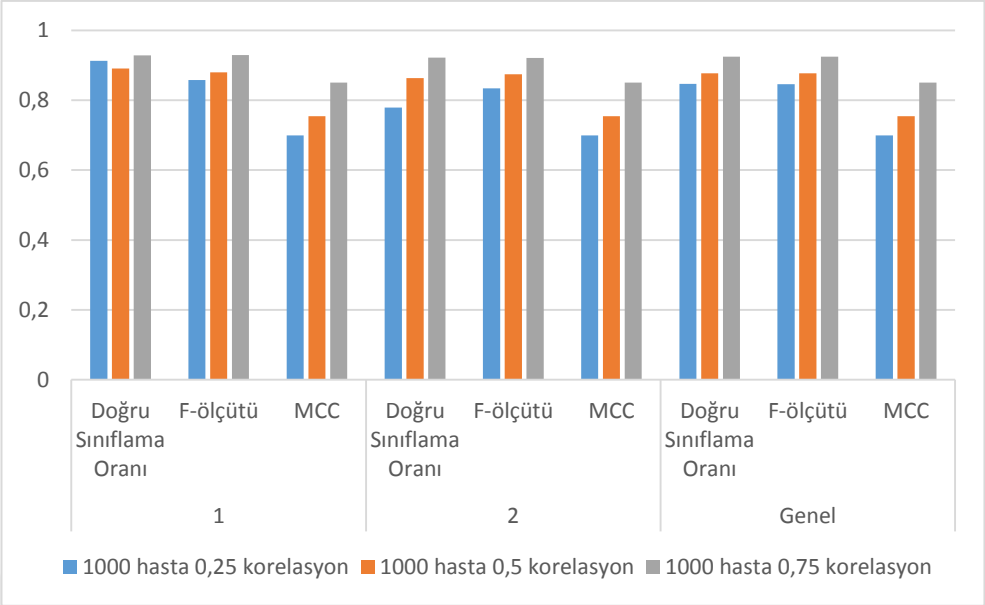
Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,738	0,733	0,728	0,458	0,802	0,799
	2	0,720	0,725	0,730	0,458	0,790	0,799
	Genel	0,729	0,729	0,729	0,458	0,796	0,799
Çok Katmanlı Algılayıcı	**1**	0,760	0,733	0,708	0,443	0,779	0,781
	2	0,681	0,708	0,736	0,443	0,759	0,781
	Genel	0,721	0,721	0,722	0,443	0,769	0,781
Destek Vektör Makinası	**1**	0,736	0,730	0,725	0,452	0,666	0,726
	2	0,716	0,722	0,727	0,452	0,662	0,726
	Genel	0,726	0,726	0,726	0,452	0,664	0,726
Naives Bayes	**1**	0,740	0,737	0,734	0,468	0,804	0,801
	2	0,728	0,731	0,734	0,468	0,789	0,801
	Genel	0,734	0,734	0,734	0,468	0,797	0,801
Hoeffding Tree	**1**	0,740	0,736	0,733	0,466	0,804	0,801
	2	0,726	0,729	0,733	0,466	0,790	0,801
	Genel	0,733	0,733	0,733	0,466	0,797	0,801
Hibrit Model	**1**	0,891	0,880	0,868	0,754	0,910	0,877
	2	0,863	0,874	0,886	0,754	0,904	0,877
	Genel	0,877	0,877	0,877	0,754	0,907	0,877

Çizelge 33 incelendiğinde; hem veri setinin dengeli ve yeterli büyüklükte olmasından hem de bağımsız değişkenler arası ilişkinin yeterli büyüklükte olmasından dolayı doğru sınıflama oranı, F-ölçütü ve MKK değerleri daha geçerli ölçütler olduğu için veri madenciliği algoritmalarının performanslarını değerlendirmek için seçilmiştir. Ancak bu ölçütlerin benzer sonuçlar vermesi durumunda diğer performans ölçütleri de değerlendirmeye dahil edilecektir. Tüm ölçütler birlikte değerlendirildiğinde en iyi performansa hibrit model ile ulaşılmakta, hibrit modeli sırasıyla Lojistik Regresyon, Random Forest, Destek Vektör Makinası, Hoeffding Tree ve Çok Katmanlı Algılayıcı izlemektedir.

Çizelge 33. 1000 Hastadan Oluşan Değişkenler Arası 0,75 Korelasyon İçeren Simüle Veri Setine ait Sonuçlar.

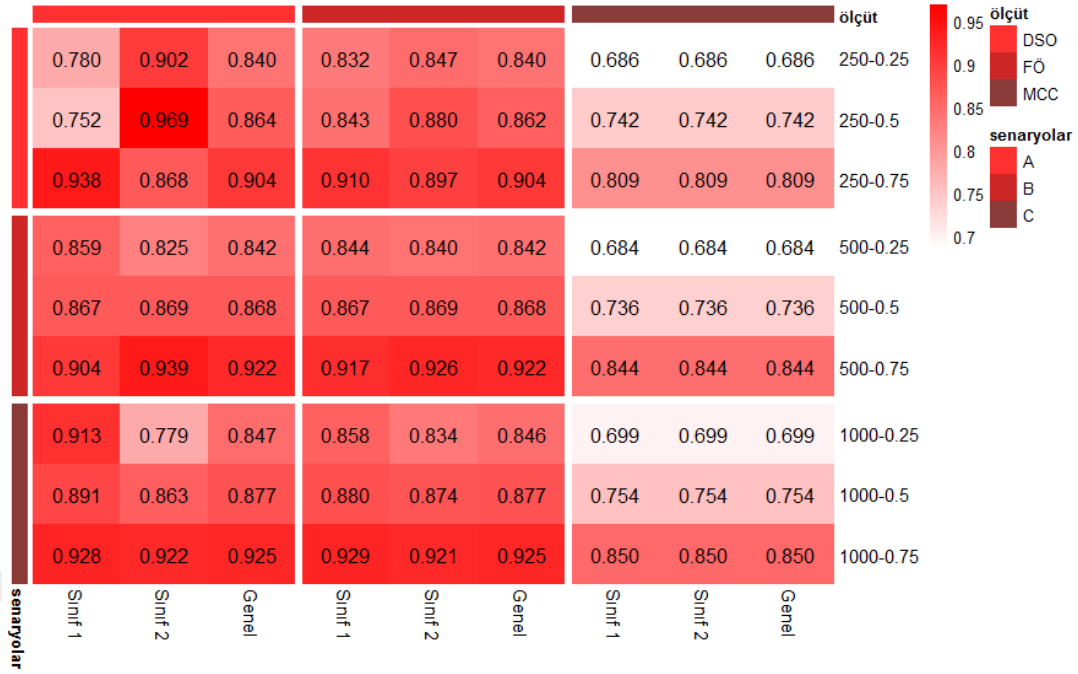
Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	**1**	0,833	0,822	0,811	0,619	0,908	0,897
	2	0,784	0,796	0,808	0,619	0,890	0,897
	Genel	0,810	0,810	0,810	0,619	0,900	0,897
Çok Katmanlı Algılayıcı	**1**	0,810	0,809	0,809	0,597	0,893	0,888
	2	0,786	0,787	0,788	0,597	0,872	0,888
	Genel	0,799	0,799	0,799	0,597	0,883	0,888
Destek Vektör Makinası	**1**	0,837	0,820	0,803	0,611	0,758	0,804
	2	0,772	0,790	0,809	0,611	0,733	0,804
	Genel	0,806	0,806	0,806	0,611	0,746	0,804
Hoeffding Tree	**1**	0,820	0,815	0,811	0,607	0,909	0,899
	2	0,786	0,791	0,797	0,607	0,893	0,899
	Genel	0,804	0,804	0,804	0,607	0,902	0,899
Random Forest	**1**	0,829	0,819	0,809	0,613	0,892	0,888
	2	0,782	0,793	0,804	0,613	0,879	0,888
	Genel	0,807	0,807	0,807	0,613	0,886	0,888
Hibrit Model	**1**	0,928	0,929	0,930	0,850	0,961	0,925
	2	0,922	0,921	0,920	0,850	0,933	0,925
	Genel	0,925	0,925	0,925	0,850	0,948	0,925

Çizelge 31, 32 ve 33 birlikte değerlendirildiğinde ise hem kategori bazlı hem de genel performans ölçütlerinin değişkenler arası korelasyon arttıkça arttığı görülmektedir. 0,5-0,5 dağılıma sahip veri seti için hibrit modele ait performans ölçütlerinin karşılaştırılması ise Şekil 14'te gösterilmiştir. Değişkenler arası korelasyon arttıkça dengesiz sınıfa ait performans ölçütlerinin arttığı ve ayrıca tüm sınıflarda korelasyon ile birlikte MKK ölçütü değerlerinin de arttığı görülmektedir (Şekil 14).



Şekil 14. 1000 Hastadan Oluşan Simüle Veri seti için Hibrit Modele ait Korelasyonlara Bağlı Performans Ölçütleri.

0,5-0,5 dağılıma göre üretilen 9 ayrı senaryoya ait performans ölçütlerinin ısı haritası üzerinde gösterimi ise Şekil 15'te verilmiştir.



Şekil 15. 0,7-0,3 Dağılım için Veri Setlerine ait Performans Ölçütlerinin Isı Haritası ile Gösterimi. DSO:Doğru Sınıflama Oranı, FÖ:F-ölçütü, A:250 hastadan oluşan veri setleri, B: 500 hastadan oluşan veri setleri, C:1000 hastadan oluşan veri setleri, 250-0.25:250 hastadan oluşan 0,25 korelasyona sahip veri seti, 250-0.5:250 hastadan oluşan 0,5 korelasyona sahip veri seti, 250-0.75:250 hastadan oluşan 0,75 korelasyona sahip veri seti, 500-0.25:500 hastadan oluşan 0,25 korelasyona sahip veri seti, 500-0.5:500 hastadan oluşan 0,5 korelasyona sahip veri seti, 500-0.75:500 hastadan oluşan 0,75 korelasyona sahip veri seti, 1000-0.25:1000 hastadan oluşan 0,25 korelasyona sahip veri seti, 1000-0.5:1000 hastadan oluşan 0,5 korelasyona sahip veri seti, 1000-0.75:1000 hastadan oluşan 0,75 korelasyona sahip veri seti.

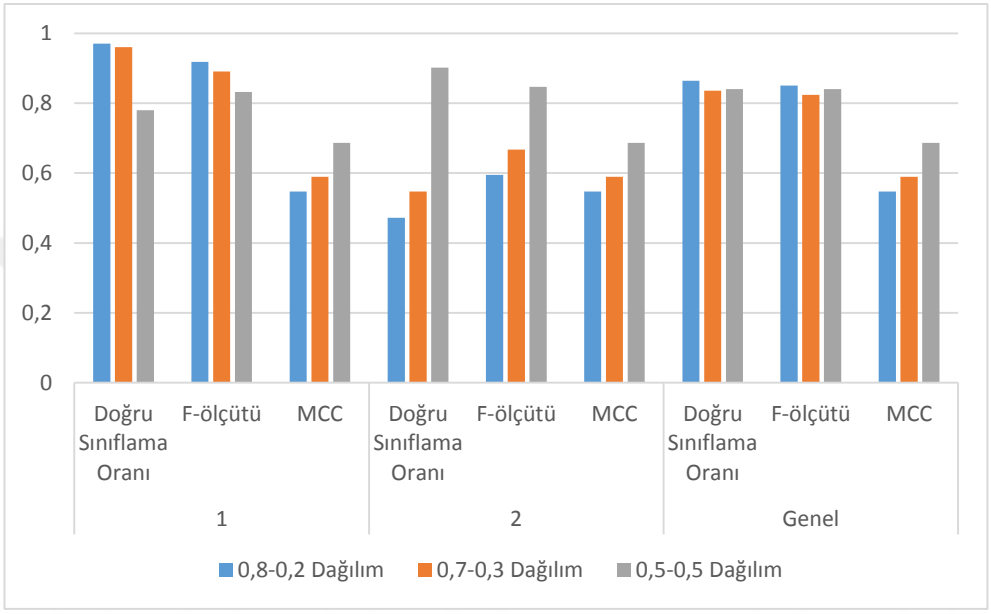
3.4. Simüle Veri Setlerine ait Genel Performans Ölçütlerinin Değerlendirilmesi

250, 500 ve 700 örnekleme 0,8-0,2, 0,7-0,3 ve 0,5-0,5 dağılıma sahip simüle veri setleri için hibrit modele ait performans değerlendirmeleri aşağıda verilmiştir.

3.4.1. 250 Hastadan Oluşan Simüle Veri Setleri için Performans Ölçütlerinin Değerlendirilmesi

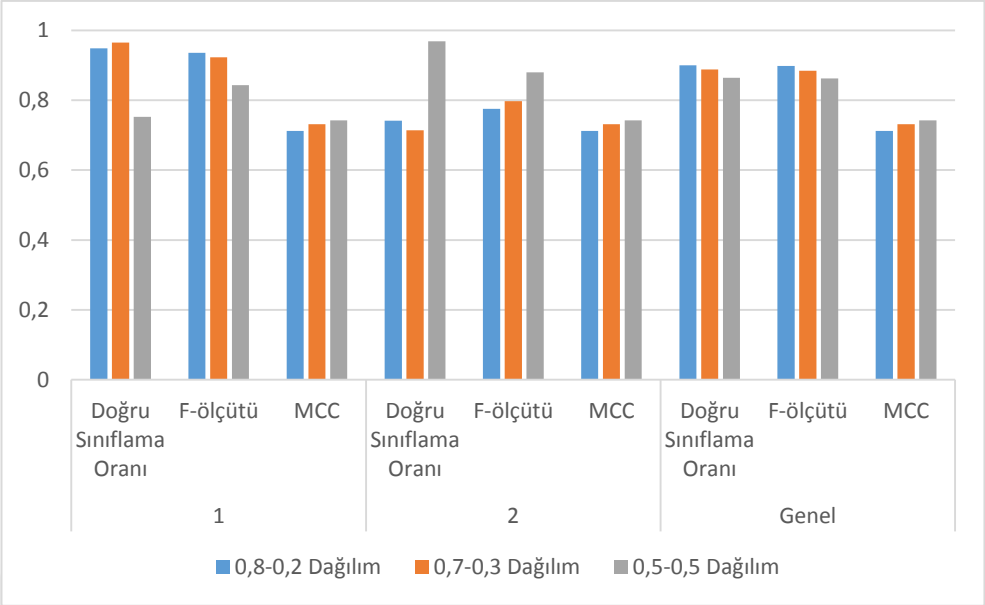
Şekil 16'da 250 hastadan oluşan bağımsız değişkenler arası 0,25 korelasyona ve 0,8-0,2, 0,7-0,3 ve 0,5-0,5 dağılımlara sahip veri setleri için hibrit model

performansları gösterilmiştir. Veri setleri dengesiz dağılıma sahip olduğu için doğru sınıflama oranı ve F-ölçütü 0,8-0,2 ve 0,7-0,3 dağılıma sahip veri setlerinde daha yüksek bulunmuştur. Ancak dengesiz dağılımlarda bu ölçütler yanıltıcı olduğu için MKK değerinin kullanılması önerilmektedir. MKK değerlerine bakıldığında ise 0,5-0,5 dağılıma sahip veri setinin daha iyi performansa sahip olduğu görülmektedir.



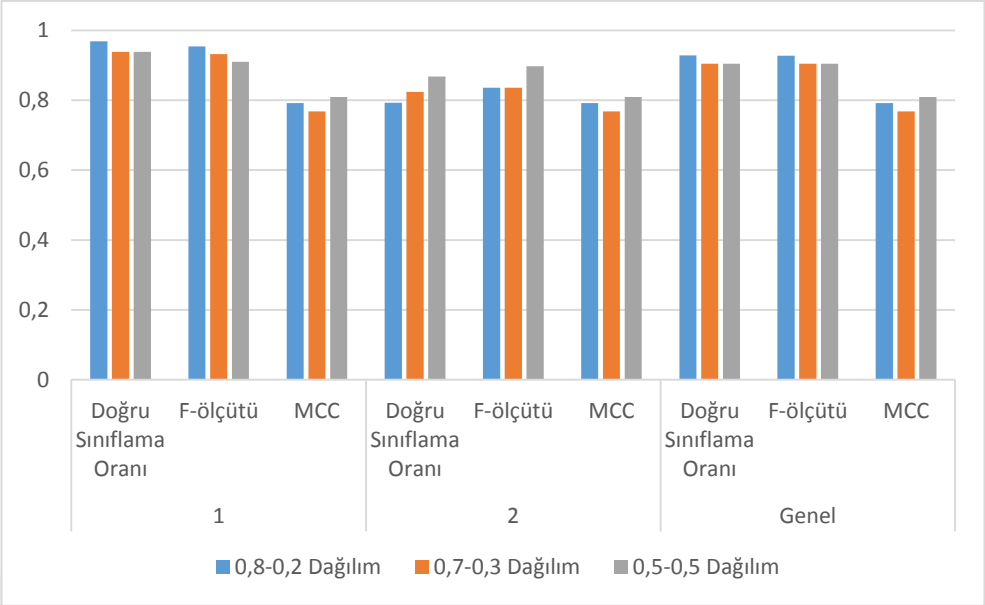
Şekil 16. 250 Hastadan Oluşan Bağımsız Değişkenler Arası 0,25 Korelasyona Sahip Tüm Veri Seti Dağılımlarına ait Performans Ölçütleri.

Şekil 17'de 250 hastadan oluşan bağımsız değişkenler arası 0,5 korelasyona ve 0,8-0,2, 0,7-0,3 ve 0,5-0,5 dağılımlara sahip veri setleri için hibrit model performansları gösterilmiştir. Veri setleri dengesiz dağılıma sahip olduğu için doğru sınıflama oranı ve F-ölçütü 0,8-0,2 ve 0,7-0,3 dağılıma sahip veri setlerinde daha yüksek bulunmuştur. Ancak dengesiz dağılımlarda bu ölçütler yanıltıcı olduğu için MKK değerinin kullanılması önerilmektedir. MKK değerlerine bakıldığında ise 0,5-0,5 dağılıma sahip veri setinin daha iyi performansa sahip olduğu görülmektedir.



Şekil 17. 250 Hastadan Oluşan Bağımsız Değişkenler Arası 0,5 Korelasyona Sahip Tüm Veri Seti Dağılımlarına ait Performans Ölçütleri.

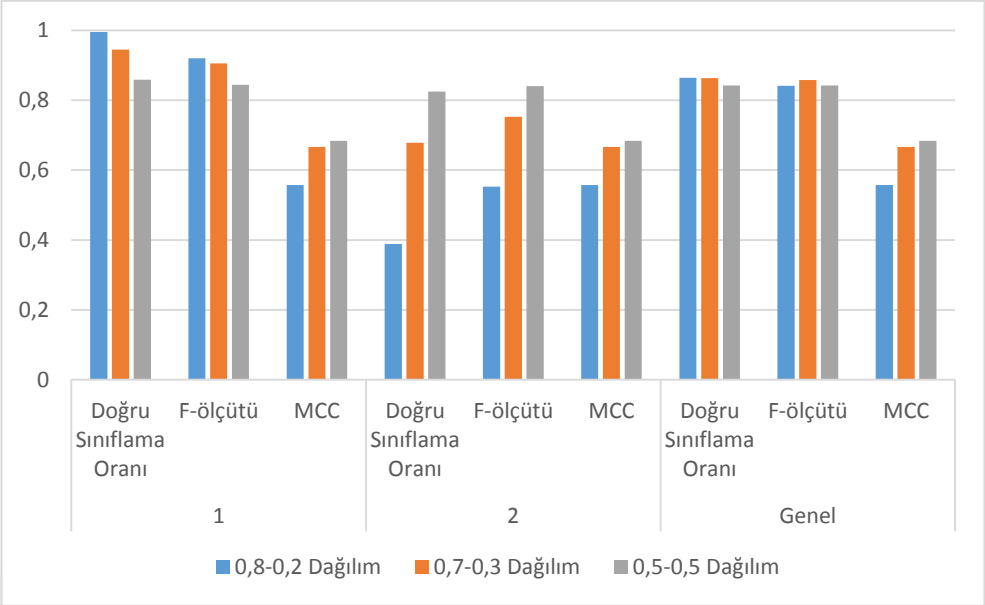
Şekil 18'de 250 hastadan oluşan bağımsız değişkenler arası 0,75 korelasyona ve 0,8-0,2, 0,7-0,3 ve 0,5-0,5 dağılımlara sahip veri setleri için hibrit model performansları gösterilmiştir. Veri setleri dengesiz dağılıma sahip olduğu için doğru sınıflama oranı ve F-ölçütü 0,8-0,2 ve 0,7-0,3 dağılıma sahip veri setlerinde daha yüksek bulunmuştur. Ancak dengesiz dağılımlarda bu ölçütler yanıltıcı olduğu için MKK değerinin kullanılması önerilmektedir. MKK değerlerine bakıldığında ise 0,5-0,5 dağılıma sahip veri setinin daha iyi performansa sahip olduğu görülmektedir.



Şekil 18. 250 Hastadan Oluşan Bağımsız Değişkenler Arası 0,75 Korelasyona Sahip Tüm Veri Seti Dağılımlarına ait Performans Ölçütleri.

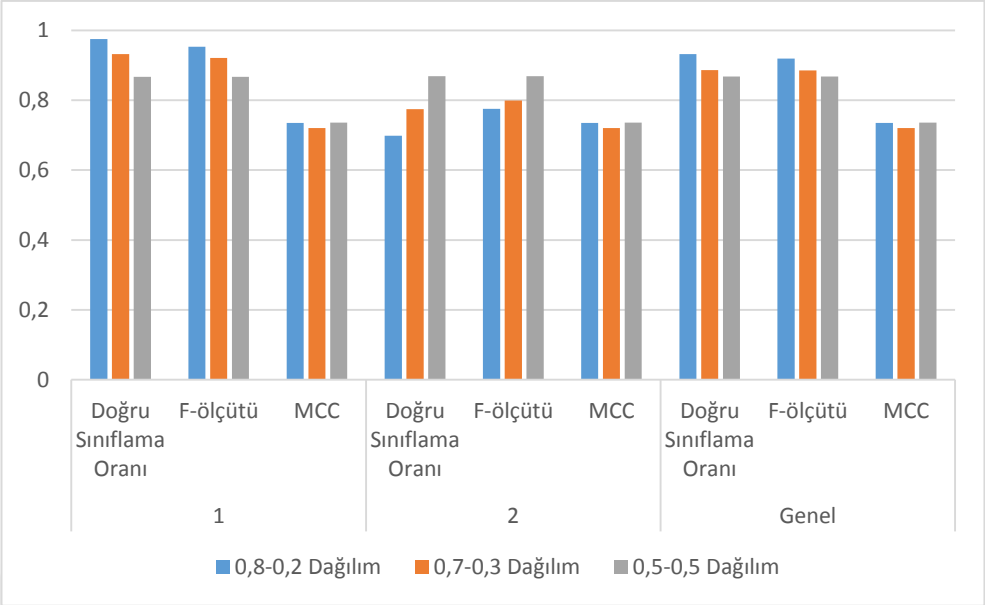
3.4.2. 500 Hastadan Oluşan Simüle Veri Setleri için Performans Ölçütlerinin Değerlendirilmesi

Şekil 19'da 500 hastadan oluşan bağımsız değişkenler arası 0,25 korelasyona ve 0,8-0,2, 0,7-0,3 ve 0,5-0,5 dağılımlara sahip veri setleri için hibrit model performansları gösterilmiştir. Veri setleri dengesiz dağılıma sahip olduğu için doğru sınıflama oranı ve F-ölçütü 0,8-0,2 ve 0,7-0,3 dağılıma sahip veri setlerinde daha yüksek bulunmuştur. Ancak dengesiz dağılımlarda bu ölçütler yanıltıcı olduğu için MKK değerinin kullanılması önerilmektedir. MKK değerlerine bakıldığında ise 0,5-0,5 dağılıma sahip veri setinin daha iyi performansa sahip olduğu görülmektedir.



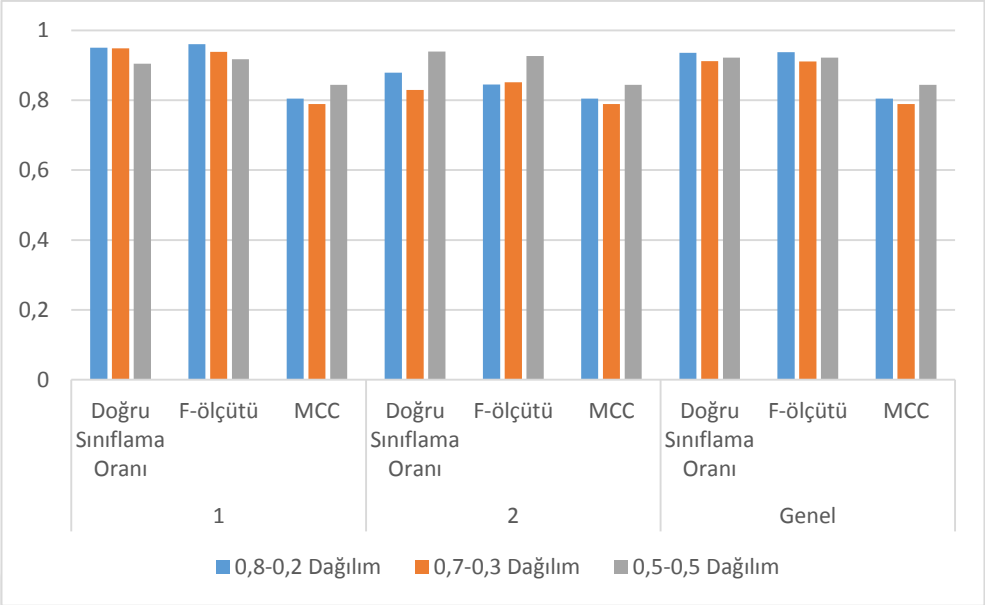
Şekil 19. 500 Hastadan Oluşan Bağımsız Değişkenler Arası 0,25 Korelasyona Sahip Tüm Veri Seti Dağılımlarına ait Performans Ölçütleri.

Şekil 20'de 500 hastadan oluşan bağımsız değişkenler arası 0,5 korelasyona ve 0,8-0,2, 0,7-0,3 ve 0,5-0,5 dağılımlara sahip veri setleri için hibrit model performansları gösterilmiştir. Veri setleri dengesiz dağılıma sahip olduğu için doğru sınıflama oranı ve F-ölçütü 0,8-0,2 ve 0,7-0,3 dağılıma sahip veri setlerinde daha yüksek bulunmuştur. Ancak dengesiz dağılımlarda bu ölçütler yanıltıcı olduğu için MKK değerinin kullanılması önerilmektedir. MKK değerlerine bakıldığında ise 0,5-0,5 dağılıma sahip veri setinin daha iyi performansa sahip olduğu görülmektedir.



Şekil 20. 500 Hastadan Oluşan Bağımsız Değişkenler Arası 0,5 Korelasyona Sahip Tüm Veri Seti Dağılımlarına ait Performans Ölçütleri.

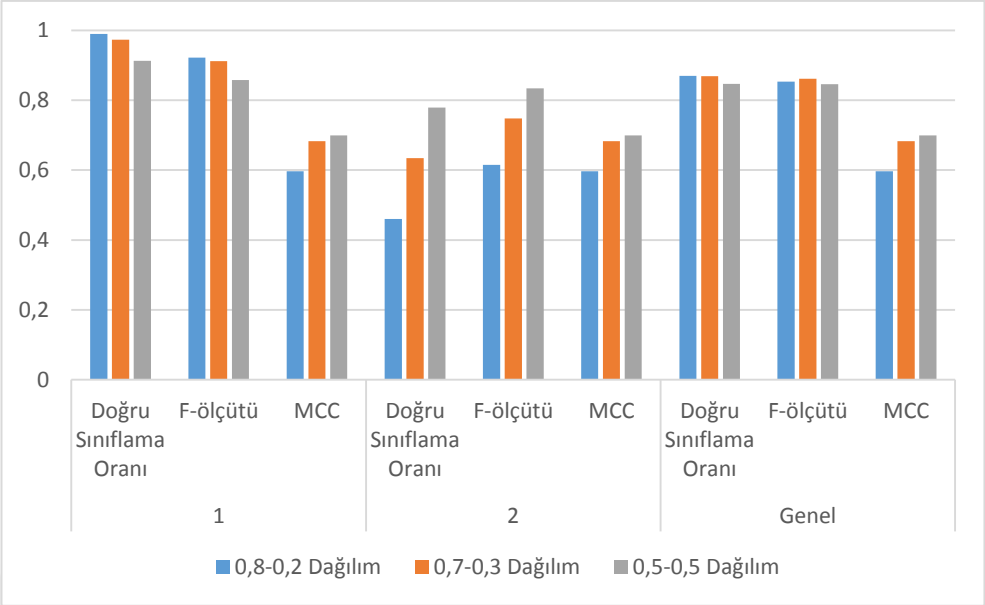
Şekil 21'de 500 hastadan oluşan bağımsız değişkenler arası 0,75 korelasyona ve 0,8-0,2, 0,7-0,3 ve 0,5-0,5 dağılımlara sahip veri setleri için hibrit model performansları gösterilmiştir. Veri setleri dengesiz dağılıma sahip olduğu için doğru sınıflama oranı ve F-ölçütü 0,8-0,2 ve 0,7-0,3 dağılıma sahip veri setlerinde daha yüksek bulunmuştur. Ancak dengesiz dağılımlarda bu ölçütler yanıltıcı olduğu için MKK değerinin kullanılması önerilmektedir. MKK değerlerine bakıldığında ise 0,5-0,5 dağılıma sahip veri setinin daha iyi performansa sahip olduğu görülmektedir.



Şekil 21. 500 Hastadan Oluşan Bağımsız Değişkenler Arası 0,75 Korelasyona Sahip Tüm Veri Seti Dağılımlarına ait Performans Ölçütleri.

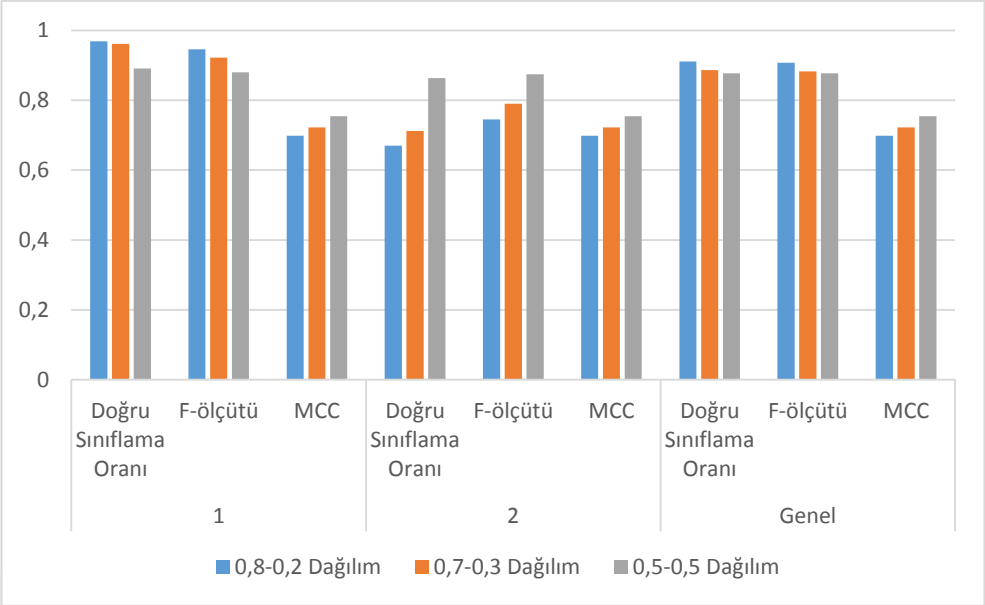
3.4.3. 1000 Hastadan Oluşan Simüle Veri Setleri için Performans Ölçütlerinin Değerlendirilmesi

Şekil 22'de 1000 hastadan oluşan bağımsız değişkenler arası 0,25 korelasyona ve 0,8-0,2, 0,7-0,3 ve 0,5-0,5 dağılımlara sahip veri setleri için hibrit model performansları gösterilmiştir. Veri setleri dengesiz dağılıma sahip olduğu için doğru sınıflama oranı ve F-ölçütü 0,8-0,2 ve 0,7-0,3 dağılıma sahip veri setlerinde daha yüksek bulunmuştur. Ancak dengesiz dağılımlarda bu ölçütler yanıltıcı olduğu için MKK değerinin kullanılması önerilmektedir. MKK değerlerine bakıldığında ise 0,5-0,5 dağılıma sahip veri setinin daha iyi performansa sahip olduğu görülmektedir.



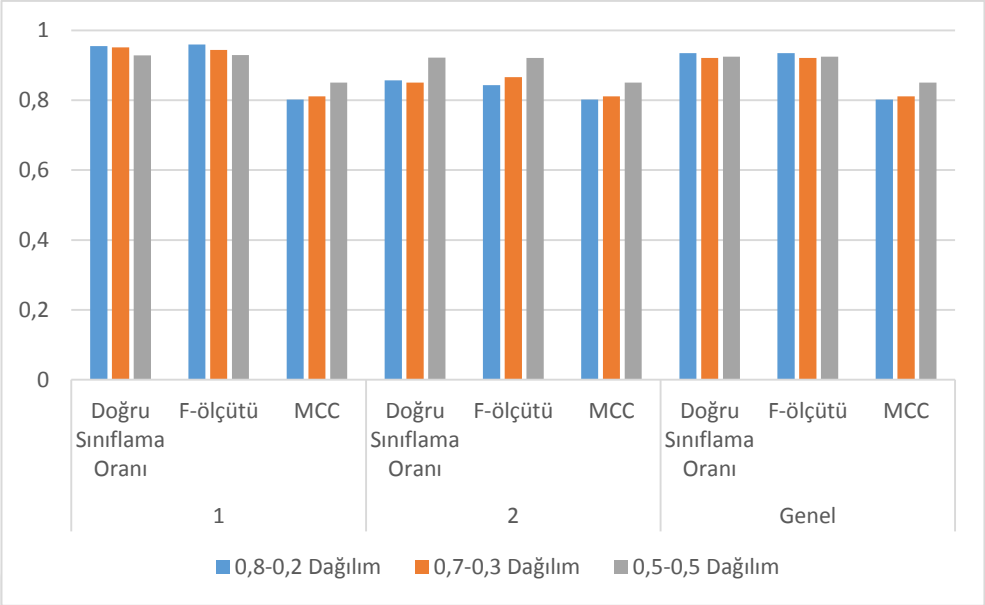
Şekil 22. 1000 Hastadan Oluşan Bağımsız Değişkenler Arası 0,25 Korelasyona Sahip Tüm Veri Seti Dağılımlarına ait Performans Ölçütleri.

Şekil 23'te 1000 hastadan oluşan bağımsız değişkenler arası 0,5 korelasyona ve 0,8-0,2, 0,7-0,3 ve 0,5-0,5 dağılımlara sahip veri setleri için hibrit model performansları gösterilmiştir. Veri setleri dengesiz dağılıma sahip olduğu için doğru sınıflama oranı ve F-ölçütü 0,8-0,2 ve 0,7-0,3 dağılıma sahip veri setlerinde daha yüksek bulunmuştur. Ancak dengesiz dağılımlarda bu ölçütler yanıltıcı olduğu için MKK değerinin kullanılması önerilmektedir. MKK değerlerine bakıldığında ise 0,5-0,5 dağılıma sahip veri setinin daha iyi performansa sahip olduğu görülmektedir.



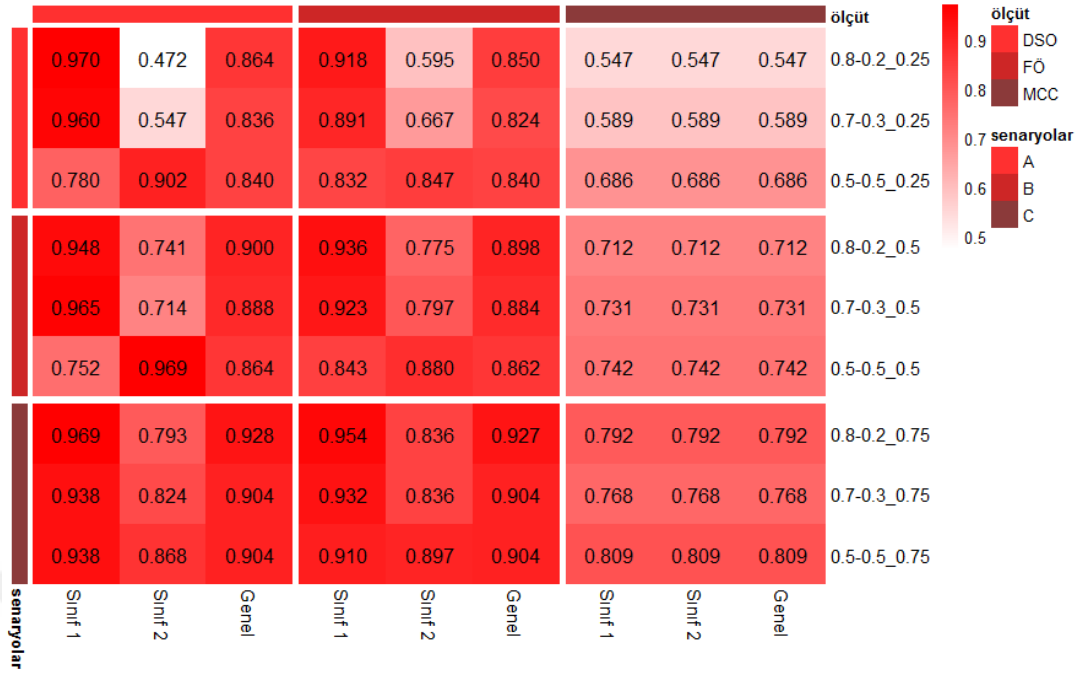
Şekil 23. 1000 Hastadan Oluşan Bağımsız Değişkenler Arası 0,5 Korelasyona Sahip Tüm Veri Seti Dağılımlarına ait Performans Ölçütleri.

Şekil 24'te 1000 hastadan oluşan bağımsız değişkenler arası 0,75 korelasyona ve 0,8-0,2, 0,7-0,3 ve 0,5-0,5 dağılımlara sahip veri setleri için hibrit model performansları gösterilmiştir. Veri setleri dengesiz dağılıma sahip olduğu için doğru sınıflama oranı ve F-ölçütü 0,8-0,2 ve 0,7-0,3 dağılıma sahip veri setlerinde daha yüksek bulunmuştur. Ancak dengesiz dağılımlarda bu ölçütler yanıltıcı olduğu için MKK değerinin kullanılması önerilmektedir. MKK değerlerine bakıldığında ise 0,5-0,5 dağılıma sahip veri setinin daha iyi performansa sahip olduğu görülmektedir.

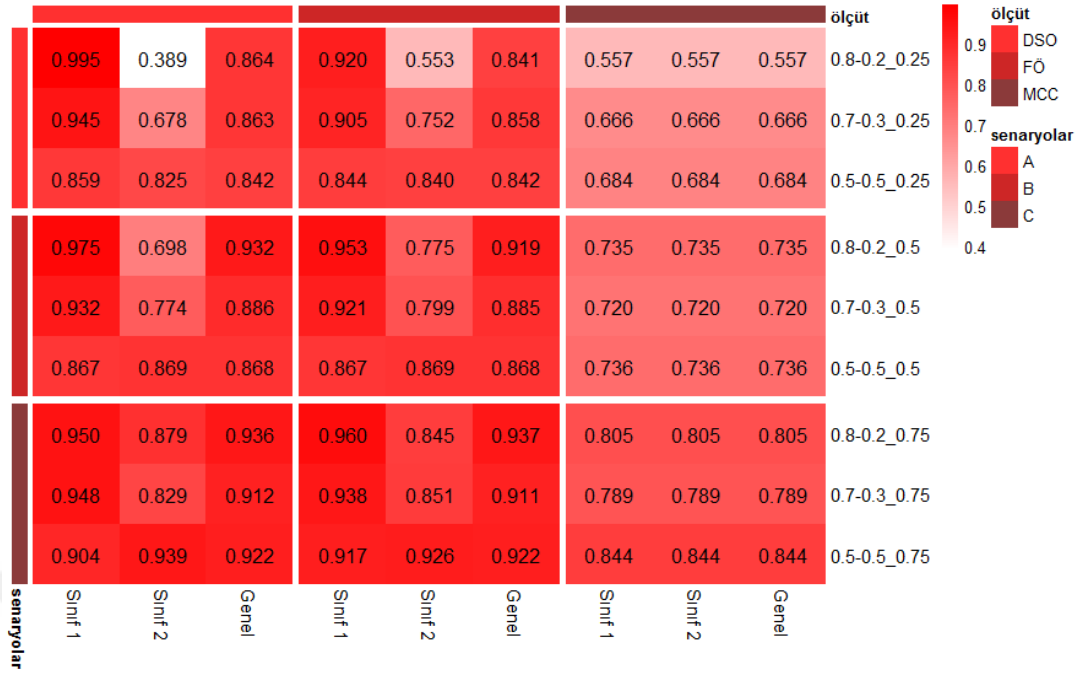


Şekil 24. 1000 Hastadan Oluşan Bağımsız Değişkenler Arası 0,75 Korelasyona Sahip Tüm Veri Seti Dağılımlarına ait Performans Ölçütleri.

Şekil 25, Şekil 26 ve Şekil 27'de ise sırasıyla 250 hastadan oluşan, 500 hastadan oluşan ve 1000 hastadan oluşan tüm senaryoları içeren veri setlerine ait performans ölçütleri ısı haritası ile gösterilmiştir.



Şekil 25. 250 Hastadan Oluşan Tüm Veri Setlerine ait Performans Ölçütlerinin Isı Haritası ile Gösterimi. DSO:Doğru Sınıflama Oranı, FÖ:F-ölçütü, A:0,25 korelasyona sahip veri setleri, B: 0,5 korelasyona sahip veri setleri, C:0,75 korelasyona sahip veri setleri, 0.8-0.2-0.25:0,8-0,2 dağılıma sahip 0,25 korelasyon içeren veri seti, 0.7-0.3-0.25:0,7-0,3 dağılıma sahip 0,25 korelasyon içeren veri seti, 0.5-0.5-0.25:0,5-0,5 dağılıma sahip 0,25 korelasyon içeren veri seti, 0.8-0.2-0.5:0,8-0,2 dağılıma sahip 0,5 korelasyon içeren veri seti, 0.7-0.3-0.5:0,7-0,3 dağılıma sahip 0,5 korelasyon içeren veri seti, 0.5-0.5-0.5:0,5-0,5 dağılıma sahip 0,5 korelasyon içeren veri seti, 0.8-0.2-0.75:0,8-0,2 dağılıma sahip 0,75 korelasyon içeren veri seti, 0.7-0.3-0.75:0,7-0,3 dağılıma sahip 0,75 korelasyon içeren veri seti, 0.5-0.5-0.75:0,5-0,5 dağılıma sahip 0,75 korelasyon içeren veri seti.



Şekil 26. 500 Hastadan Oluşan Tüm Veri Setlerine ait Performans Ölçütlerinin Isı Haritası ile Gösterimi. DSO:Doğru Sınıflama Oranı, FÖ:F-ölçütü, A:0,25 korelasyona sahip veri setleri, B: 0,5 korelasyona sahip veri setleri, C:0,75 korelasyona sahip veri setleri, 0.8-0.2-0.25:0,8-0,2 dağılıma sahip 0,25 korelasyon içeren veri seti, 0.7-0.3-0.25:0,7-0,3 dağılıma sahip 0,25 korelasyon içeren veri seti, 0.5-0.5-0.25:0,5-0,5 dağılıma sahip 0,25 korelasyon içeren veri seti, 0.8-0.2-0.5:0,8-0,2 dağılıma sahip 0,5 korelasyon içeren veri seti, 0.7-0.3-0.5:0,7-0,3 dağılıma sahip 0,5 korelasyon içeren veri seti, 0.5-0.5-0.5:0,5-0,5 dağılıma sahip 0,5 korelasyon içeren veri seti, 0.8-0.2-0.75:0,8-0,2 dağılıma sahip 0,75 korelasyon içeren veri seti, 0.7-0.3-0.75:0,7-0,3 dağılıma sahip 0,75 korelasyon içeren veri seti, 0.5-0.5-0.75:0,5-0,5 dağılıma sahip 0,75 korelasyon içeren veri seti.

Şekil 27. 1000 Hastadan Oluşan Tüm Veri Setlerine ait Performans Ölçütlerinin Isı Haritası ile Gösterimi. DSO:Doğru Sınıflama Oranı, FÖ:F-ölçütü, A:0,25 korelasyona sahip veri setleri, B: 0,5 korelasyona sahip veri setleri, C:0,75 korelasyona sahip veri setleri, 0.8-0.2-0.25:0,8-0,2 dağılıma sahip 0,25 korelasyon içeren veri seti, 0.7-0.3-0.25:0,7-0,3 dağılıma sahip 0,25 korelasyon içeren veri seti, 0.5-0.5-0.25:0,5-0,5 dağılıma sahip 0,25 korelasyon içeren veri seti, 0.8-0.2-0.5:0,8-0,2 dağılıma sahip 0,5 korelasyon içeren veri seti, 0.7-0.3-0.5:0,7-0,3 dağılıma sahip 0,5 korelasyon içeren veri seti, 0.5-0.5-0.5:0,5-0,5 dağılıma sahip 0,5 korelasyon içeren veri seti, 0.8-0.2-0.75:0,8-0,2 dağılıma sahip 0,75 korelasyon içeren veri seti, 0.7-0.3-0.75:0,7-0,3 dağılıma sahip 0,75 korelasyon içeren veri seti, 0.5-0.5-0.75:0,5-0,5 dağılıma sahip 0,75 korelasyon içeren veri seti.

3.5. Gerçek Veri Setlerine Ait Bulgular

Çalışmada gerçek veri seti olarak hepatit ve meme kanseri veri setleri kullanılmıştır. Bu veri setlerine ait sonuçlar ise aşağıda değerlendirilmiştir.

3.5.1. Hepatit Veri Seti için Bulgular

Hepatit veri seti için sonuç değişkeni bazlı tanımlayıcılar ve bu değişkene göre istatistiksel karşılaştırmalar Çizelge 34 ve Çizelge 35'te verilmiştir.

Çizelge 34. Nicel Değişkenler için Tanımlayıcılar ve İstatistiksel Karşılaştırmalar.

Değişkenler	Sınıf				p değeri
	Yaşıyor		Ölü		
	Ort.±SS	Ortanca (Min.-Maks.)	Ort.±SS	Ortanca (Min.-Maks.)	
Yaş, (n=155)	39,80±12,83	38,00 (7,00-78,00)	46,59±9,95	46,50 (30,00-70,00)	0,002[a]
Bilirubin, (n=149)	1,15±0,72	1,00 (0,30-4,60)	2,54±1,94	1,95 (0,40-8,00)	<0,001[a]
Alk Fosfat, (n=126)	101,31±50,26	85,00 (26,00-295,00)	122,38±54,35	113,50 (62,00-280,00)	0,034[a]
Sgot, (n=151)	82,44±86,51	55,00 (14,00-648,00)	99,83±101,77	66,00 (16,00-528,00)	0,222[a]
Albumin, (n=139)	3,98±0,56	4,00 (2,70-6,40)	3,15±0,60	3,30 (2,10-4,20)	<0,001[a]
Protime, (n=88)	66,57±21,91	66,00 (0,00-100,00)	43,50±16,76	39,00 (29,00-90,00)	<0,001[a]

a:Mann-Whitney U testi, Ort.:Ortalama, SS:Standart Sapma, Min.:Minimum, Maks.: Maksimum

Çizelge 34'te nicel değişkenler için sınıf değişkeni kategorilerilerine (Yaşıyor, Ölü) göre istatistiksel olarak anlamlı farklılıklar ve tanımlayıcı istatistikler verilmiştir. Nicel değişkenlerden yaş, bilirubin, Alk fosfat, Albumin ve Protime için farklar istatistiksel olarak anlamlı bulunmuştur (sırasıyla p=0,002, p<0,001, p=0,034, p<0,001 ve p<0,001).

Çizelge 35. Nitel Değişkenler için Tanımlayıcılar ve İstatistiksel Karşılaştırmalar.

Değişkenler		Sınıf				p değeri
		Yaşıyor		Ölü		
		Sayı	%	Sayı	%	
Cinsiyet	Erkek	107	87,0	32	100,0	0,044[b]
	Kadın	16	13,0	0	0,0	
Steroid	Yok	56	45,9	20	62,5	0,095[a]
	Var	66	54,1	12	37,5	
Antiviraller	Yok	22	17,9	2	6,2	0,168[b]
	Var	101	82,1	30	93,8	
Tükenmişlik	Yok	70	57,4	30	93,8	<0,001[a]
	Var	52	42,6	2	6,2	
Kırgınlık	Yok	38	31,1	23	71,9	<0,001[a]
	Var	84	68,9	9	28,1	
İştahsızlık	Yok	22	18,0	10	31,2	0,101[a]
	Var	100	82,0	22	68,8	
Karaciğerde Büyüme	Yok	22	18,6	3	11,1	0,572[b]
	Var	96	81,4	24	88,9	
Karaciğerde Sertlik	Yok	47	40,2	13	48,1	0,449[a]
	Var	70	59,8	14	51,9	
Spleen Palpable	Yok	18	15,1	12	38,7	0,003[a]
	Var	101	84,9	19	61,3	
Spiders	Yok	29	24,4	22	71,0	<0,001[a]
	Var	90	75,6	9	29,0	
Karın İltihabı	Yok	6	5,0	14	45,2	<0,001[b]
	Var	113	95,0	17	54,8	
Varis	Yok	7	5,9	11	35,5	<0,001[b]
	Var	112	94,1	20	64,5	
Histoloji	Yok	78	63,4	7	21,9	<0,001[a]
	Var	45	36,6	25	78,1	

a:Ki-kare testi, b:Fisher-exact testi

Çizelge 35'te nitel değişkenler ile Class değişkeni kategorilerileri (Yaşıyor, Ölü) arasında istatistiksel olarak anlamlı ilişkilere bakılmış ve tanımlayıcı istatistikler verilmiştir. Nitel değişkenlerden Cinsiyet, Tükenmişlik, Kırgınlık, Spleen Palpable, Spiders, Karın İltihabı, Varis ve Histoloji ile Sınıf değişkeni arasında istatistiksel olarak anlamlı ilişki bulunmuştur (sırasıyla p=0,044, p<0,001, p<0,001, p=0,003, p<0,001, p<0,001, p<0,001 ve p<0,001).

Hepatit gerçek veri seti için Chi-squared Attribute Eval., Gain Ratio Attribute Eval ve Info Gain Attribute Eval yöntemleri kullanılarak değişken önemine bakılmıştır. Klinik ve istatistiksel olarak önemsiz olduğu düşünülen Alk Fosfat ve

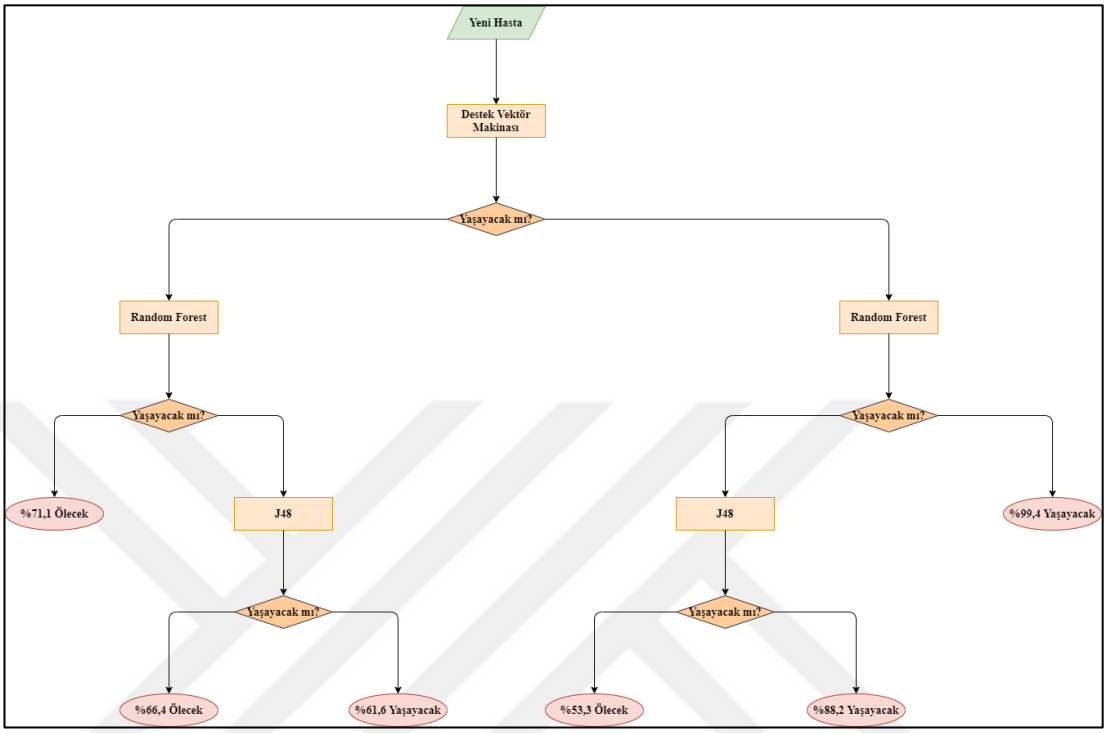
Sgot değişkenleri veri setinden çıkarılmıştır. Veri setinde analiz için 17 bağımsız ve bir bağımlı değişken kalmıştır. Bu değişkenler sırasıyla Albumin, Karın İltihabı, Bilirubin, Spiders, Varis, Histoloji, Kırgınlık, Protime, Tükenmişlik, Spleen Palpable, Yaş, Cinsiyet, Steroid, İştahsızlık, Antiviraller, Karaciğerde Büyüme, Karaciğerde Sertlik ve Sınıf'tır.

Çizelge 36. Hepatit Gerçek Veri Seti için Veri Madenciliği Yöntemlerine ait Performans Ölçütleri.

Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Çok Katmanlı Algılayıcı	Yaşıyor	0,894	0,891	0,887	0,462	0,958	0,855
	Ölü	0,563	0,571	0,581	0,462	0,600	0,855
	Genel	0,826	0,825	0,824	0,462	0,884	0,855
Destek Vektör Makinesi	Yaşıyor	0,919	0,908	0,897	0,532	0,888	0,756
	Ölü	0,594	0,623	0,655	0,532	0,473	0,756
	Genel	0,852	0,849	0,847	0,532	0,803	0,756
Lojistik Regresyon	Yaşıyor	0,919	0,893	0,869	0,426	0,898	0,802
	Ölü	0,469	0,526	0,600	0,426	0,562	0,802
	Genel	0,826	0,818	0,814	0,426	0,829	0,802
J48	Yaşıyor	0,943	0,903	0,866	0,450	0,856	0,708
	Ölü	0,438	0,528	0,667	0,450	0,585	0,708
	Genel	0,839	0,825	0,825	0,450	0,800	0,708
Random Forest	Yaşıyor	0,951	0,914	0,880	0,523	0,961	0,869
	Ölü	0,500	0,593	0,727	0,523	0,654	0,869
	Genel	0,858	0,848	0,848	0,523	0,897	0,869
Hibrit Model	Yaşıyor	0,976	0,949	0,923	0,730	0,991	0,832
	Ölü	0,688	0,772	0,880	0,730	0,836	0,832
	Genel	0,916	0,912	0,914	0,730	0,959	0,832

Çizelge 36'daki veri madenciliği yöntemlerinin performanslarını karşılaştırmak için literatürle uyumlu olması açısından en önemli karar verici performans ölçütü olarak MKK belirlenmiş ve MKK değeri benzer (yakın) olan yöntemler için bu ölçüte ek olarak doğru sınıflama oranı ve F-ölçütü değerlerine bakılması uygun görülmüştür. Bu ölçütlere bakıldığında en iyi performansa Hibrit Model ile ulaşıldığı görülmektedir. Hibrit Model'i sırasıyla Destek Vektör Makinesi, Random Forest, J48, Çok Katmanlı Algılayıcı ve Lojistik Regresyon yöntemleri izlemektedir.

Gerçek hepatit veri seti için hibrit model şekli, gelen yeni hastanın ölecek ya da yaşayacak olma tahmini olasılığı ile birlikte Şekil 28'de verilmiştir.



Şekil 28. Hepatit Gerçek Veri Seti için Oluşturulan Hibrit Model Yapısı.

Hepatit 250 hastadan oluşan simüle veri seti için Chi-squared Attribute Eval., Gain Ratio Attribute Eval ve Info Gain Attribute Eval yöntemleri kullanılarak değişken önemine bakılmıştır. Klinik ve istatistiksel olarak önemsiz olduğu düşünülen Sgot değişkeni veri setinden çıkarılmıştır. Veri setinde analiz için 18 bağımsız ve bir bağımlı değişken kalmıştır. Bu değişkenler Albumin, Karın İltihabı, Bilirubin, Protime, Spiders, Histoloji, Kırgınlık, Varis, Tükenmişlik, Yaş, Spleen Palpable, Alk fosfat, Cinsiyet, Antiviraller, Steroid, İştahsızlık, Karaciğerde Büyüme, Karaciğerde Sertlik ve Sınıf'tır.

Çizelge 37. Hepatit 250 Hastadan Oluşan Simüle Veri Seti için Veri Madenciliği Yöntemlerine ait Performans Ölçütleri.

Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Çok Katmanlı Algılayıcı	Yaşıyor	0,934	0,927	0,920	0,641	0,970	0,907
	Ölü	0,692	0,713	0,735	0,641	0,712	0,907
	Genel	0,884	0,883	0,882	0,641	0,916	0,907
Destek Vektör Makinesi	Yaşıyor	0,949	0,938	0,926	0,687	0,919	0,831
	Ölü	0,712	0,747	0,787	0,687	0,620	0,831
	Genel	0,900	0,898	0,897	0,687	0,857	0,831
Lojistik Regresyon	Yaşıyor	0,939	0,930	0,921	0,651	0,977	0,928
	Ölü	0,692	0,720	0,750	0,651	0,735	0,928
	Genel	0,888	0,886	0,885	0,651	0,928	0,928
J48	Yaşıyor	0,899	0,890	0,881	0,451	0,857	0,732
	Ölü	0,538	0,560	0,583	0,451	0,543	0,732
	Genel	0,824	0,821	0,819	0,451	0,792	0,732
Random Forest	Yaşıyor	0,965	0,920	0,880	0,557	0,977	0,926
	Ölü	0,500	0,612	0,788	0,557	0,774	0,926
	Genel	0,868	0,856	0,861	0,557	0,935	0,926
Hibrit Model	Yaşıyor	0,980	0,960	0,942	0,798	0,998	0,875
	Ölü	0,769	0,833	0,909	0,798	0,857	0,875
	Genel	0,936	0,934	0,935	0,798	0,969	0,875

Çizelge 37'deki veri madenciliği yöntemlerinin performanslarını karşılaştırmak için literatürle uyumlu olması açısından en önemli karar verici performans ölçütü olarak MKK belirlenmiş ve MKK değeri benzer (yakın) olan yöntemler için bu ölçüte ek olarak Accuracy ve F-measure değerlerine bakılması uygun görülmüştür. Bu ölçütlere bakıldığında en iyi performansa Hibrit Model ile ulaşıldığı görülmektedir. Hibrit Model'i sırasıyla Destek Vektör Makinesi, Lojistik Regresyon, Çok Katmanlı Algılayıcı, Random Forest ve J48 yöntemleri izlemektedir.

Hepatit 500 hastadan oluşan simüle veri seti için Chi-squared Attribute Eval., Gain Ratio Attribute Eval ve Info Gain Attribute Eval yöntemleri kullanılarak değişken önemine bakılmıştır. Klinik ve istatistiksel olarak önemsiz değişken olmadığı için analizler tüm veri seti üzerinden yapılmıştır. Veri setinde 19 bağımsız ve bir bağımlı değişken vardır. Bu değişkenler Albumin, Karın İltihabı, Bilirubin, Protime, Spiders, Histoloji, Kırgınlık, Varis, Tükenmişlik, Yaş, Spleen Palpable, Alk

fosfat, Cinsiyet, Antiviraller, Steroid, İştahsızlık, Karaciğerde Büyüme, Karaciğerde Sertlik, Sgot ve Sınıf'tır.

Çizelge 38. Hepatit 500 Hastadan Oluşan Simüle Veri Seti.

Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Çok Katmanlı Algılayıcı	Yaşıyor	0,947	0,945	0,942	0,731	0,986	0,953
	Ölü	0,779	0,786	0,794	0,731	0,865	0,953
	Genel	0,912	0,912	0,911	0,731	0,953	0,953
Destek Vektör Makinesi	Yaşıyor	0,962	0,950	0,938	0,750	0,933	0,861
	Ölü	0,760	0,798	0,840	0,750	0,688	0,861
	Genel	0,920	0,920	0,918	0,750	0,882	0,861
Lojistik Regresyon	Yaşıyor	0,960	0,949	0,938	0,744	0,985	0,951
	Ölü	0,760	0,794	0,832	0,744	0,857	0,951
	Genel	0,918	0,917	0,916	0,744	0,958	0,951
J48	Yaşıyor	0,949	0,934	0,919	0,665	0,916	0,846
	Ölü	0,683	0,728	0,780	0,665	0,657	0,846
	Genel	0,894	0,891	0,890	0,665	0,862	0,846
Random Forest	Yaşıyor	0,972	0,947	0,923	0,725	0,987	0,957
	Ölü	0,692	0,770	0,867	0,725	0,880	0,957
	Genel	0,914	0,910	0,912	0,725	0,965	0,957
Hibrit Model	Yaşıyor	0,990	0,984	0,978	0,920	0,998	0,952
	Ölü	0,913	0,936	0,960	0,920	0,948	0,952
	Genel	0,974	0,974	0,974	0,920	0,988	0,952

Çizelge 38'deki veri madenciliği yöntemlerinin performanslarını karşılaştırmak için literatürle uyumlu olması açısından en önemli karar verici performans ölçütü olarak MKK belirlenmiş ve MKK değeri benzer (yakın) olan yöntemler için bu ölçüte ek olarak Accuracy ve F-measure değerlerine bakılması uygun görülmüştür. Bu ölçütlere bakıldığında en iyi performansa Hibrit Model ile ulaşıldığı görülmektedir. Hibrit Model'i sırasıyla Destek Vektör Makinesi, Lojistik Regresyon, Çok Katmanlı Algılayıcı, Random Forest ve J48 yöntemleri izlemektedir.

3.5.2. Hepatit Veri Seti için Her Bir Kümeye Ait Sonuçlar

Hepatit gerçek veri setinde ayrılan birinci küme için Chi-squared Attribute Eval., Gain Ratio Attribute Eval ve Info Gain Attribute Eval yöntemleri kullanılarak değişken önemine bakılmıştır. Klinik ve istatistiksel olarak önemsiz değişken olduğu düşünülen Bilirubin, Sgot, Alk Fosfat ve Yaş değişkenleri veri setinden çıkarılmıştır. Veri setinde analiz için 15 bağımsız ve bir bağımlı değişken vardır. Bu değişkenler Albumin, Karın İltihabı, Protime, Spiders, Histoloji, Kırgınlık, Varis, Tükenmişlik, Spleen Palpable, Cinsiyet, Antiviraller, Steroid, İştahsızlık, Karaciğerde Büyüme, Karaciğerde Sertlik ve Sınıf'tır.

Çizelge 39. Hepatit Gerçek Veri Seti için 1. Kümeye ait Sonuçlar.

Yöntemler		**Performans Ölçütleri**					
		Doğru Sınıflama Oranı	**F-ölçütü**	**Kesinlik**	**MKK**	**PRC Alanı**	**EAKA**
Çok Katmanlı Algılayıcı	**Yaşıyor**	0,848	0,789	0,737	0,452	0,697	0,676
	Ölü	0,543	0,651	0,737	0,452	0,600	0,676
	Genel	0,737	0,731	0,737	0,452	0,656	0,676
Destek Vektör Makinesi	**Yaşıyor**	0,818	0,783	0,750	0,454	0,719	0,722
	Ölü	0,625	0,667	0,714	0,454	0,604	0,722
	Genel	0,737	0,734	0,735	0,454	0,671	0,722
Lojistik Regresyon	**Yaşıyor**	0,667	0,677	0,688	0,249	0,734	0,683
	Ölü	0,583	0,571	0,560	0,249	0,610	0,683
	Genel	0,632	0,633	0,634	0,249	0,682	0,683
J48	**Yaşıyor**	0,667	0,629	0,595	0,043	0,610	0,536
	Ölü	0,375	0,409	0,450	0,043	0,472	0,536
	Genel	0,544	0,536	0,534	0,043	0,552	0,536
Random Forest	**Yaşıyor**	0,727	0,762	0,800	0,472	0,705	0,718
	Ölü	0,750	0,706	0,667	0,472	0,697	0,718
	Genel	0,737	0,738	0,744	0,472	0,702	0,718
Hibrit Model	**Yaşıyor**	0,939	0,912	0,886	0,784	0,961	0,886
	Ölü	0,833	0,870	0,909	0,784	0,887	0,886
	Genel	0,895	0,894	0,896	0,784	0,930	0,886

Çizelge 39'daki veri madenciliği yöntemlerinin performanslarını karşılaştırmak için literatürle uyumlu olması açısından en önemli karar verici performans ölçütü olarak MKK belirlenmiş ve MKK değeri benzer (yakın) olan yöntemler için bu ölçüte ek olarak Accuracy ve F-measure değerlerine bakılması uygun görülmüştür.

Bu ölçütlere bakıldığında en iyi performansa Hibrit Model ile ulaşıldığı görülmektedir. Hibrit Model'i sırasıyla Random Forest, Destek Vektör Makinesi, Çok Katmanlı Algılayıcı, Lojistik Regresyon ve J48 yöntemleri izlemektedir.

Hepatit gerçek veri setinde ayrılan ikinci küme için Chi-squared Attribute Eval., Gain Ratio Attribute Eval ve Info Gain Attribute Eval yöntemleri kullanılarak değişken önemine bakılmıştır. Klinik ve istatistiksel olarak önemsiz değişken olduğu düşünülen Sgot, Bilirubin, Alk Fosfat, Protime ve Yaş değişkenleri veri setinden çıkarılmıştır. Veri setinde analiz için 14 bağımsız ve bir bağımlı değişken vardır. Bu değişkenler Albumin, Karın İltihabı, Spiders, Histoloji, Kırgınlık, Varis, Tükenmişlik, Spleen Palpable, Cinsiyet, Antiviraller, Steroid, İştahsızlık, Karaciğerde Büyüme, Karaciğerde Sertlik ve Sınıf'tır.

Çizelge 40. Hepatit Gerçek Veri Seti için 2. Kümeye ait Sonuçlar.

Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Çok Katmanlı Algılayıcı	Yaşıyor	0,956	0,950	0,945	0,351	0,974	0,860
	Ölü	0,375	0,400	0,429	0,351	0,614	0,860
	Genel	0,908	0,905	0,903	0,351	0,945	0,860
Destek Vektör Makinesi	Yaşıyor	0,978	0,967	0,957	0,546	0,956	0,739
	Ölü	0,500	0,571	0,667	0,546	0,374	0,739
	Genel	0,939	0,935	0,933	0,546	0,908	0,739
Lojistik Regresyon	Yaşıyor	0,922	0,938	0,954	0,366	0,961	0,763
	Ölü	0,500	0,421	0,364	0,366	0,406	0,839
	Genel	0,888	0,896	0,906	0,366	0,915	0,769
J48	Yaşıyor	0,944	0,955	0,966	0,515	0,923	0,630
	Ölü	0,625	0,556	0,500	0,515	0,283	0,630
	Genel	0,918	0,922	0,928	0,515	0,871	0,630
Random Forest	Yaşıyor	1,000	0,963	0,928	0,341	0,990	0,901
	Ölü	0,125	0,222	1,000	0,341	0,494	0,901
	Genel	0,929	0,902	0,934	0,341	0,949	0,901
Hibrit Model	Yaşıyor	0,978	0,978	0,978	0,728	0,998	0,864
	Ölü	0,750	0,750	0,750	0,728	0,875	0,864
	Genel	0,959	0,959	0,959	0,728	0,988	0,864

Çizelge 40'daki veri madenciliği yöntemlerinin performanslarını karşılaştırmak için literatürle uyumlu olması açısından en önemli karar verici performans ölçütü

olarak MKK belirlenmiş ve MKK değeri benzer (yakın) olan yöntemler için bu ölçüte ek olarak Accuracy ve F-measure değerlerine bakılması uygun görülmüştür. Bu ölçütlere bakıldığında en iyi performansa Hibrit Model ile ulaşıldığı görülmektedir. Hibrit Model'i sırasıyla Destek Vektör Makinesi, J48, Lojistik Regresyon, Çok Katmanlı Algılayıcı ve Random Forest yöntemleri izlemektedir. Bu veri seti için Lojistik Regresyon, Çok Katmanlı Algılayıcı ve Random Forest yöntemlerinin performanslarının üç değerlendirme ölçütü birlikte değerlendirildiğinde çok benzer olduğu söylenmektedir.

Hepatit 250 hastadan oluşan simüle veri setinde ayrılan birinci küme için Chi-squared Attribute Eval., Gain Ratio Attribute Eval. ve Info Gain Attribute Eval. yöntemleri kullanılarak değişken önemine bakılmıştır. Klinik ve istatistiksel olarak önemsiz değişken olduğu düşünülen Alk Fosfat ve Sgot değişkenleri veri setinden çıkarılmıştır. Veri setinde analiz için 17 bağımsız ve bir bağımlı değişken vardır. Bu değişkenler Albumin, Karın İltihabı, Bilirubin, Protime, Spiders, Histoloji, Kırgınlık, Varis, Tükenmişlik, Yaş, Spleen Palpable, Cinsiyet, Antiviraller, Steroid, İştahsızlık, Karaciğerde Büyüme, Karaciğerde Sertlik ve Sınıf'tır.

Çizelge 41. Hepatit 250 Hastadan Oluşan Simüle Veri Seti için 1. Kümeye Ait Sonuçlar.

Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Çok Katmanlı Algılayıcı	Yaşıyor	0,800	0,814	0,828	0,564	0,894	0,838
	Ölü	0,767	0,750	0,733	0,564	0,736	0,838
	Genel	0,786	0,787	0,788	0,564	0,828	0,838
Destek Vektör Makinesi	Yaşıyor	0,833	0,855	0,877	0,665	0,828	0,835
	Ölü	0,837	0,809	0,783	0,665	0,723	0,835
	Genel	0,835	0,836	0,838	0,665	0,784	0,835
Lojistik Regresyon	Yaşıyor	0,800	0,800	0,800	0,521	0,914	0,862
	Ölü	0,721	0,721	0,721	0,521	0,749	0,862
	Genel	0,767	0,767	0,767	0,521	0,845	0,862
J48	Yaşıyor	0,717	0,775	0,843	0,523	0,873	0,808
	Ölü	0,814	0,737	0,673	0,523	0,662	0,808
	Genel	0,757	0,759	0,772	0,523	0,785	0,808
Random Forest	Yaşıyor	0,817	0,831	0,845	0,604	0,913	0,867
	Ölü	0,791	0,773	0,756	0,604	0,819	0,867
	Genel	0,806	0,806	0,808	0,604	0,874	0,867
Hibrit Model	Yaşıyor	0,917	0,917	0,917	0,800	0,972	0,900
	Ölü	0,884	0,884	0,884	0,800	0,897	0,900
	Genel	0,903	0,903	0,903	0,800	0,941	0,900

Çizelge 41'deki veri madenciliği yöntemlerinin performanslarını karşılaştırmak için literatürle uyumlu olması açısından en önemli karar verici performans ölçütü olarak MKK belirlenmiş ve MKK değeri benzer (yakın) olan yöntemler için bu ölçüte ek olarak Accuracy ve F-measure değerlerine bakılması uygun görülmüştür. Bu ölçütlere bakıldığında en iyi performansa Hibrit Model ile ulaşıldığı görülmektedir. Hibrit Model'i sırasıyla Destek Vektör Makinesi, Random Forest, Çok Katmanlı Algılayıcı, Lojistik Regresyon ve J48 yöntemleri izlemektedir.

Hepatit 250 hastadan oluşan simüle veri setinde ayrılan ikinci küme için Chi-squared Attribute Eval., Gain Ratio Attribute Eval ve Info Gain Attribute Eval yöntemleri kullanılarak değişken önemine bakılmıştır. Klinik ve istatistiksel olarak önemsiz değişken olduğu düşünülen Yaş, Protime, Alk Fosfat, Karın İltihabı, Albumin ve Sgot değişkenleri veri setinden çıkarılmıştır. Veri setinde analiz için 13 bağımsız ve bir bağımlı değişken vardır. Bu değişkenler Bilirubin, Spiders, Histoloji,

Kırgınlık, Varis, Tükenmişlik, Spleen Palpable, Cinsiyet, Antiviraller, Steroid, İştahsızlık, Karaciğerde Büyüme, Karaciğerde Sertlik ve Sınıf'tır.

Çizelge 42. Hepatit 250 Hastadan Oluşan Simüle Veri Seti için 2. Kümeye ait Sonuçlar.

Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Çok Katmanlı Algılayıcı	Yaşıyor	0,957	0,971	0,985	0,620	0,988	0,888
	Ölü	0,778	0,636	0,538	0,620	0,532	0,888
	Genel	0,946	0,950	0,958	0,620	0,960	0,888
Destek Vektör Makinesi	Yaşıyor	0,986	0,968	0,951	0,306	0,951	0,604
	Ölü	0,222	0,308	0,500	0,306	0,159	0,604
	Genel	0,939	0,928	0,923	0,306	0,902	0,604
Lojistik Regresyon	Yaşıyor	0,964	0,967	0,971	0,494	0,987	0,874
	Ölü	0,556	0,526	0,500	0,494	0,351	0,845
	Genel	0,939	0,940	0,942	0,494	0,948	0,872
J48	Yaşıyor	0,978	0,971	0,964	0,476	0,971	0,771
	Ölü	0,444	0,500	0,571	0,476	0,286	0,771
	Genel	0,946	0,942	0,940	0,476	0,929	0,771
Random Forest	Yaşıyor	0,978	0,968	0,957	0,378	0,979	0,833
	Ölü	0,333	0,400	0,500	0,378	0,550	0,833
	Genel	0,939	0,933	0,929	0,378	0,953	0,833
Hibrit Model	Yaşıyor	0,993	0,989	0,986	0,814	0,998	0,885
	Ölü	0,778	0,824	0,875	0,814	0,910	0,885
	Genel	0,980	0,979	0,979	0,814	0,993	0,885

Çizelge 42'deki veri madenciliği yöntemlerinin performanslarını karşılaştırmak için literatürle uyumlu olması açısından en önemli karar verici performans ölçütü olarak MKK belirlenmiş ve MKK değeri benzer (yakın) olan yöntemler için bu ölçüte ek olarak Accuracy ve F-measure değerlerine bakılması uygun görülmüştür. Bu ölçütlere bakıldığında en iyi performansa Hibrit Model ile ulaşıldığı görülmektedir. Hibrit Model'i sırasıyla Çok Katmanlı Algılayıcı, Lojistik Regresyon, J48, Random Forest ve Destek Vektör Makinesi yöntemleri izlemektedir.

Hepatit 500 hastadan oluşan simüle veri setinde ayrılan birinci küme için Chi-squared Attribute Eval., Gain Ratio Attribute Eval ve Info Gain Attribute Eval yöntemleri kullanılarak değişken önemine bakılmıştır. Klinik ve istatistiksel olarak

tüm değişkenler önemli bulunduğu için hiçbir değişken veri setinden çıkarılmamıştır. Veri setinde analiz için 19 bağımsız ve bir bağımlı değişken vardır. Bu değişkenler Albumin, Karın İltihabı, Bilirubin, Protime, Spiders, Histoloji, Kırgınlık, Varis, Tükenmişlik, Yaş, Spleen Palpable, Cinsiyet, Antiviraller, Steroid, İştahsızlık, Karaciğerde Büyüme, Karaciğerde Sertlik, Alk Fosfat, Sgot ve Sınıf'tır.

Çizelge 43. Hepatit 500 Hastadan Oluşan Simüle Veri Seti için 1. Kümeye Ait Sonuçlar.

Yöntemler		**Performans Ölçütleri**					
		Doğru Sınıflama Oranı	**F-ölçütü**	**Kesinlik**	**MKK**	**PRC Alanı**	**EAKA**
Çok Katmanlı Algılayıcı	**Yaşıyor**	0,909	0,905	0,902	0,771	0,970	0,950
	Ölü	0,860	0,865	0,871	0,771	0,918	0,950
	Genel	0,889	0,889	0,889	0,771	0,948	0,950
Destek Vektör Makinesi	**Yaşıyor**	0,876	0,883	0,891	0,723	0,853	0,862
	Ölü	0,849	0,839	0,830	0,723	0,767	0,862
	Genel	0,865	0,865	0,865	0,723	0,817	0,862
Lojistik Regresyon	**Yaşıyor**	0,893	0,885	0,878	0,721	0,953	0,924
	Ölü	0,826	0,835	0,845	0,721	0,870	0,924
	Genel	0,865	0,864	0,864	0,721	0,918	0,924
J48	**Yaşıyor**	0,818	0,818	0,818	0,562	0,841	0,797
	Ölü	0,744	0,744	0,744	0,562	0,669	0,797
	Genel	0,787	0,787	0,787	0,562	0,770	0,797
Random Forest	**Yaşıyor**	0,843	0,850	0,857	0,643	0,947	0,927
	Ölü	0,802	0,793	0,784	0,643	0,907	0,927
	Genel	0,826	0,826	0,827	0,643	0,931	0,927
Hibrit Model	**Yaşıyor**	0,959	0,959	0,959	0,901	0,998	0,950
	Ölü	0,942	0,942	0,942	0,901	0,932	0,950
	Genel	0,952	0,952	0,952	0,901	0,971	0,950

Çizelge 43'deki veri madenciliği yöntemlerinin performanslarını karşılaştırmak için literatürle uyumlu olması açısından en önemli karar verici performans ölçütü olarak MKK belirlenmiş ve MKK değeri benzer (yakın) olan yöntemler için bu ölçüte ek olarak Accuracy ve F-measure değerlerine bakılması uygun görülmüştür. Bu ölçütlere bakıldığında en iyi performansa Hibrit Model ile ulaşıldığı görülmektedir. Hibrit Model'i sırasıyla Çok Katmanlı Algılayıcı, Destek Vektör Makinesi, Lojistik Regresyon, Random Forest ve J48 yöntemleri izlemektedir.

Hepatit 500 hastadan oluşan simüle veri setinde ayrılan ikinci küme için Chi-squared Attribute Eval., Gain Ratio Attribute Eval ve Info Gain Attribute Eval yöntemleri kullanılarak değişken önemine bakılmıştır. Klinik ve istatistiksel olarak önemsiz değişken olduğu düşünülen Sgot, Karın İltihabı ve Alk Fosfat değişkenleri veri setinden çıkarılmıştır. Veri setinde analiz için 16 bağımsız ve bir bağımlı değişken vardır. Bu değişkenler Albumin, Bilirubin, Protime, Spiders, Histoloji, Kırgınlık, Varis, Tükenmişlik, Yaş, Spleen Palpable, Cinsiyet, Antiviraller, Steroid, İştahsızlık, Karaciğerde Büyüme, Karaciğerde Sertlik ve Sınıf'tır.

Çizelge 44. Hepatit 500 Hastadan Oluşan Simüle Veri Seti için 2. Kümeye Ait Sonuçlar.

Yöntemler		**Performans Ölçütleri**					
		Doğru Sınıflama Oranı	**F-ölçütü**	**Kesinlik**	**MKK**	**PRC Alanı**	**EAKA**
Çok Katmanlı Algılayıcı	**Yaşıyor**	0,975	0,978	0,982	0,663	0,995	0,941
	Ölü	0,722	0,684	0,650	0,663	0,689	0,941
	Genel	0,959	0,960	0,961	0,663	0,977	0,941
Destek Vektör Makinesi	**Yaşıyor**	0,996	0,986	0,975	0,736	0,975	0,804
	Ölü	0,611	0,733	0,917	0,736	0,584	0,804
	Genel	0,973	0,970	0,971	0,736	0,951	0,804
Lojistik Regresyon	**Yaşıyor**	0,989	0,987	0,986	0,788	0,988	0,908
	Ölü	0,778	0,800	0,824	0,788	0,661	0,913
	Genel	0,976	0,976	0,976	0,788	0,968	0,908
J48	**Yaşıyor**	0,985	0,973	0,961	0,473	0,954	0,742
	Ölü	0,389	0,483	0,636	0,473	0,443	0,742
	Genel	0,949	0,943	0,941	0,473	0,923	0,742
Random Forest	**Yaşıyor**	1,000	0,980	0,962	0,611	0,985	0,865
	Ölü	0,389	0,560	1,000	0,611	0,701	0,865
	Genel	0,962	0,955	0,964	0,611	0,968	0,865
Hibrit Model	**Yaşıyor**	1,000	0,995	0,989	0,908	0,999	0,917
	Ölü	0,833	0,909	1,000	0,908	0,942	0,917
	Genel	0,990	0,989	0,990	0,908	0,995	0,917

Çizelge 44'deki veri madenciliği yöntemlerinin performanslarını karşılaştırmak için literatürle uyumlu olması açısından en önemli karar verici performans ölçütü olarak MKK belirlenmiş ve MKK değeri benzer (yakın) olan yöntemler için bu ölçüte ek olarak Accuracy ve F-measure değerlerine bakılması uygun görülmüştür. Bu ölçütlere bakıldığında en iyi performansa Hibrit Model ile ulaşıldığı

görülmektedir. Hibrit Model'i sırasıyla Lojistik Regresyon, Destek Vektör Makinesi, Çok Katmanlı Algılayıcı, Random Forest ve J48 yöntemleri izlemektedir.

3.5.3. Meme Kanseri Veri Seti için Bulgular

Meme kanseri veri seti için sonuç değişkeni bazlı tanımlayıcılar ve bu değişkene göre istatistiksel karşılatırmalar Çizelge 45'te verilmiştir.

Çizelge 45. Nicel Değişkenler için Tanımlayıcılar ve İstatistiksel Karşılaştırmalar.

Değişkenler	Sınıf				p değeri
	Sağlıklı (n=52)		Hasta (n=64)		
	Ort.±SS	Ortanca (Min.-Maks.)	Ort.±SS	Ortanca (Min.-Maks.)	
Yaş	58,08±18,96	65,00 (24,00-89,00)	56,67±13,49	53,00 (34,00-86,00)	0,477[a]
VKİ	28,32±5,43	27,69 (18,67-38,58)	26,98±4,62	27,41 (18,37-37,11)	0,201[a]
Glukoz	88,23±10,19	87,00 (60,00-118,00)	105,56±26,56	98,50 (70,00-201,00)	<0,001[a]
İnsülin	6,93±4,86	5,48 (2,71-26,21)	12,51±12,32	7,58 (2,43-58,46)	0,026[a]
HOMA	1,55±1,22	1,14 (0,47-7,11)	3,62±4,59	2,05 (0,51-25,05)	0,003[a]
Leptin	26,64±19,33	21,49 (4,31-83,48)	26,60±19,21	18,88 (6,33-90,28)	0,947[a]
Adiponektin	10,33±7,63	8,13 (2,19-38,04)	10,06±6,19	8,45 (1,66-33,75)	0,764[a]
Resistin	11,61±11,45	8,93 (3,29-82,10)	17,25±12,64	14,37 (3,21-55,22)	0,002[a]
MCP1	499,73±292,24	471,32 (45,84-1256,08)	563,02±384,00	465,37 (90,09-1698,44)	0,502[a]

a:Mann-Whitney U testi, Ort.:Ortalama, SS:Standart Sapma, Min.:Minimum, Maks.: Maksimum

Çizelge 45'te nicel değişkenler için Sınıf değişkeni kategorilerilerine (Sağlıklı, Hasta) göre istatistiksel olarak anlamlı farklılıklar ve tanımlayıcı istatistikler verilmiştir. Nicel değişkenlerden glukoz, insülin, HOMA ve Resistin için farklar

istatistiksel olarak anlamlı bulunmuştur (sırasıyla p<0,001, p=0,026, p=0,003 ve p=0,002).

Meme kanseri gerçek veri seti için Chi-squared Attribute Eval., Gain Ratio Attribute Eval ve Info Gain Attribute Eval yöntemleri kullanılarak değişken önemine bakılmıştır. Klinik ve istatistiksel olarak önemsiz değişken olduğu düşünülen MCP1, Insulin, Adiponektin, BMI, Resistin ve Leptin değişkenleri veri setinden çıkarılmıştır. Veri setinde analiz için 3 bağımsız ve bir bağımlı değişken vardır. Bu değişkenler Glukoz, HOMA, Yaş ve Sınıf'tır.

Çizelge 46. Meme Kanseri Gerçek Veri Seti.

Yöntemler		**Performans Ölçütleri**					
		Doğru Sınıflama Oranı	**F-ölçütü**	**Kesinlik**	**MKK**	**PRC Alanı**	**EAKA**
Lojistik Regresyon	**Sağlıklı**	0,712	0,725	0,740	0,511	0,663	0,756
	Hasta	0,797	0,785	0,773	0,511	0,808	0,756
	Genel	0,759	0,758	0,758	0,511	0,743	0,756
Destek Vektör Makinesi	**Sağlıklı**	0,250	0,371	0,722	0,236	0,517	0,586
	Hasta	0,922	0,728	0,602	0,236	0,598	0,586
	Genel	0,621	0,568	0,656	0,236	0,562	0,586
Ada Boost	**Sağlıklı**	0,635	0,641	0,647	0,354	0,718	0,736
	Hasta	0,719	0,713	0,708	0,354	0,758	0,736
	Genel	0,681	0,681	0,681	0,354	0,740	0,736
Decision Table	**Sağlıklı**	0,577	0,600	0,625	0,299	0,569	0,635
	Hasta	0,719	0,697	0,676	0,299	0,649	0,635
	Genel	0,655	0,654	0,653	0,299	0,613	0,635
Hoeffding Tree	**Sağlıklı**	0,885	0,676	0,548	0,324	0,789	0,782
	Hasta	0,406	0,542	0,813	0,324	0,805	0,782
	Genel	0,621	0,602	0,694	0,324	0,798	0,782
Hibrit Model	**Sağlıklı**	0,904	0,931	0,959	0,879	0,943	0,936
	Hasta	0,969	0,947	0,925	0,879	0,961	0,936
	Genel	0,940	0,939	0,941	0,879	0,953	0,936

Çizelge 46'daki veri madenciliği yöntemlerinin performanslarını karşılaştırmak için literatürle uyumlu olması açısından en önemli karar verici performans ölçütü olarak MKK belirlenmiş ve MKK değeri benzer (yakın) olan yöntemler için bu ölçüte ek olarak Accuracy ve F-measure değerlerine bakılması uygun görülmüştür. Bu ölçütlere bakıldığında en iyi performansa Hibrit Model ile ulaşıldığı

görülmektedir. Hibrit Model'i sırasıyla Lojistik Regresyon, Ada Boost, Hoeffding Tree, Decision Table ve Destek Vektör Makinesi yöntemleri izlemektedir.

Meme kanseri 250 hastadan oluşan simüle veri seti için Chi-squared Attribute Eval., Gain Ratio Attribute Eval ve Info Gain Attribute Eval yöntemleri kullanılarak değişken önemine bakılmıştır. Klinik ve istatistiksel olarak önemsiz değişken olduğu düşünülen MCP1, Adiponektin, BMI ve Leptin değişkenleri veri setinden çıkarılmıştır. Veri setinde analiz için 5 bağımsız ve bir bağımlı değişken vardır. Bu değişkenler Glukoz, HOMA, Yaş, Insülin, Resistin ve Sınıf'tır.

Çizelge 47. Meme Kanseri 250 Hastadan Oluşan Simüle Veri Seti.

Yöntemler		Performans Ölçütleri					
		Doğru Sınıflama Oranı	F-ölçütü	Kesinlik	MKK	PRC Alanı	EAKA
Lojistik Regresyon	Sağlıklı	0,768	0,717	0,672	0,461	0,739	0,813
	Hasta	0,696	0,738	0,787	0,461	0,851	0,813
	Genel	0,728	0,729	0,735	0,461	0,801	0,813
Destek Vektör Makinesi	Sağlıklı	0,857	0,736	0,644	0,479	0,616	0,737
	Hasta	0,616	0,711	0,842	0,479	0,730	0,737
	Genel	0,724	0,722	0,753	0,479	0,679	0,737
IBk	Sağlıklı	0,741	0,712	0,686	0,463	0,624	0,733
	Hasta	0,725	0,749	0,775	0,463	0,714	0,733
	Genel	0,732	0,733	0,735	0,463	0,674	0,733
Bagging	Sağlıklı	0,714	0,696	0,678	0,437	0,801	0,807
	Hasta	0,725	0,741	0,758	0,437	0,827	0,807
	Genel	0,720	0,721	0,722	0,437	0,815	0,807
Hoeffding Tree	Sağlıklı	0,920	0,746	0,628	0,500	0,830	0,844
	Hasta	0,558	0,688	0,895	0,500	0,856	0,844
	Genel	0,720	0,714	0,776	0,500	0,844	0,844
Hibrit Model	Sağlıklı	0,964	0,904	0,850	0,822	0,912	0,913
	Hasta	0,862	0,912	0,967	0,822	0,918	0,913
	Genel	0,908	0,908	0,915	0,822	0,915	0,913

Çizelge 47'deki veri madenciliği yöntemlerinin performanslarını karşılaştırmak için literatürle uyumlu olması açısından en önemli karar verici performans ölçütü olarak MKK belirlenmiş ve MKK değeri benzer (yakın) olan yöntemler için bu ölçüte ek olarak Accuracy ve F-measure değerlerine bakılması uygun görülmüştür. Bu ölçütlere bakıldığında en iyi performansa Hibrit Model ile ulaşıldığı

görülmektedir. Hibrit Model'i sırasıyla Hoeffding Tree, Destek Vektör Makinesi, IBk, Lojistik Regresyon ve Bagging yöntemleri izlemektedir.

Meme kanseri 500 hastadan oluşan simüle veri seti için Chi-squared Attribute Eval., Gain Ratio Attribute Eval ve Info Gain Attribute Eval yöntemleri kullanılarak değişken önemine bakılmıştır. Klinik ve istatistiksel olarak önemsiz değişken olduğu düşünülen Adiponektin değişkeni veri setinden çıkarılmıştır. Veri setinde analiz için 8 bağımsız ve bir bağımlı değişken vardır. Bu değişkenler Glukoz, HOMA, Yaş, Insülin, Resistin, MCP1, Leptin, BMI ve Sınıf'tır.

Çizelge 48. Meme Kanseri 500 Hastadan Oluşan Simüle Veri Seti.

Yöntemler		**Performans Ölçütleri**					
		Doğru Sınıflama Oranı	**F-ölçütü**	**Kesinlik**	**MKK**	**PRC Alanı**	**EAKA**
Naive Bayes	**Sağlıklı**	0,942	0,805	0,703	0,629	0,866	0,894
	Hasta	0,678	0,786	0,935	0,629	0,918	0,894
	Genel	0,796	0,795	0,831	0,629	0,895	0,894
Ada Boost	**Sağlıklı**	0,902	0,794	0,709	0,604	0,762	0,843
	Hasta	0,699	0,786	0,898	0,604	0,891	0,843
	Genel	0,790	0,790	0,813	0,604	0,833	0,843
Hoeffding Tree	**Sağlıklı**	0,821	0,785	0,751	0,597	0,807	0,872
	Hasta	0,779	0,810	0,843	0,597	0,900	0,872
	Genel	0,798	0,799	0,802	0,597	0,858	0,872
Random Forest	**Sağlıklı**	0,835	0,801	0,770	0,629	0,849	0,896
	Hasta	0,797	0,826	0,856	0,629	0,910	0,896
	Genel	0,814	0,814	0,817	0,629	0,883	0,896
Bagging	**Sağlıklı**	0,844	0,791	0,744	0,605	0,840	0,885
	Hasta	0,764	0,808	0,858	0,605	0,905	0,885
	Genel	0,800	0,801	0,807	0,605	0,876	0,885
Hibrit Model	**Sağlıklı**	0,960	0,900	0,846	0,814	0,906	0,909
	Hasta	0,859	0,908	0,963	0,814	0,917	0,909
	Genel	0,904	0,904	0,911	0,814	0,912	0,909

Çizelge 48'deki veri madenciliği yöntemlerinin performanslarını karşılaştırmak için literatürle uyumlu olması açısından en önemli karar verici performans ölçütü olarak MKK belirlenmiş ve MKK değeri benzer (yakın) olan yöntemler için bu ölçüte ek olarak Accuracy ve F-measure değerlerine bakılması uygun görülmüştür. Bu ölçütlere bakıldığında en iyi performansa Hibrit Model ile ulaşıldığı

görülmektedir. Hibrit Model'i sırasıyla Random Forest, Naive Bayes, Bagging, Ada Boost ve Hoeffding Tree yöntemleri izlemektedir.

4. TARTIŞMA

Dijital teknolojideki gelişmeler, toplanan verilerin boyutu, karmaşıklığı ve miktarında ciddi bir büyümeye yol açmıştır. Kullanılan sistemler bu sayede farklı tipteki verilerin yüksek kalitede saklanmasına olanak sağlamıştır. Bu gelişmelere paralel olarak sağlık alanındaki verilerin boyutunda da öngörülemez bir atış olmuştur. Ayrıca, algoritmalar ve görselleştirme araçları gibi veri madenciliği algoritmalarındaki son gelişmeler, her türlü ham verinin üst düzey bilgiye dönüştürülmesini mümkün kılmıştır. Artan veri ile birlikte veri madenciliği algoritmalarında ortaya çıkan sorun ise her yöntemin veri yapısı, şekli ve geçerliliği ile ilgili kendi yaklaşımına sahip olmasıdır. Bu sınırlama sınıflandırma sistemlerinin performansını etkilemektedir. Bu yüzden hibrit bir veri madenciliği yaklaşımına duyulan ihtiyaç, veri madenciliği topluluğu tarafından kabul görmüş ve bu konuda yapılan çalışma sayısı artmıştır.

Ha ve Joo (2010), bir hastanenin acil servisinde göğüs ağrısı şikayetiyle tedavi gören 478 hasta üzerinde yaptıkları çalışmada, C5.0 adında boosting tabanlı oluşturdukları hibrit model ile Sinir Ağı ve Destek Vektör Makinası yöntemlerinin sonuçlarını karşılaştırmıştır. Yöntemlere ait doğru sınıflama oranlarını ise hibrit model için %94,18, Destek Vektör Makinası için %85,19 ve Sinir Ağı için %88,89 olarak bulmuşlar ve hibrit modelin en iyi performansa sahip yöntem olduğunu bildirmişlerdir.

Simsek ve ark. (2020), 240 meme kanseri hastasına ait 7399 gen üzerinde yaptıkları çalışmada bir regresyon yöntemi (LASSO) ve meta-sezgisel optimizasyon yöntemini birleştirerek Genetik Algoritma adını verdikleri bir hibrit model oluşturmuşlardır, Hibrit model sonuçlarını ise Yapay Sinir Ağı ve Lojistik Regresyon yöntemleri ile karşılaştırmışlardır. Sonuç olarak ise verileri 1, 3 ve 5 yıllık sağkalım olmak üzere 3 ayrı veri setine bölmüşler ve bu veri setleri üzerinde toplamda 360 analiz gerçekleştirmişler. Analizler sonucunda ise genel olarak hibrit modelin Yapay

Sinir Ağı ve Lojistik Regresyon yöntemlerine göre daha iyi performansa sahip olduğunu bildirmişlerdir.

El-Rashidy ve ark. (2011), Wisconsin Meme Kanseri veri setinde bulunan 699 hasta ile yaptıkları çalışmada Destek Özellik Makinası (Support Feature Machine), Bulanık C-ortalaması (Fuzzy C-means) ve maksimum-minimum yaklaşımını birleştirerek OCSFM adını verdikleri bir hibrit model oluşturmuşlardır. Veri setini 20 ayrı kümeye böldükten sonra her bir küme için ayrı ayrı hibrit modelin performansını test etmişlerdir. Sonuç olarak ise her kümede ve genel sınıflama performansı olarak doğru sınıflama oranı ve MKK değerlerinin yaklaşık %90 ve üzerinde olduğunu bildirmişlerdir.

Verma ve ark. (2016), 335 koroner arter hastası üzerinde yaptıkları çalışmada veri madenciliği yöntemi olarak Çok Katmanlı Algılayıcı, Lojistik Regresyon, Bulanık sırasız kural indüksiyon algoritması (FURIA) ve C4.5 algoritması kullanmışlardır. Bu yöntemlerin her birine korelasyona dayalı özellik alt kümesi ve parçacık sürüsü optimizasyonu (particle swan optimization) uygulayarak 4 ayrı hibrit model oluşturmuşlar ve her bir yöntemin performansını hibrit hali ile karşılaştırmışlardır. Çok Katmanlı Algılayıcı yöntemine ait doğru sınıflama oranını %77,0, hibrit model bazlı Çok Katmanlı Algılayıcı yöntemi performansını ise %79,7 olarak bulmuşlardır. Lojistik Regresyon yöntemine ait doğru sınıflama oranını %83,5, hibrit model bazlı Lojistik Regresyon yöntemi performansını ise %84,17 olarak bulmuşlardır. FURIA yöntemine ait doğru sınıflama oranını %77,9, hibrit model bazlı FURIA yöntemi performansını ise %79,7 olarak bulmuşlardır. C4.5 yöntemine ait doğru sınıflama oranını %77,3, hibrit model bazlı C4.5 yöntemi performansını ise %77,9 olarak bulmuşlardır.

Bayat ve Motevalli Almouti (2016), UCI veri tabanında bulunan 200 hepatit B hastası içeren veri seti ile yaptıkları çalışmada, Sinir Ağı, Naive Bayes, Rough Set yöntemini birleştirerek hibrit model oluşturmuşlardır. Hibrit modelin performansını ise Karar Ağacı, Destek Vektör Makinası, Sinir Ağı, Naive Bayes ve Rough set yöntemleri ile karşılaştırmışlardır. Performans ölçütlerinden doğru sınıflama oranı ve

F-ölçütü değerlerini ise Karar Ağacı için %75,0 ve %78,8, Destek Vektör Makinası için %77,0 ve %80,3, Sinir Ağı için %88,0 ve %87,1, Naive Bayes için %87,0 ve %85,6, Rough Set için %89,0 ve %86,7, hibrit model için ise %90,0 ve %88,8 olarak bulmuşlardır. Sonuç olarak ise hibrit model performansının diğer yöntemlere göre daha iyi olduğunu bildirmişlerdir.

Rathi ve Parek (2016), çalışmalarında üç ayrı veri seti kullanmışlardır. Bunlar 1171 hastadan oluşan Diyabet veri seti, 699 hastadan oluşan Wisconsin Meme Kanseri veri seti ve 3163 hastaadan oluşan Hipotiroid veri setleridir. Araştırmacılar Destek Vektör Makinası ve Bagging yöntemlerini birleştirerek oluşturdukları hibrit modeli, Destek Vektör Makinası ve J48 yöntemi birleşiminden oluşan yöntemi, Destek Vektör Makinası ve Rep Tree yöntemi birleşiminden oluşan yöntemi ve Destek Vektör Makinası yöntemini bu üç veri setinde test etmişlerdir. Diyabet veri setinde doğru sınıflama oranlarını ise hibrit model için 0,843, Destek Vektör Makinası ve J48 yönteminden oluşan model için 0,825, Destek Vektör Makinası ve Rep Tree yönteminden oluşan model için 0,825 ve Destek Vektör Makinası için ise 0,825 olarak bulmuşlardır. Wisconsin Meme Kanseri veri setinde doğru sınıflama oranlarını ise hibrit model için 0,988, Destek Vektör Makinası ve J48 yönteminden oluşan model için 0,971, Destek Vektör Makinası ve Rep Tree yönteminden oluşan model için 0,971 ve Destek Vektör Makinası için ise 0,965 olarak bulmuşlardır. Hipotirioid veri setinde ise doğru sınıflama oranlarını ise hibrit model için 0,992, Destek Vektör Makinası ve J48 yönteminden oluşan model için 0,992, Destek Vektör Makinası ve Rep Tree yönteminden oluşan model için 0,992 ve Destek Vektör Makinası için ise 0,972 olarak bulmuşlardır. Araştırmacılar sonuç olarak hibrit bir veri madenciliği modeli oluşturmanın daha iyi sonuçlar verdiğini bildirmişlerdir.

Wang ve ark. (2006), 2934 kadın hasta üzerinde yaptıkları çalışmada 3 ayrı veri madenciliği yöntemini baz alarak 9 ayrı hibrit model oluşturmuşlardır. Verileri eğitim ve test verisi olarak ikiye ayırmışlar ve sonuçları eğitim ve test verileri için ayrı ayrı vermişlerdir. Sonuç olarak eğitim veri setinde modellere ait ortalama doğru sınıflama oranını %85,79 olarak, test veri setinden ise %65,29 olarak bulmuşlardır.

Ayrıca hibrit model yaklaşımlarının performansta yaklaşık %5'lik iyileştirme sağladığını bildirmişlerdir.

Çalışmamızda literatürle uyumlu olarak hibrit modele ait performans ölçütleri vei madenciliği yöntemlerine göre daha yüksek bulunmuştur. Üretilen simüle veriler üzerindeki 27 senaryoda ve gerçek veri setleri üzerinde yapılan analizler sonucunda hibrit modelin daha iyi sonuç verdiği, performansı ciddi oranda iyileştirdiği görülmüştür.

5. SONUÇ VE ÖNERİLER

Çalışmada üretilen simüle veri setleri incelendiğinde dengesiz dağılıma sahip veri setlerinde literatürle uyumlu olarak doğru sınıflama oranı ve F-ölçütü ölçütleri dengeli dağılıma sahip veri setlerine kıyasla daha yüksek bulunmuştur. Bunun nedeni genel performansın iki grubun ağırlıklı ortalaması alınarak hesaplanmasından dolayı ölçüt değerlerinin yüksek örnekleme sahip grubun değerlerine daha yakın ve yüksek olmasıdır. MKK ölçütü dengesiz dağılımdan etkilenmediği için bu ölçüte göre değerlendirildiğinde ise dengeli gruba sahip değerler daha yüksek bulunmuştur. Ayrıca örneklem büyüklüğü arttıkça MKK değerlerinin de arttığı görülmüştür.

Simüle veri setlerinde bağımlı değişken ile bağımsız değişkenler arası korelasyon arttıkça MKK değerlerinin de arttığı görülmüştür. Ayrıca dengesiz veri setlerinde bağımlı değişken ile bağımsız değişkenler arası korelasyon arttıkça düşük örnekleme sahip gruba ait performans ölçütlerinde de gözle görülür bir artış gözlemlenmiştir. Ayrıca dengesiz dağılımda gruplar arası dengesizlik azaldığında da tüm performans ölçütleri daha iyi sonuç vermektedir.

Gerçek veri setleri incelendiğinde de durumun benzer olduğu görülmektedir. Çalışmada kullanılan hepatit veri seti ilk olarak gerçek değerler kullanılarak değerlendirilmiş, daha sonra bu verinin tüm özellikleri korunarak örneklem büyüklüğü 250 ve 500'e yükseltilmiştir. Ayrıca veri setleri dengesiz dağılıma sahip olduğu için kümeleme yöntemi kullanılarak veri setleri kümelere ayrılmış ve bu kümelerin performansları ayrı ayrı değerlendirilmiş, sonrasında ise bu kümelerin performansları birleştirilerek genel veri seti performansına ulaşılmıştır. Elde edilen tüm veri setlerine ait sonuçlar değerlendirildiğinde ise örneklem büyüklüğü arttıkça performans ölçütlerinde de iyileşmeler olduğu gözlenmiştir.

Meme kanseri veri setinde de benzer olarak ilk aşamada gerçek değerler kullanılarak değerlendirme yapılmış, daha sonra bu verinin tüm özellikleri korunarak örneklem büyüklüğü 250 ve 500'e yükseltilmiştir. Üç veri seti için performans

ölçütleri hesaplanıp karşılaştırmalar yapıldığında ise örneklem büyüklüğü arttıkça performans ölçütlerinin de daha iyi sonuçlar verdiği görülmüştür.

Simüle veri setindeki tüm senaryolarda ve gerçek veri setlerinde hibrit model daha iyi sonuçlar vermiştir. Buradan uygun veri madenciliği yöntemlerinin birlikte kullanımı durumunda, yapılacak tüm çalışmalarda hibrit model yaklaşımının daha iyi modeller elde etmemizi sağlayacağı sonucuna ulaşılmaktadır.

ÖZET

Veri Madenciliğinde Hibrit Model Yaklaşımı

Son yıllarda, tıp disiplinlerinde toplanan veri miktarı giderek artmaktadır. Dijital teknolojideki gelişmeler, toplanan verilerin boyutu, karmaşıklığı ve miktarında, yani tıbbi raporlarda ve ilgili görüntülerde benzeri görülmemiş bir büyümeye yol açmıştır. Dünya çapında her yıl milyarlarca sağlık kaydı işlemi yapılmaya başlanmıştır. Özellikle, sinir ağları, istatistiksel modelleme, evrimsel algoritmalar ve görselleştirme araçları gibi veri madenciliği algoritmalarındaki son gelişmeler, her türlü ham verinin üst düzey bilgiye dönüştürülmesini mümkün kılmıştır. Ancak asıl sorun, her yöntemin veri yapısı, şekli ve geçerliliği ile ilgili kendi yaklaşımına sahip olmasıdır. Bu sınırlama sınıflandırma sistemlerinin performansını etkiler. Sonuç olarak, hibrit bir veri madenciliği yaklaşımına duyulan ihtiyaç, veri madenciliği topluluğu tarafından kabul görmektedir ve bu konuda son yıllarda yapılan çalışma sayısı gittikçe artmaktadır. Hibrit veri madenciliği, her bir tekniğin gücünü kullanmak ve birbirlerinin zayıflıklarını telafi etmek için çeşitli veri madenciliği tekniklerinin etkili bir kombinasyonu olarak tanımlanmaktadır. Bu çalışmanın amacı, son teknoloji veri madenciliği algoritmalarını ve uygulamalarını sunmak ve tıbbi verilerin kümelenmesi ve sınıflandırılması için yeni bir hibrit veri madenciliği yaklaşımı önermektir. Ayrıca çalışmada, denetimli ve denetimsiz öğrenme yöntemlerinin dengeli ve dengesiz veri setlerinde, farklı örneklem büyüklüklerinde ve farklı değişkenler arası ilişkiler olması durumunda performans ölçütlerinin hesaplanması ve bu ölçütlerin hibrit modelden elde edilen ölçütler ile karşılaştırılması amaçlanmıştır.

Çalışmada çeşitli senaryolar baz alınarak üretilen simüle veri setleri ve UCI veri tabanından alınan hepatit ve meme kanseri veri setleri kullanılmıştır. Sık kullanılan ve veri setlerinde en iyi performansa sahip denetimli öğrenme algoritmalarından Karar Ağaçları, Destek Vektör Makinesi, Random Forest, Naive Bayes ve K-en yakın komşu'nun yanı sıra Lojistik Regresyon ve Yapay Sinir Ağları algoritmaları, denetimsiz öğrenme algoritmalarından ise K-ortalama kullanılmıştır. Ayrıca kullanılan denetimli ve denetimsiz öğrenme algoritmaları birleştirilerek hibrit modeller oluşturulmuştur.

Simüle veri setlerinde bağımsız değişkenler arası korelasyon ve örneklem büyüklüğü arttıkça MKK değerlerinin de arttığı görülmüştür. Ayrıca dengesiz veri setlerinde bağımsız değişkenler arası korelasyon arttıkça düşük örnekleme sahip gruba ait performans ölçütlerinde de gözle görülür bir artış gözlemlenmiştir. Gerçek veri setleri incelendiğinde de durumun benzer olduğu görülmektedir.

Anahtar Sözcükler: Hibrit Model, Veri Madenciliği, Performans Ölçütleri

SUMMARY

Hybrid Model Approach in Data Mining

In recent years, the amount of data collected in medical disciplines has been increasing. Advances in digital technology have led to an unprecedented growth in the size, complexity and amount of collected data in medical reports and related images. Every year, billions of health records are made worldwide. In particular, recent advances in data mining algorithms such as neural networks, statistical modeling, evolutionary algorithms, and visualization tools have made it possible to transform any raw data into high-level information. However, the main problem is that each method has its own approach to data structure, form and validity. This limitation affects the performance of classification systems. As a result, the need for a hybrid data mining approach is recognized by the data mining community, and the number of studies on this subject has been increasing in recent years. Hybrid data mining is defined as an effective combination of various data mining techniques to harness the power of each technique and compensate for each other's weaknesses. The purpose of this study is to present state-of-the-art data mining algorithms and applications and to propose a new hybrid data mining approach for clustering and classifying medical data. In addition, in the study, it was aimed to calculate performance criteria of supervised and unsupervised learning methods in balanced and unbalanced data sets, different sample sizes and in case of different relationships between variables and to compare these criteria with the criteria obtained from the hybrid model.

In the study, simulated data sets produced on the basis of various scenarios and hepatitis and breast cancer data sets obtained from the UCI database were used. From supervised learning algorithms Decision Trees, Support Vector Machine, Random Forest, Naive Bayes, K-nearest neighbor, Logistic Regression and Artificial Neural Networks algorithms were used and K-mean was used for unsupervised learning algorithms, which are frequently used and have best performance in data sets. In addition, hybrid models were created by combining the supervised and unsupervised learning algorithms used.

In simulated data sets, it was observed that as the correlation between independent variables and sample size increased, MCC values also increased. In addition, as the correlation between independent variables increased in unbalanced data sets, a noticeable increase was observed in the performance criteria of the group with low sampling. When the actual data sets are examined, it is seen that the situation is similar.

Keywords: Data mining, Hybrid Model, Performance Criterias

KAYNAKLAR

ALBON C (2018). Python Machine Learning Cookbook. 1st Ed. O'Reilly Media.

BAYAT P, MOTEVALLI ALMOUTI M (2016). Presenting a Hybrid Model for Early Diagnosis. *Int. J. of Comp. & Info. Tech.*, **4(3):** 63-69.

BIAU G, SCORNET E (2016). A random forest guided tour. *Test*, **25(2):** 197-227.

BIFET A, FRANK E, HOLMES G, PFAHRINGER B (2010). Accurate ensembles for data streams: Combining restricted hoeffding trees using stacking. *In Proceedings of 2nd asian conference on machine learning*, 225-240.

BREIMAN L (1996). Bagging predictors. *Machine Learning*, **24(2):** 123-140.

BREIMAN L (2001). Random forests. *Machine Learning*, **45(1):** 5-32.

CHICCO D, JURMAN G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, **21(1)**: 6.

DANGETI P (2017). Statistics for Machine Learning. Packt Publishing Ltd.

DAVIS J, GOADRICH M (2006). The relationship between Precision-Recall and ROC curves. *In Proceedings of the 23rd international conference on Machine learning*, p. 233-240.

EL-RASHIDY MA, TAHA TE, AYAD NM, SROOR HS (2011). An Efficient Hybrid Data Mining Approach for Breast Tumors Diagnosis. *International Journal of Computer Information Systems*, **3(4):** 61-69.

FERNANDEZ A, GARCIA S, GALAR M, PRATI RC, KRAWCZYK B, HERRERA F (2018). Learning from Imbalanced Data Sets. Berlin: Springer. p.: 1-377

GELBUKH A, REYES-GARCIA CA (2006). Advances in Artificial Intelligence. Springer-Verlag Berlin/Heidelberg.

HA SH, JOO SH (2010). A hybrid data mining method for the medical classification of chest pain. *International Journal of Computer and Information Engineering*, **4(1):** 33-38.

HUANG TM, KECMAN V, KOPRIVA I (2006). Kernel Based Algorithms for Mining Huge Data Sets. Heidelberg: Springer.

KANTARDZIC, M (2011). Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons.

KOHAVI R (1995). The power of decision tables: In European conference on machine learning. Springer. p.: 174-189.

OLSON DL (2018). Data Mining Models. 2nd Ed. Business Expert Press.

QUINLAN JR (1986). Induction of decision trees. *Machine Learning*, **1(1):** 81-106.

RATHI M, PAREEK V (2016). Disease prediction tool: an integrated hybrid data mining approach for healthcare. *IRACST Int J Comput Sci Inf Technol Secur.*, **6(6):** 32-40.

SAITO T, REHMSMEIER M (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS One*, **10(3)**: e0118432.

SAITO T, REHMSMEIER M (2017). Precrec: fast and accurate precision–recall and ROC curve calculations in R. *Bioinformatics*, **33(1):** 145-147.

SIMSEK S, KURSUNCU U, KIBIS E, ANISABDELLATIF M, DAG A (2020). A hybrid data mining approach for identifying the temporal effects of variables associated with breast cancer survival. *Expert Systems with Applications*, **139:** 1-13.

SYED MR, SYED SN. (2008). Handbook of Research on Modern Systems Analysis and Design Technologies and Applications. IGI Global.

VERMA L, SRIVASTAVA S, NEGI PC (2016). A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. *Journal of medical systems*, **40(7):** 178-184.

WANG W, RICHARDS G, REA S (2006). Hybrid data mining ensemble for predicting osteoporosis risk. *In 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, p.: 886-889.

WU X, KOTAGIRI R, KORB KB (2006). Research and Development in Knowledge Discovery and Data Mining: Second Pacific-Asia Conference, PAKDD'98, Melbourne, Australia. Springer.

ÖZGEÇMİŞ

I- Bireysel Bilgiler

Adı	: Batuhan
Soyadı	: BAKIRARAR
Doğum yeri ve tarihi	:
Uyruğu	: T.C.
Medeni durumu	: Evli
Askerlik durumu	: Yaptı
E-posta	:
Telefon	:

II- Eğitimi

Doktora	: Ankara Üniversitesi, Tıp Fakültesi, Biyoistatistik AD, Mezuniyet yılı: -
Yüksek Lisans	: Ankara Üniversitesi, Tıp Fakültesi, Biyoistatistik AD, Mezuniyet yılı: 2016
Lisans	: İstanbul Kültür Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği, Mezuniyet yılı: 2011
Lise	: Antalya Çağlayan Lisesi, Fen Bilimleri, Hazırlık + 3 yıl, Mezuniyet yılı: 2005
Yabancı Dil	: İngilizce – KPDS : 76,25

III- Ünvanları

Araştırma Görevlisi

IV- Mesleki Deneyimi

2012-2013	: Kütahya Porselen Bilgi İşlem Merkezi, Bilgisayar Mühendisi

2013-2015 : Talya Bilişim, Bilgisayar Mühendisi

2015- : Ankara Üniversitesi, Tıp Fakültesi, Biyoistatistik AD, Araştırma Görevlisi

V- Üye Olduğu Bilimsel Kuruluşlar

Biyoistatistik Derneği

VI- Bilimsel İlgi Alanları

Uluslararası Yayınlar

1. Altınboğa Orhan, Yakıştıran Betül, Erol Seyit Ahmet, Oğuz Yüksel, Bakirarar Batuhan, Gülerman Cavidan, Çelen Şevki, Çağlar Ali Turhan (2020). Selection of a Correct Treatment Protocol in Caesarean Scar Pregnancies. Archives of Gynecology And Obstetrics, 302(6), 1375-1380., Doi: 10.1007/S00404-020-05800-2
2. Kalem Ziya, Namli Kalem Müberra, Bakirarar Batuhan, Kent Erkin, Makrigiannakis Antonios, Gürgan Timur (2020). Intrauterine G-Csf Administration In Recurrent Implantation Failure (Rif): An Rct. Scientific Reports, 10(1), Doi: 10.1038/S41598-020-61955-7
3. Yüksel Cemil, Erşen Ogün, Mercan Ümit, Başçeken Salim İlksen, Bakırarar Batuhan, Bayar Sancar, Ünal Ali Ekrem, Demirci Salim (2020). Long-Term Results and Current Problems in Laparoscopic Gastrectomy: Single-Center Experience. Journal of Laparoendoscopic Advanced Surgical Techniques, 30(11), 1204-1214., Doi: 10.1089/Lap.2020.0180
4. Tezcan Aydemir Sabiha, Kuzu Kumcu Müge, Ulukan Cagrı, Bakırarar Batuhan, Akbostancı Muhittin Cenk (2020). Patient Preference of Device-Based Treatment of Parkinson's Disease. International Journal of Neuroscience, Doi: 10.1080/00207454.2020.1853723
5. Kurt Mehmet Hakan, Bozkurt Poyzan, Görürgöz Cansu, Bakırarar Batuhan, Orhan Kaan (2020). Accuracy of Using Different Voxel Sizes to Detect Osseous Defects in Mandıbular Condyle. Journal of Stomatology, 73(5), 217-224.
6. Gürsoy Çoruh Ayşegül, Uzun Çağlar, Bozca Esra, Bozca Büşra, Demir İhsan Batuhan, Atasever Hilal Gülseren, Göçtürk Berna, Bakırarar Batuhan, Akyol Cihangir(2020). Is It Possible to Predict the Side of Hepatic Metastases According to The Primary Location of Colorectal Cancer?. Polish Journal of Radiology, 85(1), 595-599.
7. Kuzu Kumcu Müge, Bakırarar Batuhan, Yücesan Canan (2020). Quality of Life in Neuro- Behçet's Disease: A Cross-Sectional Study. Neurological Sciences, Doi: 10.1007/S10072-020-04630-Z
8. Yüksel Cemil, Çulcu Serdar, Bakırarar Batuhan, Doğan Lütfi (2020). The Effect of Breast Parenchymal Density on Breast Cancer Subtypes and Prognostic Factors. Journal of Contemporary Medical Science, 6(4), 181-186.
9. Isler Sila Cagri, Soysal Fatma, Akca Gülcin, Bakırarar Batuhan, Ozcan Gonen, Ünsal Fatma Berrin (2020). The Effects of Decontamination Methods of

Dental İmplant Surface on Cytokine Expression Analysis in the Reconstructive Surgical Treatment of Peri-İmplantitis. Odontology, Doi: 10.1007/S10266-020-00520-0

10. Günay Fatih, Cullas Ilarslan Nisa, Bakar Ates Filiz, Deniz Kiymet, Kadioglu Yusuf Kagan, Kiran Sibel, Bakırarar Batuhan, Cobanoglu Nazan (2020). Evaluation of Hair Cotinine and Toxic Metal Levels in Children Who Were Exposed to Tobacco Smoke. Pediatric Pulmonology, 55(4), 1012- 1019., Doi: 10.1002/Ppul.24692

11. Turgut Çankaya Zeynep, Ünsal Fatma Berrin, Gürbüz Sühan, Bakırarar Batuhan, Tamam Evşen (2020). Efficiency of Concentrated Growth Factor in the Surgical Treatment of Multiple Adjacent Papillary Losses: A Randomized, Controlled, Examiner-Blinded Clinical Trial Using Cad/Cam. International Journal of Periodontics Restoratıie Dentistry, 40(2), 73- 83., Doi: 10.11607/Prd.4359

12. Turgut Çankaya Zeynep, Gürbüz Sühan, Bakırarar Batuhan, Ünsal Fatma Berrin, Kurtiş Mazlum Bülent (2020). Evaluation of The Effect of the Application of Hyaluronic Acid Following Laser-Assisted Frenectomy: An Examiner-Blind, Randomized, Controlled Clinical Study. Quıntessence Internatıonal, 51(3), 188-201., Doi: 10.3290/J.Qi.A43667

13. Turgut Çankaya Zeynep, Gürbüz Sühan, Bakırarar Batuhan, Kurtiş Mazlum Bülent (2020). Evaluation of The Effect of Hyaluronic Acid Application On The Vascularization of Free Gingival Graft for Both Donor and Recipient Sites with Laser Doppler Flowmetry: A Randomized, Examiner-Blinded, Controlled Clinical Trial. International Journal of Periodontics Restorative Dentistry, 40(2), 233-243., Doi: 10.11607/Prd.4494

14. Yakar Fatih, Egemen Emrah, Dere Umit Akin, Celtikci Emrah, Dogruel Yücel, Sahinoglu Defne, Cuneyit Ibrahim, Bakırarar Batuhan, Adiguzel Esat, Coskun Erdal (2020). Subdural Hematoma Evacuation Via Rigid Endoscopy System. Journal of Craniofacial Surgery, Doi: 10.1097/Scs.0000000000007031

15. Coskun Erdal, Yakar Fatih, Baykara Eyup, Civlan Serkan, Bakırarar Batuhan, Egemen Emrah (2020). Experience of Fully Awake Craniotomy for Supratentorial Lesions: A Single-Institution Study. Turkish Neurosurgery, 30(6), 907-913., Doi: 10.5137/1019-5149.Jtn.30747-20.2

16. Candeniz Şeyda, Çıtaker Seyit, Bakırarar Batuhan (2019). Cross-Cultural Adaptation, Reliability and Validity of The Turkish Version of the Neck Outcome Score. Turkish Journal of Medical Sciences, Doi: 10.3906/Sag-1907-87

17. Şimşir Coşkun, Namlı Kalem Müberra, Kalem Ziya, Var Turgut, Bakırarar Batuhan, Coşkun Buğra, Coşkun Bora (2019). Circulating Spexin Levels in Pregnant Women with and without Gestational Diabetes. International Journal of Reproduction, Contraception, Obstetrics And Gynecology, 8(10), 4056-4061., Doi: 10.18203/2320-1770.İjrcog20194380

18. Kalem Ziya, Kaya Ellıbes Askı, Bakırarar Batuhan, Başbuğ Alper, Namlı Kalem Müberra (2019). An Optimal Uterine Closure Technique for Better Scarhealing and Avoiding Isthmocele in Cesarean Section: A Randomized Controlled Study. Journal of Investigative Surgery, Doi: 10.1080/08941939.2019.1610530

19. Kalem Ziya, Kaya Ellıbes Askı, Bakırarar Batuhan, Namlı Kalem Müberra (2019). Fetal Nuchal Translucency: Is There an Association with Birthweight and Neonatal Wellbeing?. Journal of Turkish Society of Obstetric and Gynecology, 16(1), 35-40., Doi: 10.4274/Tjod.Galenos.2019.21384

20. Kalem Ziya, Namlı Kalem Müberra, Akgün Nilüfer, Kaya Ellıbes Askı, Bakırarar Batuhan, Aydın S (2019). The Relationship Between The Levels of Anti-Müllerian Hormone, Vaspin, Visfatin, and The Patterns of Nutrition and Menstruation Innon-Polycystic Ovary Syndrome and Non- Obese Young Women. Clinical And Experimental Obstetrics Gynecology

21. Kalem Ziya, Namlı Kalem Müberra, Bakırarar Batuhan, Kent Erkin, Gürgan Timur (2018). Natural Cycle Versus Hormone Replacement Therapy Cycle in Frozen-Thawed Embryo Transfer. Saudi Medical Journal, 39(11), 1102-1108., Doi: 10.15537/Smj.2018.11.23299

22. Kalem Ziya, Namlı Kalem Müberra, Anadol Elvan, Bakırarar Batuhan, Yılmaz Canan, Elmas Çiğdem, Yalçınkaya Perihan, Ruso Halil, Gürgan Timur (2018). Maternal Nutrition and Reproductive Functions of Female and Male Offspring. Reproductıon, 156(4), 353-364., Doi: 10.1530/Rep-18-0070

23. Isler Sila Cagri, Soysal Fatma, Ceyhanlı Tuğçe, Bakırarar Batuhan, Ünsal Fatma Berrin (2018). Regenerative Surgical Treatment of Peri-Implantitis Using Either a Collagen Membrane or Concentrated Growth Factor: A 12-Month Randomized Clinical Trial. Clinical Implant Dentistry And Related Research, 20(5), 703-712., Doi: 10.1111/Cid.12661

24. Doğan İhsan, Bozkurt Melih, Kahiloğulları Gökmen, Yakar Fatih, Zaimoğlu Murat, Bakırarar Batuhan, Kırcıl Cihan, Eroğlu Ümit, Özgüral Onur, Ucer Melih, Kılınç Mustafa Cemil, Demirel Altan, Güner Efe, Uğur Hasan Çağlar, Çağlar Yusuf Şükrü (2018). Is a Unilateral Surgical Approach Effective in Patients with Bilateral Leg Pain with Unilateral Lumbar Disc Herniation? A Prospective Nonrandomized Clinical and Surgical Study. World Neurosurgery, 117, 316-322., Doi: 10.1016/J.Wneu.2018.06.022

25. Kalem Ziya, Namlı Kalem Müberra, Ruso Halil, Bakırarar Batuhan, Gürgan Timur (2018).Fresh Versus Frozen-Thawed Blastocyst Transfer in High Responders. Ginekologia Polska, 89(8), 407-413., Doi: 10.5603/Gp.A2018.0070

26. Kalem Ziya, Namlı Kalem Müberra, Seval Murat, Bakırarar Batuhan, Şimşir Coşkun, Yılmaz Canan, Atabekoğlu Cem Somer, Gürgan Timur (2018). Serum and Follicular Fluid Concentration of Stem Cell Factor in PCOS. International Journal of Reproduction, Contraception, Obstetrics and Gynecology, 7(9), 3467-3472., Doi: 10.18203/2320-1770.İjrcog20180001

27. Namlı Kalem Müberra, Kalem Ziya, Bakırarar Batuhan, Ergün Ali, Gürgan Timur (2018). The Effect of Progesterone Use in the First Trimester on Fetal Nuchal Translucency. Journal of The Turkish-German Gynecological Association(19), 29-33., Doi: 10.4274/Jtgga.2017.0056

28. Engin-Ustun Yaprak, Yılmaz Nafiye, Akgun Nilüfer, Aktulay Ayla, Tuzluoğlu Ahmet Deniz, Bakırarar Batuhan (2018). Body Mass Index Effects Kruger's Criteria in Infertile Men. International Journal of Fertility Sterility, 11(4), 258-262., Doi: 10.22074/İjfs.2018.4888

29. Sorgun Mine Hayriye, Kuzu Müge, İnci Şule Özer, Yilmaz Volkan, Ulukan Cagrı, Cotur Levent Hafize, Rzayev Sefer, Rawandi Anwar, Bakırarar Batuhan, Togay Işıkay Canan (2018). Risk Factors, Biomarkers, Etiology, Outcome and Prognosis of Ischemic Stroke in Cancer Patients. Asian Pacific Journal of Cancer Prevention(19), 649-653., Doi: 10.22034/Apjcp.2018.19.3.649

30. Gulsahi Ayse, Kulah Cemal Kivanc, Bakırarar Batuhan, Gulen Orhan, Kamburoglu Kivanc (2018). Age Estimation Based on Pulp/Tooth Volume Ratio

Measured on Cone-Beam CT Images. Dentomaxillofacial Radiology, Doi: 10.1259/Dmfr.20170239

31. Namlı Kalem Müberra, Kalem Ziya, Yüce Tuncay, Bakırarar Batuhan, Söylemez Feride (2017). Comparison of Melatonin Levels in The Colostrum Between Vaginal Delivery and Cesarean Delivery. American Journal of Perinatology, Doi: 10.1055/S-0037-1608681

32. Namli Kalem Müberra, Kalem Ziya, Bakırarar Batuhan, Demircan Kadir (2017). Adamts 1, 4, 5, 8, and 9 in Early Pregnancies. Fetal and Pediatric Pathology

33. Namlı Kalem Müberra, Akgün Nilüfer, Kalem Ziya, Bakırarar Batuhan, Çelik Tuğrul (2017). Chemokine (C-C Motif) Ligand-2 (Ccl2) and Oxidative Stress Markers in Recurrent Pregnancy Loss and Repeated Implantation Failure. Journal of Assisted Reproduction and Genetics, Doi: 10.1007/S10815-017-0992-5

34. Kalem Muberra Namlı, Kalem Ziya, Akgun Nilufer, Bakırarar Batuhan (2017). The Relationship Between Postmenopausal Women's Sclerostin Levels and Their Bone Density, Age, Body Mass Index, Hormonal Status, and Smoking and Consumption of Coffee and Dairy Products. Archives of Gynecology and Obstetrics, 295(3), 785-793., Doi: 10.1007/S00404-017-4288-X

35. Akgun Nilufer, Keskin Esra, Kalem Muberra Namlı, Bakırarar Batuhan (2017). Intrauterine Levobupivacaine for Pain Control During Intrauterine Device Insertion. International Journal of Reproduction, Contraception, Obstetrics and Gynecology, 6(3), 1117, Doi: 10.18203/2320- 1770.İjrcog20170596

Ulusal Yayınlar

1. Kalem Ziya, Şimşir Çoşkun, Bakırarar Batuhan, Namlı Kalem Müberra (2020). The additional diagnostic value of NLR and PLR for CA-125 in the differential diagnosis of endometrioma and benign ovarian cysts in women of reproductive age: a retrospective case- control study. The European Research Journal, 6(2), 111-119., Doi: 10.18621/eurj.483442

2. Bakırarar Batuhan, Kar İrem, Gökmen Derya, Elhan Atilla Halil, Genç Volkan (2019).The Prediction of Breast Biopsy Outcomes Using Two Data Mining Algorithms Based on Parameter Variations. Turkiye Klinikleri Journal of Biostatistics, 11(2), 133-142., Doi: 10.5336/biostatic.2019-64754

3. Isler Sila Cagri, Uraz Ahu, Şengül Janset, Çakıroğlu Miray, Bakırarar Batuhan, Çetiner Fitnat Deniz (2019). Evalution of the Patiens Oral Health Related Quality of Life After Harvesting Free Gingival Graft. Cumhuriyet Dental Journal, 22(1), 11-22., Doi: 10.7126/cumudj.452909

4. Nalcı Hilal, Gündüz Ahmet Kaan, Tüntaş Bilen Feyza, Köse Serdal Kenan, Bakırarar Batuhan (2018). Konjonktiva Malign Melanomlarında Klinik AJCC Sınıflandırmasına Göre Tedavi Sonuçlarının Değerlendirilmesi. Turkiye Klinikleri Journal of Ophthalmology, 27(2), 114-122., Doi: 10.5336/ophthal.2017-56132

5. Şimşir Coşkun, Kalem Ziya, Bakırarar Batuhan, Namlı Kalem Müberra (2018). Comparison of pregnancy and neonatal outcome in spontaneous and invitro fertilization (IVF) twin pregnancies. Journal of Harran University Medical Faculty, 15(3), 250-254

6. Şimşir Coşkun, Kalem Ziya, Bakırarar Batuhan, Namlı Kalem Müberra (2018). Comparison of pregnancy and neonatal outcome in spontaneous and invitro

fertilization (IVF) twin pregnancies. Harran Üniversitesi Tıp Fakültesi Dergisi (Journal of Harran University Medical Faculty), 15(3), 250-254

7. Bulmuş Tüccar Tuğçe, Bakırarar Batuhan, Köksal Eda (2017). Pre ve Postmenopozal Kadınlarda Besin Tüketim DurumuFRAX® Kırık Riski ile İlişkili midir?. Beslenme ve Diyet Dergisi, 45(2), 116-127

Ulusal ve Uluslararası Kongrelerde Sunulan Bildiriler ve Posterler

1. Isler Sila Cagri, Soysal Fatma, Ceyhanlı Tuğçe, Bakırarar Batuhan, Ünsal Fatma Berrin (2020). Reconstructive Surgical Therapy of Periimplantitis: 3 Year Results of a Randomized Clinical Trial. 29th Annual Scientific Meeting of The European Association for Osseointegration, 31(S20), 35-35., Doi: 10.1111/Clr.30_13643
2. Ünal Ali Ekrem, Yüksel Cemil, Başçeken Salim İlksen, Erşen Ogün, Mercan Ümit, Bakırarar Batuhan (2019). Cytoreductive Surgery and Hyperthermic Intraperitoneal Chemotherapy in Patients with Appendix Cancer. 39th Congress of The European Society of Surgical Oncology, 46(2), 104-105., Doi: Https://Doi.Org/10.1016/J.Ejso.2019.11.260
3. Yüksel Cemil, Basçeken Salim İlksen, Mercan Ümit, Erşen Ogün, Yalkın Ömer, Aydın Ferit, Bakırarar Batuhan, Bayar Sancar, Ünal Ali Ekrem, Demirci Salim (2019). Safety and Efficacy of Laparoscopic Gastrectomy in Our Series. 39th Congress of The European Society of Surgical Oncology, 46(2), 151-152., Doi: https://Doi.Org/10.1016/J.Ejso.2019.11.399
4. İşler Sıla Çağrı, Çetiner Fitnat Deniz, Bakırarar Batuhan, Uraz Ahu (2019). A Prediction Model for Diagnosing Peri-Implant Health and Disease. 28th Annual Scientific Meeting of the European Association for Osseointegration, 30(S19), 20-20., Doi: 10.1111/Clr.32_13508
5. Coşarderelioğlu Seçer Çağlar, Uçkun Özkan Ayşegül, Bahşi Remzi, Selvi Öztorun Hande, Atmış Volkan, Bakırarar Batuhan, Yalçın Ahmet, Aras Sevgi, Varlı Murat (2019). Asymptomatic Vertebral Fracture Risk Assessment with FRAX Plushandgrip Strength: A New Screening Algorithm. 15th International Congress of The Eugms Krakow/Poland, 10(S1), 1-325., Doi: 10.1007/S41999-019-00221-0
6. Duru Birgi Sümerya, Akyürek Serap, Gümüştepe Esra, Arslan Yakup, Bakırarar Batuhan, Çakır Gökçe Şaban (2019). Dosimetric Investigation of Radiation Induced Trigeminal Nerve Toxicity in Patients with Parotid Tumor. International Journal of Radiation Oncology Biology Physics, 105(1), 171-171
7. Duru Birgi Sümerya, Akyürek Serap, Gümüştepe Esra, Bakırarar Batuhan, Gökçe Şaban Çakır (2019). Results of Definitive Treatment in Patients with Synchronous Oligometastatic Non Small Cell Lung Cancer. Ulusal Kanser Kongresi, 14(10), 996, Doi: 10.1016/J.Jtho.2019.08.2192
8. Atasever Melahat, Erişgin Züleyha, Sönmez Ciğdem, Özer Çiğdem, Coşkun Bora, Arıkan Murat, Bakırarar Batuhan, Çetinkaya Kadir (2018). Sıçanlarda Karboplatin ile İndüklenen Gonadotoksisitemodelinde Nigella Sativanın Over Rezervi Üzerine Etkileri. 1st International Gynecologic Oncology Congress. November 21st - 25th, 201, 109
9. Bakırarar Batuhan, Kar İrem, Doğanay Erdoğan Beyza, Elhan Atilla Halil (2018). Eksik Veri Yerine Değer Atama Yöntemlerinin Performanslarının Veri

Madenciliği Sınıflama Algoritması Kullanılarak Değerlendirilmesi. 20. Ulusal 3. Uluslararası Biyoistatistik Kongresi

10. Öztürk Ebru, Bakırarar Batuhan, Karahan Sevilay, Karaağaoğlu Ahmet Ergun (2018). An Extensive Web Interface for Validity and Reliability with RServe. 20. Ulusal ve 3. Uluslararası Biyoistatistik Kongresi

11. Kar İrem, Bakırarar Batuhan, Elhan Atilla Halil, Köse Serdal Kenan (2018). Çok Yüzeyli Rasch Modeli ve Bir Uygulaması. 20. Ulusal ve 3. Uluslararası Biyoistatistik Kongresi

12. Ünsal Fatma Berrin, Turgut Çankaya Zeynep, Gürbüz Sühan, Tamam Evşen, Bakırarar Batuhan (2018). Çolku Papil Kayıplarının Cerrahi Tedavisinde Konsantre Büyüme Faktörü Etkinliğin Değerlendirilmesi Kontrollu Tek Kör Klinik Çalışma. Türk Periodontoloji Derneği 48. Bilimsel Kongresi

13. Turgut Çankaya Zeynep, Gürbüz Sühan, Bakırarar Batuhan, Ünsal Fatma Berrin, Tamam Evşen (2018). Çoklu Papil Kayıplarının Cerrahi Tedavisinde Konsantre Büyüme Faktörü Etkinliğinin Değerlendirilmesi: Kontrollü Tek-Kör Klinik Çalışma. Türk Periodontoloji Derneği 48. Uluslararası Bilimsel Kongresi

14. Bakırarar Batuhan, Kar İrem, Elhan Atilla Halil (2018). Rserve ile Kümeleme Algoritmalarına ait Web Tabanlı Yazılım Geliştirilmesi. 20. Ulusal ve 3. Uluslararası Biyoistatistik Kongresi

15. Kuzu Müge, Tezcan Sabiha, Ulukan Cagrı, Bakırarar Batuhan, Akbostancı Muhittin Cenk (2018). Patient's Choice of Device-Based Treatments in Parkinson's Disease. 2018 International Congress of Parkinson's Disease and Movement Disorder.

16. Büyükkıdık Serap, Bakırarar Batuhan, Bulut Okan (2018). Comparing the Performance of Data Mining Methods in Classifying Successful Students with Scientific Literacy in PISA 2015. 6th International Congress on Measurement and Evaluation in Education and Psychology

17. Kar İrem, Bakırarar Batuhan, Doğanay Erdoğan Beyza, Gökmen Derya, Köse Serdal Kenan, Elhan Atilla Halil (2017). A Comparison of Maximum Likelihood and Expected a Posteriori Estimation in Computerized Adaptive Testing. 10th International Statistics Congress

18. Bakırarar Batuhan, Kar İrem, Gökmen Derya, Doğanay Erdoğan Beyza, Elhan Atilla Halil (2017). Demonstration of a Computerized Adaptive Testing Application Over a Simulated Data. 10th International Statistics Congress

19. Bakırarar Batuhan, Kar İrem, Elhan Atilla Halil (2017). LMS Yönteminin Uyum İyiliğinin Ölçülmesinde Kullanılan Yöntemlerin Karşılaştırılması. 19. Ulusal Biyoistatistik Kongresi Ve 2. Uluslararası Biyoistatistik Kongresi

20. Kar İrem, Bakırarar Batuhan, Köse Serdal Kenan (2017). Bulanık Çıkarsama Sistemleri ile Veri Madenciliği Yöntemlerinin Sınıflama Performansının Benzetim Çalışması ile Karşılaştırılması. 19. Ulusal Biyoistatistik Kongresi ve 2. Uluslararası Biyoistatistik Kongresi

21. Bakırarar Batuhan, Kar İrem, Elhan Atilla Halil (2017). Rserve ile Veri Madenciliği Algoritmalarının Web Uygulaması. 19. Ulusal Biyoistatistik Kongresi ve 2. Uluslararası Biyoistatistik Kongresi

22. Köse Serdal Kenan, Kar İrem, Bakırarar Batuhan (2017). Bulanık Çıkarsama Sistemleri ile Lojistik Regresyon Sınıflandırıcılarının Performanslarının Apandisit Verisinde Karşılaştırılması. 19. Ulusal Biyoistatistik Kongresi ve 2. Uluslararası Biyoistatistik Kongresi

23. Namlı Kalem Müberra, Kalem Ziya, Akgün Nilüfer, Bakırarar Batuhan (2017). The Relatıonshıp Between The Levels of Anti-Müllerian Hormone, Vaspın and Visfatin and The Patterns of Nutrition and Menstruation Innon-Polycystic Ovary Syndrome and Non-Obese Young Women. 19th World Congress on Ivf in Conjunction with 6th Society of Reproductive Medicine And Surgery Congress, 193-193

24. Kalem Namlı Müberra, Kalem Ziya, Anadol Elvan, Yılmaz Canan, Elmas Çiğdem, Yalçınkaya Perihan, Ruso Halil, Bakırarar Batuhan, Gürgan Timur (2017). The Effects of Cafeteria Diet Induced Maternal Obesity on Off-Springs Fertility. 19th World Congress on Ivf in Conjunction with 6th Society of Reproductive Medicine and Surgery Congress, 68-68

25. Bakırarar B**,** Kar İ, Yavuz Y, Köse SK, "Veri Madenciliği'ne Genel Bakış ve Çok Katmanlı Algılayıcı Yönteminin WEKA Programında İncelenmesi: Sağlık Alanında Bir Uygulama", XVII. Ulusal Biyoistatistik Kongresi, Girne, Kıbrıs, 05-09 Kasım 2015

26. Kar İ, Bakırarar B, Köse SK, "İki Sürekli Değişken Arasındaki Uyum Araştırmalarında En Uygun Kesim Noktasının Tespiti: ROC Analizi Mi, Tek Değişkenli Lojistik Regresyon Analizi Mi?", XVII. Ulusal Biyoistatistik Kongresi, Girne, Kıbrıs, 05-09 Kasım 2015

27. Bakırarar B, Kar İ, Yavuz Y, "Büyük Veriye Genel Bakış: Sağlık Alanında Örnek Uygulama", XVIII. Ulusal Biyoistatistik Kongresi ve I. Uluslararası Biyoistatistik Kongresi, Beldibi, Antalya, 26-29 Ekim 2016

28. Kar İ, Bakırarar B, Köse SK, "Veri Madenciliği Sınıflandırıcılarının Performanslarının Karşılaştırılması", XVIII. Ulusal Biyoistatistik Kongresi ve I. Uluslararası Biyoistatistik Kongresi, Beldibi, Antalya, 26-29 Ekim 2016

Seminerler

1. "Sağlık Alanında Büyük Veri Uygulamaları", Ankara Üniversitesi, Ekim 2016
2. "Veri Madenciliği ve Uygulama Alanları", Ankara Üniversitesi, Nisan 2018