YAPAY ZEKÂ

SES İLE CİNSİYET TANIMA VERİ SETİNİN KARAR AĞACI, NAİVE BAYES, K-EN YAKIN KOMŞU VE YAPAY SİNİR AĞLARI-ÇOK KATMANLI ALGILAYICI İLE TAHMİN EDİLMESİ

ÖĞRETİM GÖREVLİSİ

MUHAMMET HANEFİ CALP

MUKADDER YILMAZ 368714

ÖZET

Yapay Zeka bilgisayar ve robot gibi insan yapımı araçlar kullanarak insanlar ve hayvanlar gibi doğal sistemleri taklit etmekle alakalıdır. Bu yöntem, bilginin - özellikle de kesin olmayan belirsiz bilginin- bilgisayar hafızasında depolanabilmesi ve bu bilgiden otomatik olarak çıkarımlar yapılabilmesi amacıyla bilginin nasıl temsil edilebileceğini anlamayı içermektedir. Ayrıca depolanan bilgiyi esas alarak kararların nasıl yapılabileceği ve eylem planlarının nasıl oluşturulabileceği ve örnek veriden öğrenerek veya insan uzmanları sorgulayarak bilgisayarda işlenebilir bilginin nasıl edinebileceğini anlamayı da içermektedir.

Bu projede Ses ve Konuşma Analizi ile Cinsiyet Tanıma(voice.csv)isimli bir veri seti kullanılmıştır. Bu veri seti sesin ve konuşmanın akustik özelliklerine göre sesi erkek veya kadın olarak tanımlamak için oluşturulmuştur. Veri setindeki verileri en doğru şekilde tahmin edilmesi için K-En Yakın Komşu, Naive Bayes, Karar Ağacı ve Yapay Sinir Ağları-Çok Katmanlı Algılayıcı sınırlandırma yöntemleri kullanılmıştır. Tüm yöntemlerde "Meanfun, Sd, Q25, IQR, Label" parametreleri temel alınarak ilerlenmiştir. Yöntemlerde makineye veri setinin sonucu bilinen değerlerin %60 eğitim %40 test, %25 eğitim %75 test ve %11 eğitim %89 test kümesi alınarak 3 ayrı model oluşturulmuştur. Yöntemler için gerekli olan hesaplamalar yapılmış, sonuçlar ve grafikler kullanılmıştır. Her modelin hem kendi içerisinde hem de modeller arasında kıyaslaması yapılıp gerekli çıkarımlar ortaya koyulmuştur. Uygulamanın detaylı açıklaması 3. Bölümde bulunmaktadır.

1.GİRİŞ

Günümüzde teknolojinin gelişmesi ve teknoloji okuryazarlığının artmasıyla birlikte bilgisayar artık sadece bilgi edinme gibi nedenlerle değil karar verme uygulamalarında da kullanılmaktadır. İşlem hızı, hiçbir detayı gözden kaçırmaması, karar verme mekanizmasındaki maliyeti düşürmesi ve matematiksel olarak çözülemeyen durumlarda çözüm sunması gibi sebeplerle Yapay Zeka uygulamaları her alanda yaygınlaşmaktadır. Yapay zeka, insan tarafından yapıldığında zeka olarak adlandırılan davranışların(akıllı davranışların) makine tarafından da yapılmasıdır. Yapay zekanın insan aklının nasıl çalıştığını gösteren bir kuram olduğu da söylenebilir. Yapay zekanın amacı, makinaları daha akıllı hale getirmek, zekanın ne olduğunu anlamak, insan zekasını bilgisayar aracılığı ile taklit etmek, bu anlamda belli bir ölçüde bilgisayarlara öğrenme yeteneği kazandırabilmektir.

‘Yapay Zeka’ kavramı ilk duyuşta ister akademisyen, öğretmen, öğrenci olsun ister işadamı olsun birçok kişi üzerinde merak uyandırmaktadır. Literatürde "Artificial Intelligence" olarak adlandırılan yapay zeka ilk bakışta herkese farklı bir şeyin çağrışımını yaptırmaktadır.[2] Kimilerine göre, yapay zeka kavramı, insanoğlunun yerini alan insan gibi elektromekanik bir robotu çağrıştırmaktadır. Fakat bu alanla ilgili olan herkes, insanoğlu ile makinalar arasında kesin bir farklılığın olduğu bilincindedir.

Bilgisayarlar mevcut teknoloji ile hiçbir zaman insanoğlunun yaratıcılık, duygu ve mizacının benzeşimini aktarabilme becerisine sahip olamayacağı görülmektedir. Bununla beraber, bilgisayarların belirli fiziksel insan davranışlarını yapan robotlar gibi makinalara yön vermesi ve veri hesaplaması, tıbbi teşhis gibi belirli bir uzmanlık alanı ile ilgili beşeri düşünme sürecinin benzeşimini yapan sistemlere beyin olma becerisine sahip olması mümkündür. Bu şekilde başarılı uygulamaları da mevcut olup ticari getirisi de olmuştur.

Bilgisayarlar, insan zekâsının bazı basit fonksiyonlarını daha iyi yapabilirler; matematik hesapları yapabilirler, rakamları ve harfleri işleyebilirler, basit kararlar verebilirler ve değişik bilgi saklama ve erişim fonksiyonlarını yürütebilirler. Bu tür uygulamalarda bilgisayarlar oldukça iyidir ve genellikle de insanların performansını geçerler. Çünkü evlerde kullanılan bilgisayarlar bile 5-6 haneli iki sayının çarpa işlemini nano saniyeler seviyesinde yapmaktadır. Veya bir kitap içindeki belli bir kelimenin kaç tane olduğunu insanın algılamasının çok altında hemen çıkarmaktadır. Bilgisayarlar, insanın düşünme işlemlerini bazı yönlerini büyük ölçüde basitleştirir ve hızlandırır. Yapay zeka teknoloji uygulamaları ile elle yapılan kompleks işlemleri daha da genişletmemize yada otomatik olarak yapmamıza imkan sağlar. Ayrıca yapay zeka teknolojileri, diğer bilgisayar tabanlı bilgi sistemleri ile bütünleştirilerek bilgisayarların yetenekleri ve uygulanabilirlikleri hızla arttırılmaktadır.[1]

Yapay zekayı anlamak bilgisayarda klasik şekilde veri işlemek düşüncesinden uzaklaşmayı gerektirir. Burada söz konusu olan bilgisayarların programları ile klasik algoritmik işlemleri yapmasından öte daha can alıcı özelliklerle ortaya çıkmasıdır. Şöyle ki, bir bilgisayarınız var ama klavyesi yok, dolayısıyla konuştuğunuzu anlıyor, ona göre ilgili komutları yerine getiriyor ve cevap veriyor. Tabiî ki sonucu size kullanıcı uyumlu, sesli ve/veya zengin içerikli grafik ekran gibi çıkış birimlerinden sunuyor. Daha ileri giderek, bilgisayardan tanımladığınız işlevleri yerine getirecek program üretmesini sağlayabiliyorsunuz. Evet, bunlar yakın zamana kadar her biri birer hayal olmaktan öteye gitmeyen düşüncelerdi. Fakat bunların tümü olmasa da kısmen gerçekleştiğini görmekteyiz. Bu alandaki gelişmenin bilgisayar teknolojisinin gelişmesi ve kullanımdaki yaygınlığının artmasının önemli rolü bulunmaktadır.

"Bilgi" çağımıza damgasını vuran bir terim olmaktan da öte, yeni bir teknoloji olma yolunda hızla ilerlemektedir. Genellikle bilgi bilgisayarda depolanmış bir varlık gibi düşünülür. Bilgi veritabanı uygulamalarında olduğu gibi yapısal olduğu sürece depolanması ve kullanılması bilgisayar ortamında kolaydır. Ayrıca bilgi en alt seviyede bir veri olarak değerlendirilmesi söz konusu olduğu gibi yol yöntemleri tanımlayan bilgilere ve bu bilgilerin değerlendirilmesini sağlayacak bilgilere (bilgi hakkında bilgilere) de ihtiyaç vardır. Diğer yandan Bilginin en etkin bir şekilde depolanıp saklanması ve ihtiyaç duyulduğunda yüksek performanstaki bir hızla bulunup getirilmesi, yeni yöntemler geliştirilmesini gerektirmektedir. Verileri depolama ve işleme araçlarının sayısı oldukça artmıştır. Bu artış beraberinde ilk yıllarda depolama ortamlarının yetersizliğini gündeme getirse de yaklaşık olarak her yıl defalarca katlanan depolama kapasitesi sayesinde bu sorun büyük ölçüde giderilmiştir. Ancak bu defa da depolanan büyük verilerin analiz edilmesi ve faydalı bilginin ortaya çıkarılması aşamasında, verinin büyüklüğü ve dolayısı ile analizi gerçekleştiren mikroişlemci ve RAM bellek sorunları ortaya çıkmıştır. Özellikle internetin her gün devasa veriler ürettiği günümüzde bu sorun daha da belirgin bir hal almıştır. Büyük veri kümelerinin analiz edilerek işe yarar bilginin ortaya çıkarılması ve bu verileri kullanan otomatik veya yarı otomatik sistemlerin tasarlanması veri madenciliği ve yapay zeka bilim dallarının başlıca uğraşı içerisinde yer almaktadır. Yapay zeka ve veri madenciliği teknikleri büyük verilerden anlamlı ve faydalı bilgiler elde etmeyi amaçlayan birbiri ile derin ilişkili iki disiplindir.[1] Büyük veriler üzerinde bu disiplinlerin ortaya koyduğu metodlar, araştırmacılar tarafından sıklıkla kullanılmaktadır. Metodları en doğru şekilde kullanmanın yolu hızlı, çalışma şekliyle insan beyninin çalışma anlayışıyla olabildiğince bütünleştirilmiş bilgisayarların yaratılması ile mümkün olabilecektir. Belki de klasik sayısal hesaplama yapabilen bilgisayar yapıları dışında bilgi işleyen, çıkarım yapan bilgisayar mimari yapıları doğabilecektir.

2.LİTERATÜR TARAMASI

Çalışmada çocuk işçiliğinin genel bir tanımı yapılarak çocuk işçiliğinin nedenleri belirtilmiş. Bu kapsamda gelişmiş ve gelişmekte olan ülkeler değerlendirilerek çocuk işçiliği ele alınmıştır. Bu çalışmanın amacı; karar ağacı algoritmalarından CART ve CHAID ile 114 ülke üzerinde çocuk işçiliğine etki eden faktörlerin önem sırasına göre belirlenmesidir. 114 ülkeye ait veriler Dünya Bankası'ndan elde edilmiştir. Bu ülkelere ait eksik veriler ILO, UNİCEF, OECD, TÜİK'den elde edilen veriler sayesinde tamamlanmıştır. Çalışmada büyük sayıdaki veri kümesi içerisinde gizlenmiş, geçerli ve kullanılabilen bilgileri ortaya koyma özelliği bulunan Veri

Madenciliği ele alınmıştır. VM tekniklerinden en sık kullanılan karar ağacı algoritması yardımıyla da çalışmanın uygulama kısmı gerçekleştirilmiştir. Veri setinin sırasıyla %70, %50 ve %30'luk kısmı dahil edilerek 3 farklı kritere göre ağaçlar oluşturulmuştur. Bunlar sırasıyla karşılaştırmalı olarak yorumlanmıştır. En uygun model %70-%30 oranlarıyla elde edilen model olmuştur. Çalışmanın bulgularına göre CART ve CHAID algoritmalarında en önemli değişken SGP olarak saptanmıştır. Sonuçlara göre ülkelerin satın alma gücü paritesi arttıkça çocuk işçiliği oranlarında azalma olduğu görülmüştür. Yapılan araştırma sonucunda literatürü doğrular nitelikte sonuçlar elde edilmiştir. Çocuk işçiliğini etkileyen en önemli değişkenler arasında SGP, GSYH ve yoksulluk oranı değişkenleri bulunmaktadır. Bu nedenle, öncelikli olarak ailelerin gelir seviyesini arttıracak ekonomik önlemlerin alınması önerilebilir. Geniş kitleler için uygulanacak sosyal politikalar çerçevesinde ekonomik tedbirler alınmalıdır. Maliye, gelir dağılımı, sanayileşme, istihdam, verimlilik gibi dallarda planlama yapılmalıdır[3].

Bu çalışmada, bir demir çelik fabrikasında yaşanan iş kazalarına ilişkin, belirli alt gruplara özgü olan ilişkilerin tanımlanması, vakaların yüksek, orta, düşük risk grupları gibi kategorilendirmesi ve gelecekteki olayların tahmin edilebilmesi için kurallar oluşturulması amaçlanmaktadır. Bu çalışmada, bir demir çelik fişletmesinde kazalanan 205 çalışana ait veri tabanı üzerinden iş kazalarına ilişkin, belirli alt gruplara özgü olan ilişkilerin tanımlanması, vakaların yüksek, orta, düşük risk grupları gibi kategorilendirmesi ve gelecekteki olayların tahmin edilebilmesi için kurallar oluşturulması amaçlanmaktadır. Çalışmada görsel, anlaşılır, basit yorumlanabilir ve kural çıkarımına imkân tanıması nedenleriyle veri madenciliği yöntemlerinden karar ağaçları tekniklerinden yararlanılmıştır. Karar ağacı algoritmaları, demir çelik sektöründeki iş kazası şiddeti tahmini için uygun modeller oluşturmuştur. Chaid, CRT ve C5.0 algoritma sonuçları incelendiğinde; işyerindeki iş kazalarının şiddetini etkileyen en önemli değişken "çalışma alanı" olarak ortaya çıkmıştır. Daha sonra bunu "kaza nedeni" ve "tecrübe" faktörleri izlemiştir[4].

Bu çalışmada veri madenciliği yöntemlerinden biri olan karar ağacı kullanılarak sağlık harcaması tahmini yapılmış ve sonuçlar analiz edilmiştir. Açık erişimli Kaggle veri bilimi depolama platformundan alınan veri kümesindeki yaş, cinsiyet, çocuk sayısı, vücut kitle indeksi, sigara kullanma ve bölge bilgileri karar ağacının giriş değerlerini oluşturmaktadır. Bu çalışmada sağlık harcamalarının tahmini için veri madenciliği yöntemlerinden Karar Ağacı kullanılmıştır. Bu çalışmada bireylere ait verilerden yola çıkılarak karar ağacı yöntemi ile kişiye ait sağlık harcaması tahmin edilmeye çalışılmış ve ortaya çıkan sonuçlar karşılaştırılmıştır. Çalışmanın veri madenciliği yöntemleri kullanarak sağlık harcamaları veya diğer verilerin

tahmini konusunda yapılacak araştırmalara Normalizasyon, yöntem ve parametre seçimi konularında yol gösterici olacağı düşünülmektedir[5].

Çalışmada, akciğer kanserinin erken tanısına katkıda bulunabilmek amaçlanmıştır. Genel olarak hastalara hastalık belirtileri doğrultusunda akciğer kanseri olup olmadıklarına dair teşhis konulmaktadır. Bu çalışma ile sağlık veritabanında mevcut olan, önceden teşhisi konulmuş vakaların anonim verileri kullanılarak, WEKA veri madenciliği yazılımında hangi algoritmanın bu alanda daha başarılı olabileceğine dair bir çalışma yapılmıştır. WEKA'da bulunan veri madenciliği algoritmaları arasından karşılaştırılacak algoritmalar seçilirken popülerlik ve literatürde benzer konuda yapılan çalışmalar dikkate alındıktan sonra on adet algoritma seçilmiş ve veri setine uygulanmıştır. Bu algoritmalar Naive Bayes, BayesNet, Lojistik Regresyon, Multilayer Perceptron, KStar, Bagging, OneR, ZeroR, J48 ve Random Tree algoritması şeklindedir. Bu çalışmada akciğer kanseri hastalığının teşhisinde fikir sunabilecek veri seti elde edilmiş ve bu veri setine WEKA veri madenciliği yazılımı ile çeşitli algoritmalar uygulanmıştır. Bu kapsamda personelin sisteme yanlış veya uç değer olarak girdiği verilerin teker teker kontrol edilip tek bir standart haline dönüştürülmesinden ve bütün önişleme süreçlerinin tamamlanmasından sonra açık kaynak kodlu veri madenciliği yazılımı olan WEKA ile veri seti üzerinde çeşitli algoritmalar uygulanarak modeller oluşturulmuş ve buna göre en başarılı algoritma olarak Naive Bayes algoritması bulunmuştur[6].

Çalışmanın temel amacı veri madenciliği tekniklerinin internet bankacılığı kullanıcı profilinin çıkarılması için kullanılmasıdır. Analizler sonucunda öncelikle internet bankacılığı kullanan banka müşterilerinin profilleri incelenmiş, Birliktelik Kuralları Analizi ile müşterilerin hangi bankaları tercih ettikleri, ne tür işlemler gerçekleştirdikleri, işlemleri gerçekleştirirken hangi kalite unsurlarına önem verdikleri belirlenmiştir. Veri setine Kümeleme Analizi de uygulanmıştır. Banka müşterileri rekabet ortamında çeşitli ürünleri, daha iyi servis ve daha uygun fırsatlarla kullanmak istemektedir. Bu durum sonucunda, bankaların pazarlama tekniklerini geliştirmeleri ve müşteriye farklı alternatifler sunmaları gerekmektedir. Sepet analizi ve kümeleme analizi sonucunda elde edilen sonuçlar, reklam stratejileri belirlemede, CRM, müşteri profillerinin analiz edilerek çapraz satış tahminlerinin yapılması, yeni müşterilere ulaşabilmek için etkili faktörlerin belirlenmesi, hedef pazarın belirlenmesi, müşteri değerleme, müşteri segmentasyonu ve müşteri ilişkileri yönetimi için kullanılabilir. Elde edilen kümelerle hangi müşteri gruplarının hedeflenmesi gerektiği belirlenmiştir[7].

Bu çalışmanın amacı; bir süpermarketten elde edilen veriler yardımıyla yapay sinir ağları ve zaman serisi tahmin yöntemlerinin uygulanması, uygulanan iki modelin geçmişe dönük tahmin

doğruluklarının kıyas edilerek en uygun sonuç sağlayan modelin seçilmesi, seçilen bu model yardımıyla gelecek dönem taleplerinin haftalar itibariyle tahmin edilmesidir. Yapılan işlemler neticesinde yapay sinir ağı modellerinin, üç et türü de dâhil olmak üzere, zaman serisi yöntemlerinden ARIMA modeline göre daha az tahmin hatası yaptığı ve daha iyi sonuçlar sunduğu görülmektedir. Bu nedenden ötürü, çalışma yapılan firmaya yapay sinir ağları ile talep tahmin modelleri kurması önerilmiştir[8].

Çalışmanın amacı kümelemeyi otomatik hale getirmek ve dışarıdan K parametresinin girilmesine gerek kalınmadan verileri uygun küme sayısınca kümelere yerleştirmektir. Geliştirilen otomatik K-Means algoritması sayısal veriler ve görüntüler üzerinde test edilmiş ve başarılı sonuçlara ulaşılmıştır. Kullanılacak olan kümeleme algoritmasının seçimi, amaca ve veri tipine bağlıdır. Bu çalışmada küme sayısına ilişkin bir varsayım yaparak kümeleme yapan bir algoritma geliştirilmiştir. Uygun küme değerinin belirlenmesi doğru sınıflandırma için oldukça önemlidir. Bankacılık sektöründe kaç farklı müşteri profilinin olduğu, resimde kaç farklı bölge olduğu, tıpta kaç farklı hasta tipi olduğu doğru küme sayısı ile tespit edilebilir. Bu çalışma diğer kümeleme algoritmalarından fuzzy c-means algoritmasına da adapte edilebilir. Daha karmaşık veri setleri üzerinde de denenerek kümeleme performansları tespit edilir[9].

Bu çalışmada, Naive Bayes ile birlikte literatürde Naive Bayes metodunun doğruluk değerini iyileştirdiği ifade edilen Jelinek-Mercer, Dirichlet, Two-Stage teknikleri ve Mutlak Bağlantılı Ağırlıklandırılmış Naive Bayes tekniğinden oluşan 5 ayrı metot ile LOC ve CK metrik kümesine dayalı oluşturulan 3 veri setindeki 34 ayrı ölçüt grubunun çalışma uyumlulukları değerlendirilip, sınıflandırma başarıları kıyaslanmıştır. Kullanılan veri setleri/ölçüt gruplarına göre araştırılan metotlardan Naive Bayes metodu üzerine uygulanan bazı tekniklerin (Dirichlet, Two-Stage) sınıflandırma performansını diğer sınıflandırma metotlarına kıyasla daha da iyileştirdiği sonucunu gösterdi. Yazılım metriklerine yönelik 3 veri seti üzerinde NB, JM NB, D NB, TS NB ve ACWNB teknikleri denenmiştir. Sonucunda yazılım hata sınıflandırması çerçevesinde, uygulayıcılara yüksek doğruluk değerlerine ulaşan metotlar ve metotların uyumlu olarak çalışabildikleri az sayıdaki metrik gruplarına ait bir özet sunulmaktadır[10].

Çalışmada literatürde en çok kullanılan 3 uzaklık ölçüsünden bahsedilmiştir. Bunlar; Öklid uzaklığı, Manhattan uzaklığı ve Chebyshev uzaklık ölçüleridir. Bu çalışmada hava kompresörlerinde kullanıcı kaynaklı piston segmanı aşınması gibi durumlarda oluşan yağ taşınımı arızası araştırılmış ve kompresör üzerindeki etkisi incelenmiştir. Veri toplama sistemi olarak Dewesoft firmasına ait kontrolcü ve yazılımlar kullanılmaktadır. Dijital ve analog kontrollerin yanı sıra analog ölçümler için Sirius modülü kullanılmaktadır. Karmaşıklık matrisi

ile performans metrikleri sonuçlarına göre en doğru cevabı verecek uzaklık ölçütü ve komşu sayısı belirlenebilir. En çok kullanılan uzaklık ölçüsünün ise öklid uzaklığının olduğu belirtilmiştir[11].

Amacı kümeleme algoritmaları kullanılarak önceden belirlenmiş parametrelere göre ülkelerin aldığı değerlerin karşılaştırılması ve anlamlı kümeler oluşturulmasıdır. Dünya Bankası'nın internet sitesinden alınan veriler üzerinde, kümeleme algoritmalarından K-Means ve SOM uygulanarak ülkeler değerlendirilmiştir. Dünya Bankası verilerinin bulunduğu internet sitesinden 2015 yılına ait veriler incelenebilmesi için excel formatında indirilmiştir. Bu algoritmalar sonucunda oluşan kümeler ve Türkiye' nin bu kümelerdeki yeri incelenmiştir. Bu değerlere genel olarak bakıldığında ülkemizin genel olarak iyi bir noktada olduğu söylenebilir. Türkiye'nin daha iyi bir noktaya gelebilmesi için T.C. Kalkınma Bakanlığı tarafından kalkınma planları düzenlenmektedir[12].

3.YÖNTEM VE TEKNİKLER

Karar Ağacı

Karar ağaçları, sınıflandırma ve tahmin için sıkça kullanılan bir veri madenciliği yaklaşımıdır. Karar ağaçlarının sunduğu mantıksal modelin yansıttığı karar kuralları, insanlar tarafından kolayca anlaşılabilecek kadar açıktır[13]. Yüksek sınıflandırma doğruluk oranı ve üretilen basit kurallar gibi özelliklere sahip olduğundan dolayı bu yöntem geniş bir uygulama yelpazesine sahiptir[14]. Karar ağaçları; kişilerin kredi geçmişlerini kullanarak kredi tercihinde bulunması, geçmişte işletmeye en faydalı olan bireylerin özelliklerini kullanarak işe alma süreçlerinin tespit edilmesi, tıbbi gözlem verilerinden hareketle en etkin kararların verilmesi, satışı etkileyen değişkenlerin saptanması, üretim verilerini inceleyerek ürün hatalarına yol açan değişkenlerin belirlenmesi gibi uygulamalarda kullanılmaktadır[15].

Karar ağaçları;

- Düşük maliyetli olması,
- Anlaşılmasının, yorumlanmasının ve veri tabanları ile entegrasyonun kolaylığı,
- Güvenilirliklerinin iyi olması gibi nedenlerden ötürü en yaygın kullanılan sınıflandırma tekniklerinden biridir.

Naive Bayes

Naive Bayes, Bayes Teoremi baz alınarak oluşturulan kolay uygulanabilirlik ve anlaşılabilirlik yönünden avantajlı olan basit makine öğrenme algoritmalarındandır. Eldeki

verilerin belirlenmiş olan sınıflara ait olma olasılıklarını öngören bir algoritmadır. Temeli, istatistikteki Bayes teoremine dayanır. Bu teorem; belirsizlik taşıyan herhangi bir durumun modelinin oluşturularak, bu durumla ilgili evrensel doğrular ve gerçekçi gözlemler doğrultusunda belli sonuçlar elde edilmesine olanak sağlar. Belirsizlik taşıyan durumlarda karar verme konusunda oldukça başarılıdır. Genellikle belirsizlik durumlarında sınıflandırma ve tahmin yapmak için kullanılır. En önemli dezavantajı değişkenler arası ilişkinin modellenmemesi ve değişkenlerin birbirinden tamamen bağımsız olduğu varsayımıdır [16].

K-NN

K-en yakın komşu algoritması (K-NN), gerçekleştiriminin basit ve kolay, öğrenme sürecinin güçlü ve kullanışlı olmasından dolayı sınıflandırmada yaygın biçimde kullanılmaktadır. Makine öğrenmesi, veri madenciliği gibi çok çeşitli alanlarda uygulanmaktadır. K-NN algoritması, en temel örnek tabanlı öğrenme algoritmaları arasındadır. Örnek tabanlı öğrenme algoritmalarında, öğrenme işlemi eğitim setinde tutulan verilere dayalı olarak gerçekleştirilmektedir. Yeni karşılaşılan bir örnek, eğitim setinde yer alan örnekler ile arasındaki benzerliğe göre sınıflandırılmaktadır [17]. K-NN algoritmasında, eğitim setinde yer alan örnekler n boyutlu sayısal nitelikler ile belirtilir. Her örnek n boyutlu uzayda bir noktayı temsil edecek biçimde tüm eğitim örnekleri n boyutlu bir örnek uzayında tutulur. Bilinmeyen bir örnek ile karşılaşıldığında, eğitim setinden ilgili örneğe en yakın k tane örnek belirlenerek yeni örneğin sınıf etiketi, k en yakın komşusunun sınıf etiketlerinin çoğunluk oylamasına göre atanır [18].

Yapay Sinir Ağları

Yapay sinir ağları (YSA), insan beyninin öğrenme yolunu taklit ederek beynin öğrenme, hatırlama, genelleme yapma yolu ile topladığı verilerden yeni veri üretebilme gibi temel işlevlerin gerçekleştirildiği bilgisayar yazılımlarıdır. Yapay sinir ağları; insan beyninden esinlenerek, öğrenme sürecinin matematiksel olarak modellenmesi uğraşı sonucu ortaya çıkmıştır [19].

Yapay sinir ağları aşağıdaki temel özelliklere sahiptir:

- Doğrusal Olmama
- Paralel Çalışma
- Öğrenme
- Genelleme
- Hata Toleransı ve Esneklik

- Eksik Verilerle Çalışma
- Çok Sayıda Değişken ve Parametre Kullanma
- Uyarlanabilirlik

Yapay Sinir Ağları uygulamaları en çok tahmin, sınıflandırma, veri ilişkilendirme, veri yorumlama ve veri filtreleme işlemlerinde kullanılmaktadır [20]. YSA'lar her geçen gün gelişen teknoloji ile birlikte hayatımızın her alanına girmeye başlamışlardır. Özellikle sağlık alanı başta olmak üzere otomotiv, elektronik, enerji, uzay bilimleri, bankacılık, finans ve askeri alanlarında etkin rol almaya başlamıştır[21]. İnsansı robotlarla birlikte bu teknolojiye ilgi daha da artacaktır.

Yapay sinir ağları trafik kontrolünde [22], tıp ve sağlık hizmetlerinde veri madenciliği üzerine [23], İstatistiksel tahmin yöntemleri[24], Meteorolojik yağış verilerinin tahmini[25], Endüstriyel problemlerin çözümünde [26], Güç sistemlerinde yük akış analizi [27]gibi bek çok uygulamasına rastlamak mümkündür.

Çok Katmanlı Algılayıcı Çalışma Şekli

Örneklerin toplanması: Ağın çözmesi istenen olay için daha önce gerçekleşmiş örneklerin bulunması adımıdır. Ağın eğitilmesi için örnekler toplandığı gibi (eğitimseti), ağın test edilmesi içinde örneklerin (testseti) toplanması gerekmektedir[28].

Ağın topolojik yapısının belirlenmesi: Öğrenilmesi istenen olay için oluşturulacak olan ağın yapısı belirlenir. Kaç tane girdi ünitesi, kaç tane ara katman, her ara katmanda kaç tane hücre elemanı ve kaç tane çıktı elemanı olması gerektiği belirlenmektedir[28].

Öğrenme parametrelerinin belirlenmesi: Ağın öğrenme katsayısı, proses elemanlarının toplama ve aktivasyon fonksiyonları, momentum katsayısı gibi parametreler bu adımda belirlenmektedir[28].

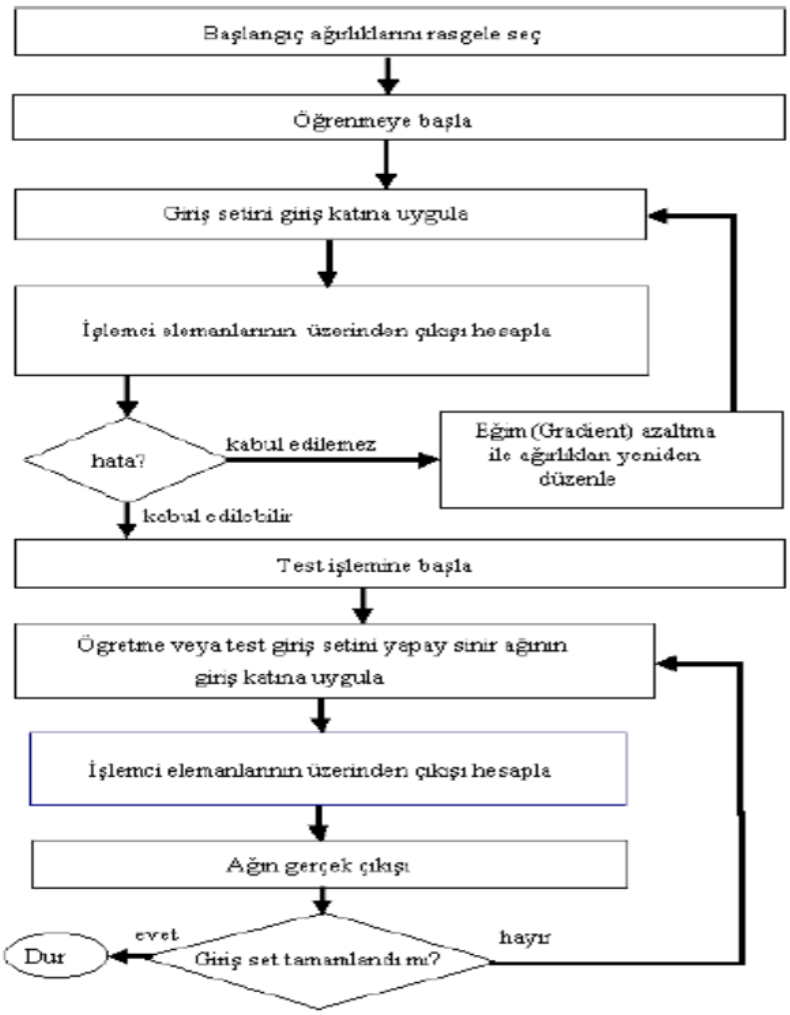
Ağın başlangıç değerlerinin atanması: Hücre elemanlarını bir birine bağlayan ağırlık değerlerinin ve eşik değere başlangıç değerinin atanması[28]

Öğrenme setinden örneklerin seçilmesi ve ağa gösterilmesi: Ağın öğrenmeye başlaması, öğrenme kuralına uygun olarak ağırlıkların değiştirilmesi için ağa örneklerin gösterilmesi[28].

Öğrenme sırasında ileri hesaplamaların yapılması: Verilen girdi için ağın çıktı değerinin hesaplanması[28].

Gerçekleşen çıktının beklenen çıktı ile karşılaştırılması: Ağın ürettiği hata değerlerinin hesaplanması[28].

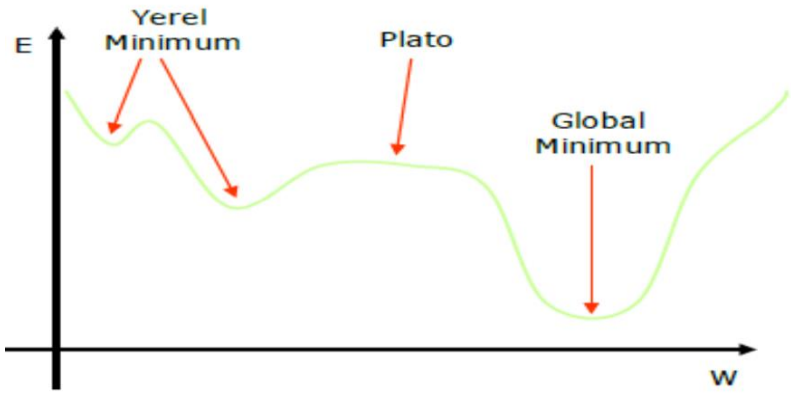
Ağırlıkların değiştirilmesi: Geri hesaplama yöntemi uygulanarak üretilen hatanın azalması için ağırlıkların değiştirilmesi[28].



Momentum Katsayısı

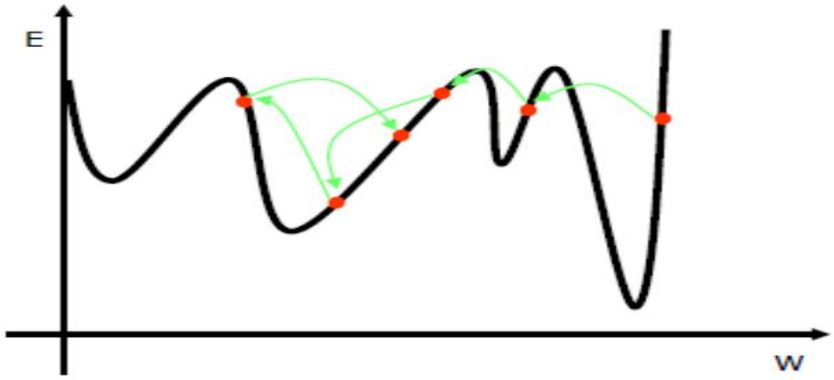
Çok katmanlı ağların yerel sonuçlara takılıp kalmaması için momentum katsayısı geliştirilmiştir. Bu katsayının iyi kullanılması yerel çözümleri kabul edilebilir hata düzeyinin altına çekebilmektedir. Çok katmanlı ağların diğer bir sorunu ise öğrenme süresinin çok uzun olmasıdır. Ağırlık değerleri başlangıçta büyük değerler olması durumunda ağın yerel sonuçlara düşmesi ve bir yerel sonuçtan diğerine sıçramasına neden olmaktadır. Eğer ağırlıklar küçük aralıkta seçilirse o zamanda ağırlıkların doğru değerleri bulması uzun sürmektedir. Momentum katsayısı, yerel çözümlere takılmayı önler. Bu değerin çok küçük

seçilmesi yerel çözümlerden kurtulmayı zorlaştırır. Değerin çok büyük seçilmesi ise tek bir çözüme ulaşmada sorunlar yaratabilir[28].

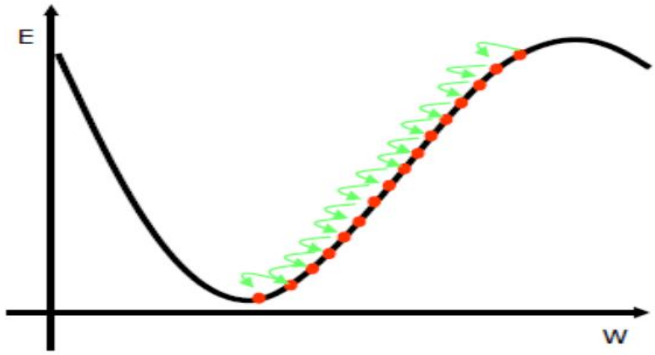


Öğrenme Katsayısı

Öğrenme katsayısı ağırlıkların değişim miktarını belirler. Eğer öğrenme katsayısı gereğinden büyük olursa problem uzayında rastgele gezinme olur. Bunun da ağırlıkları rastgele değiştirmekten farkı olmaz[28].



Eğer öğrenme katsayısı çok küçük olursa çözüme ulaşmak daha uzun sürer[28].



Çok Katmanlı Ağlarda Dikkat Edilmesi Gereken Noktalar

- Momentum katsayısı, bir önceki iterasyon değişiminin belirli bir oranının yeni değişim miktarını etkilemesidir[28].
- Bu özellikle yerel çözüme takılan ağların bir sıçrama ile daha iyi sonuçlar bulmasını sağlamak amacı ile geliştirilmiştir[28].
- Momentum katsayısı, bir önceki iterasyon değişiminin belirli bir oranının yeni değişim miktarını etkilemesidir. Çok Katmanlı Ağlarda Dikkat Edilmesi Gereken Noktalar Bu özellikle yerel çözüme takılan ağların bir sıçrama ile daha iyi sonuçlar bulmasını sağlamak amacı ile geliştirilmiştir[28].
- Değerin küçük olması yerel çözümlerden kurtulmayı zorlaştırır. Çok büyük değerler ise bir çözüme ulaşmada sorunlar yaşanabilir[28].

3.1.ÇALIŞMANIN AMACI

Ses ile cinsiyet tanıma veri setini Karar Ağacı, K-NN, Naive Bayes, Yapay Sinir Ağları-Çok Katmanlı Algılayıcı algoritmaları ile sınırlandırmak. Algoritmaları hem birbirleri ile hem de kendi içerisinde farklı özellikler ile kıyaslayarak başarı oranını etkileyen faktörleri tespit etmek ve en başarılı algoritmayı seçmek.

3.2.VERİ SETİNİN TANITIMI

Bu veritabanı, sesin ve konuşmanın akustik özelliklerine dayalı olarak bir sesi erkek veya kadın olarak tanımlamak için oluşturulmuştur. Veri seti, erkek ve kadın konuşmacılardan toplanan 3.168 kayıtlı ses örneğinden oluşmaktadır. Ses örnekleri, seewave ve tuneR paketleri kullanılarak R'de akustik analiz ile önceden işlenir, analiz edilen frekans aralığı 0hz-280hz'dir.

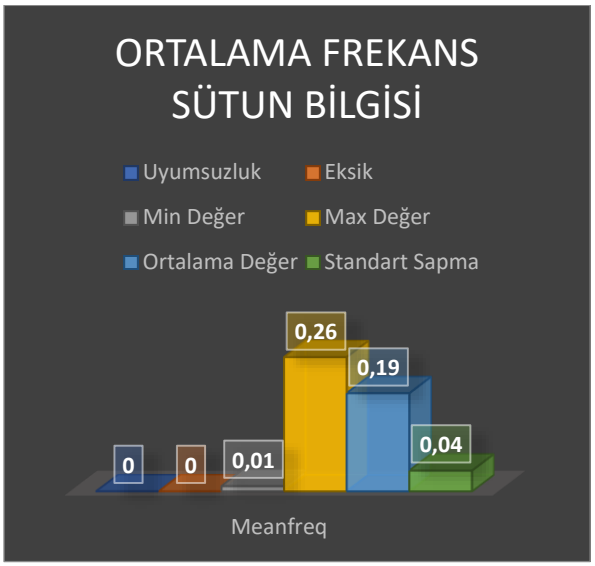
Veri Kümesi

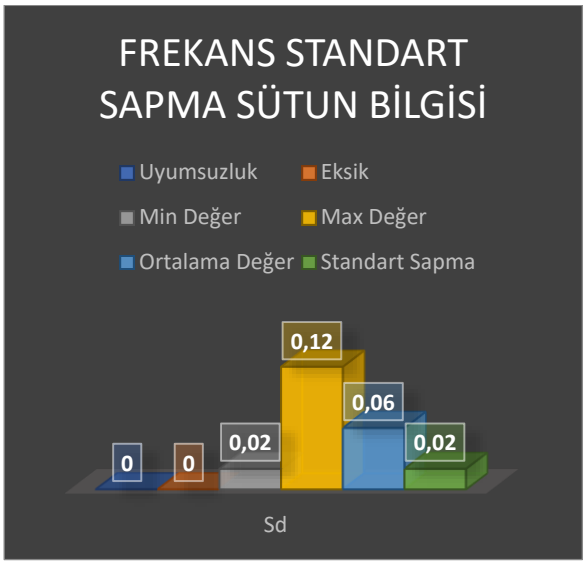
Her sesin aşağıdaki akustik özellikleri ölçülür ve CSV'ye dahil edilir:

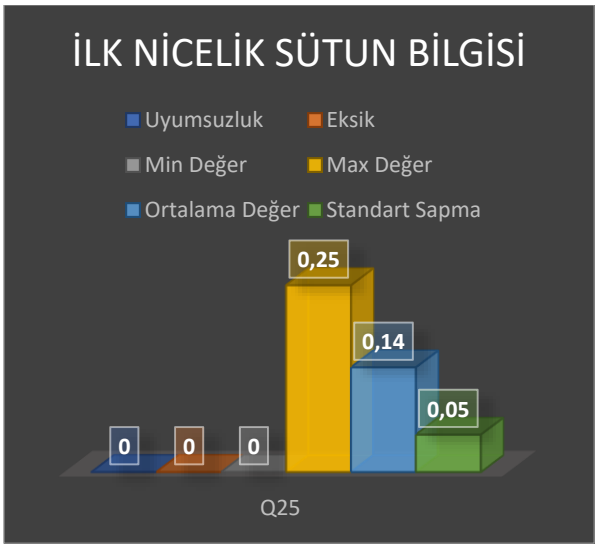
SÜTUNLAR	AÇIKLAMASI
Meanfreq	Ortalama frekans (kHz cinsinden)
Sd	Frekansın standart sapması
Q25	İlk nicelik (kHz cinsinden)
Q75	Üçüncü kuantil (kHz cinsinden)
IQR	Çeyrekler arası aralık (kHz cinsinden)
Sfm	Spektral düzlük

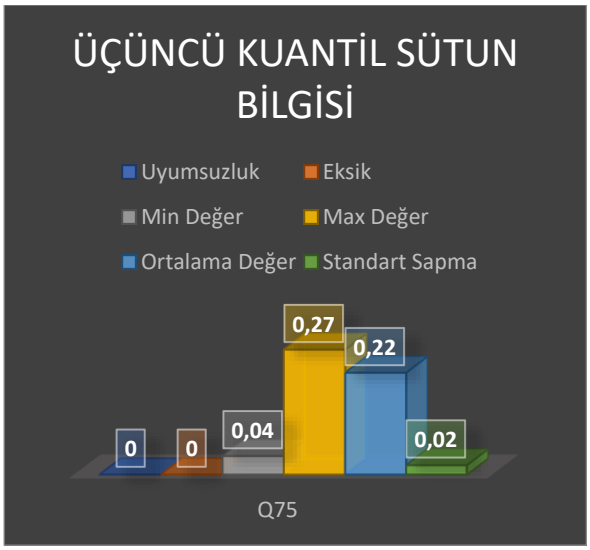
Centroid	Frekans centroid (spesifikasyona bakınız)
Meanfun	Akustik sinyalde ölçülen temel frekansın ortalaması
Minfun	Akustik sinyalde ölçülen minimum temel frekans
Maxfun	Akustik sinyalde ölçülen maksimum temel frekans
Meandom	Akustik sinyalde ölçülen baskın frekansın ortalaması
Mindom	Akustik sinyalde ölçülen minimum baskın frekans
Maxdom	Akustik sinyalde ölçülen maksimum baskın frekans
Dfrange	Akustik sinyalde ölçülen baskın frekans aralığı
Modindx	Modülasyon indeksi. Temel frekansların bitişik ölçümleri arasındaki birikmiş mutlak farkın frekans aralığına bölünmesiyle hesaplanır
Label	Kadın veya Erkek

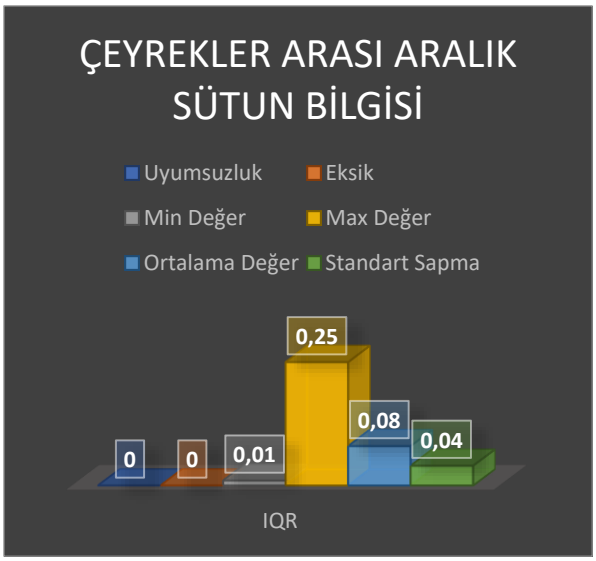
Parametreler Ve Özellikleri



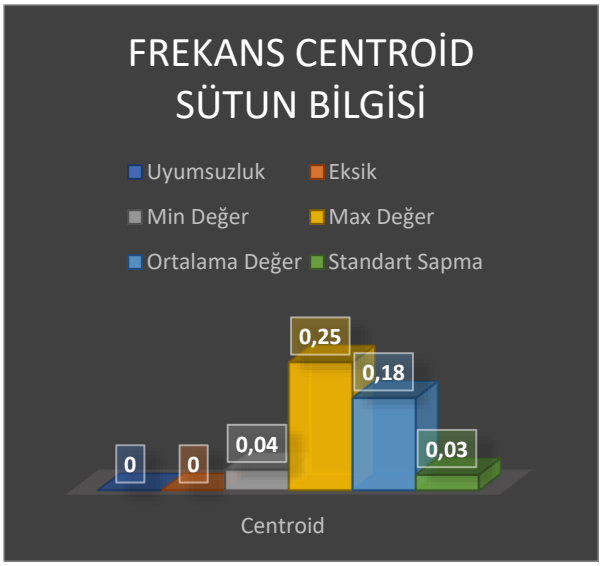


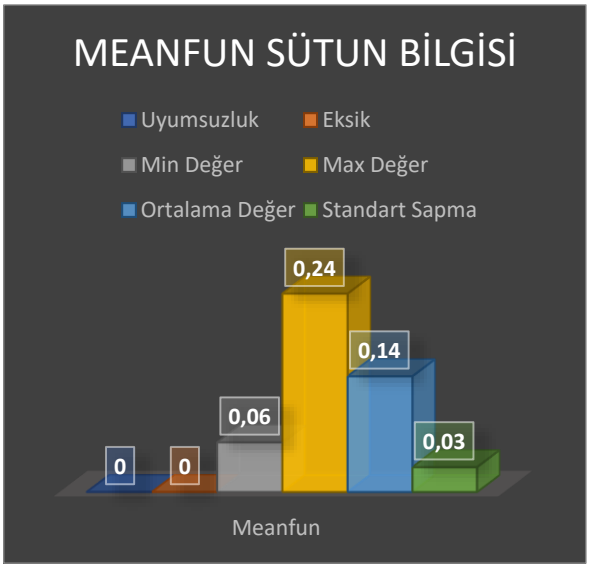

İLK NİCELİK SÜTUN BİLGİSİ
Uyumsuzluk
Eksik
Min Değer
Max Değer
Ortalama Değer
Standart Sapma
0
0
0
0,25
0,14
0,05
Q25

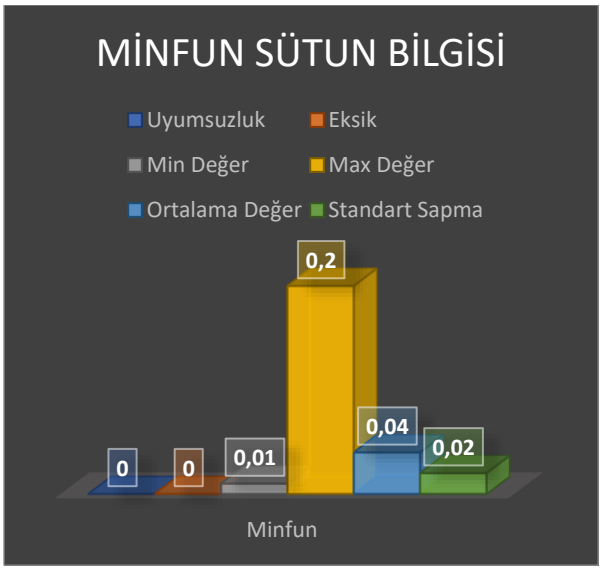

ÜÇÜNCÜ KUANTİL SÜTUN BİLGİSİ
Uyumsuzluk
Eksik
Min Değer
Max Değer
Ortalama Değer
Standart Sapma
0
0
0,04
0,27
0,22
0,02
Q75

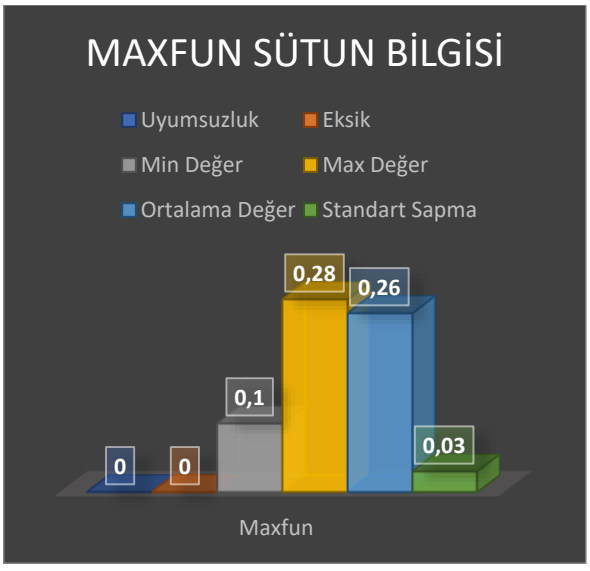

ÇEYREKLER ARASI ARALIK SÜTUN BİLGİSİ
Uyumsuzluk
Eksik
Min Değer
Max Değer
Ortalama Değer
Standart Sapma
0
0
0,01
0,25
0,08
0,04
IQR

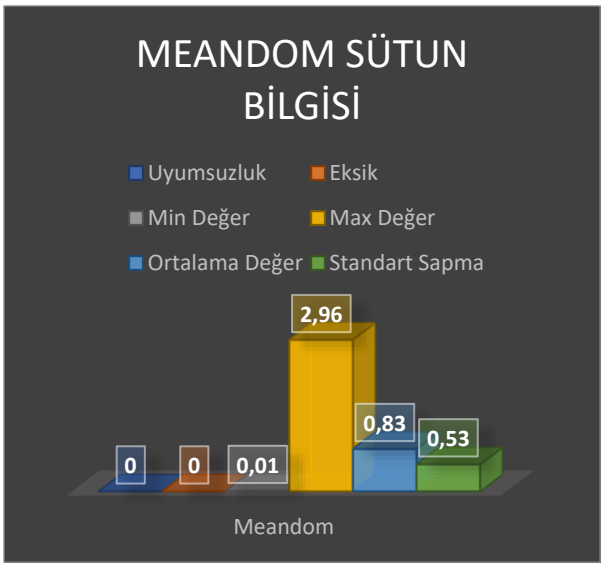

SPEKTRAL DÜZLÜK SÜTUN BİLGİSİ
Uyumsuzluk
Eksik
Min Değer
Max Değer
Ortalama Değer
Standart Sapma
0
0
0,04
0,84
0,41
0,18
Sfm

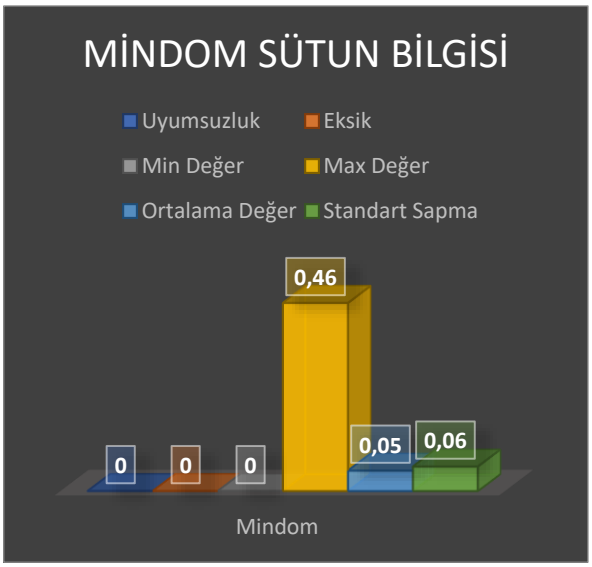

FREKANS CENTROİD SÜTUN BİLGİSİ
Uyumsuzluk
Eksik
Min Değer
Max Değer
Ortalama Değer
Standart Sapma
0
0
0,04
0,25
0,18
0,03
Centroid

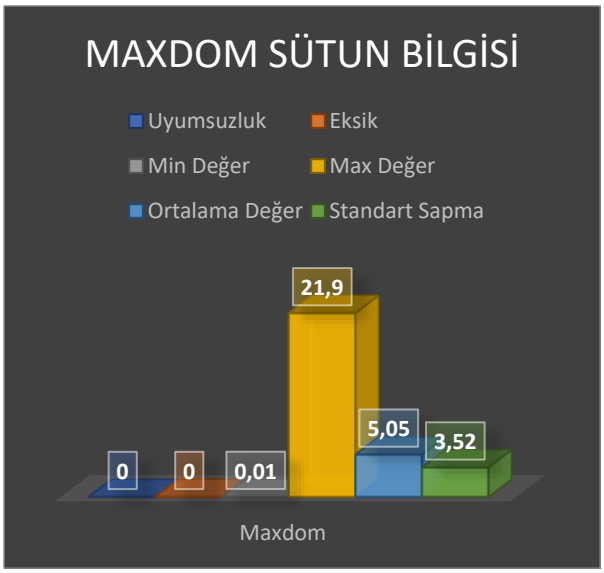

MEANFUN SÜTUN BİLGİSİ
Uyumsuzluk
Eksik
Min Değer
Max Değer
Ortalama Değer
Standart Sapma
0
0
0,06
0,24
0,14
0,03
Meanfun

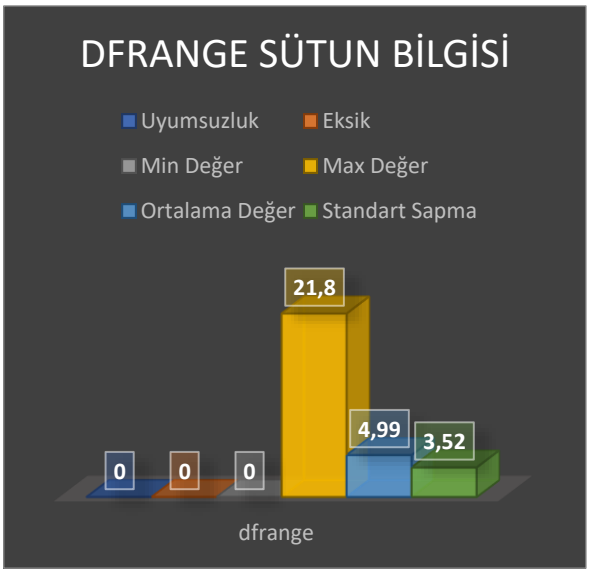

MİNFUN SÜTUN BİLGİSİ
Uyumsuzluk
Eksik
Min Değer
Max Değer
Ortalama Değer
Standart Sapma
0
0
0,01
0,2
0,04
0,02
Minfun

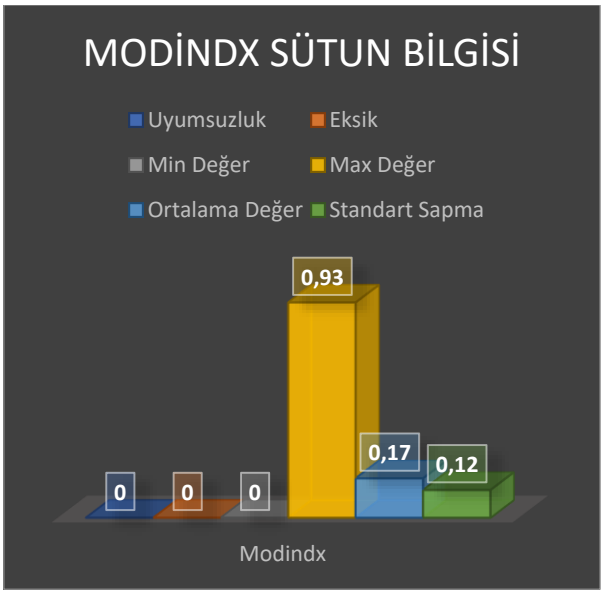

MAXFUN SÜTUN BİLGİSİ
Uyumsuzluk
Eksik
Min Değer
Max Değer
Ortalama Değer
Standart Sapma
0
0
0,1
0,28
0,26
0,03
Maxfun


MEANDOM SÜTUN BİLGİSİ
Uyumsuzluk
Eksik
Min Değer
Max Değer
Ortalama Değer
Standart Sapma
0
0
0,01
2,96
0,83
0,53
Meandom

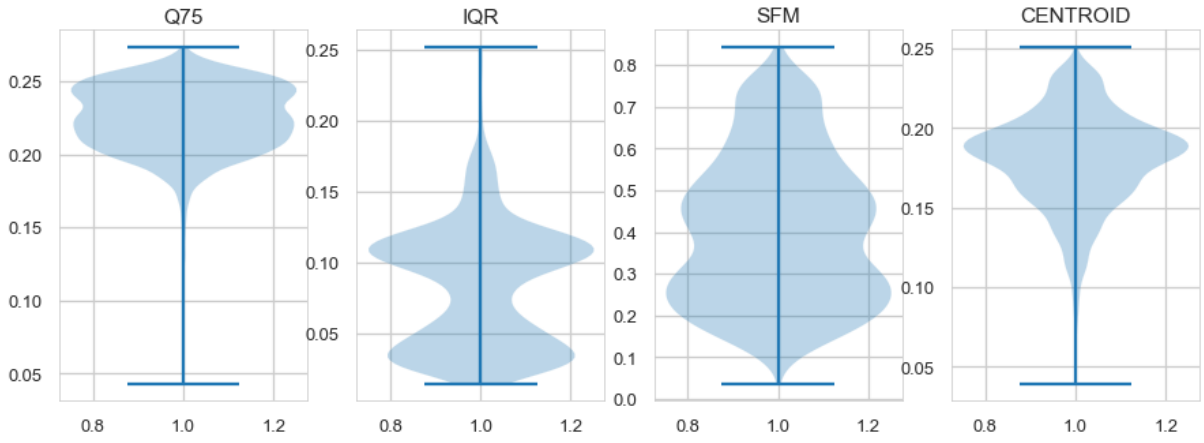

MİNDOM SÜTUN BİLGİSİ
Uyumsuzluk
Eksik
Min Değer
Max Değer
Ortalama Değer
Standart Sapma
0
0
0
0,46
0,05
0,06
Mindom

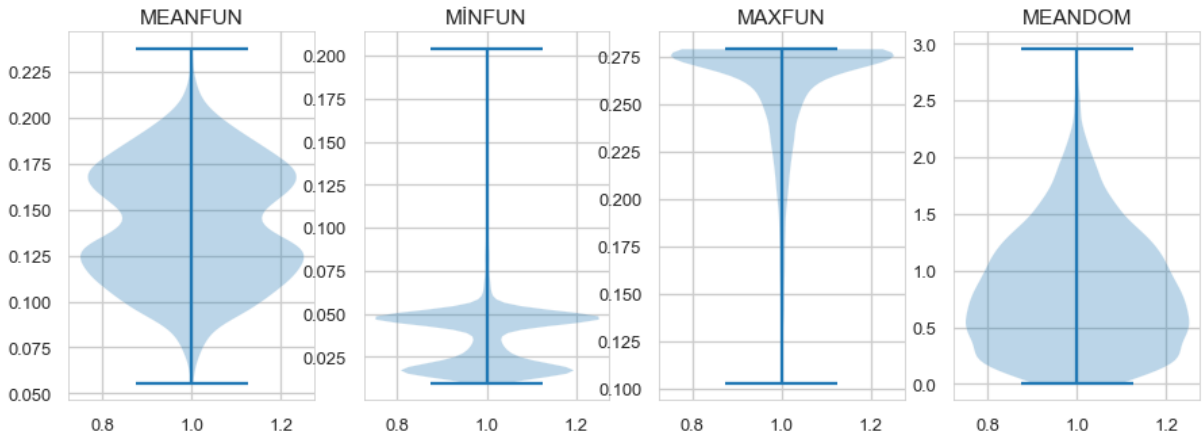

MAXDOM SÜTUN BİLGİSİ
Uyumsuzluk
Eksik
Min Değer
Max Değer
Ortalama Değer
Standart Sapma
0
0
0,01
21,9
5,05
3,52
Maxdom

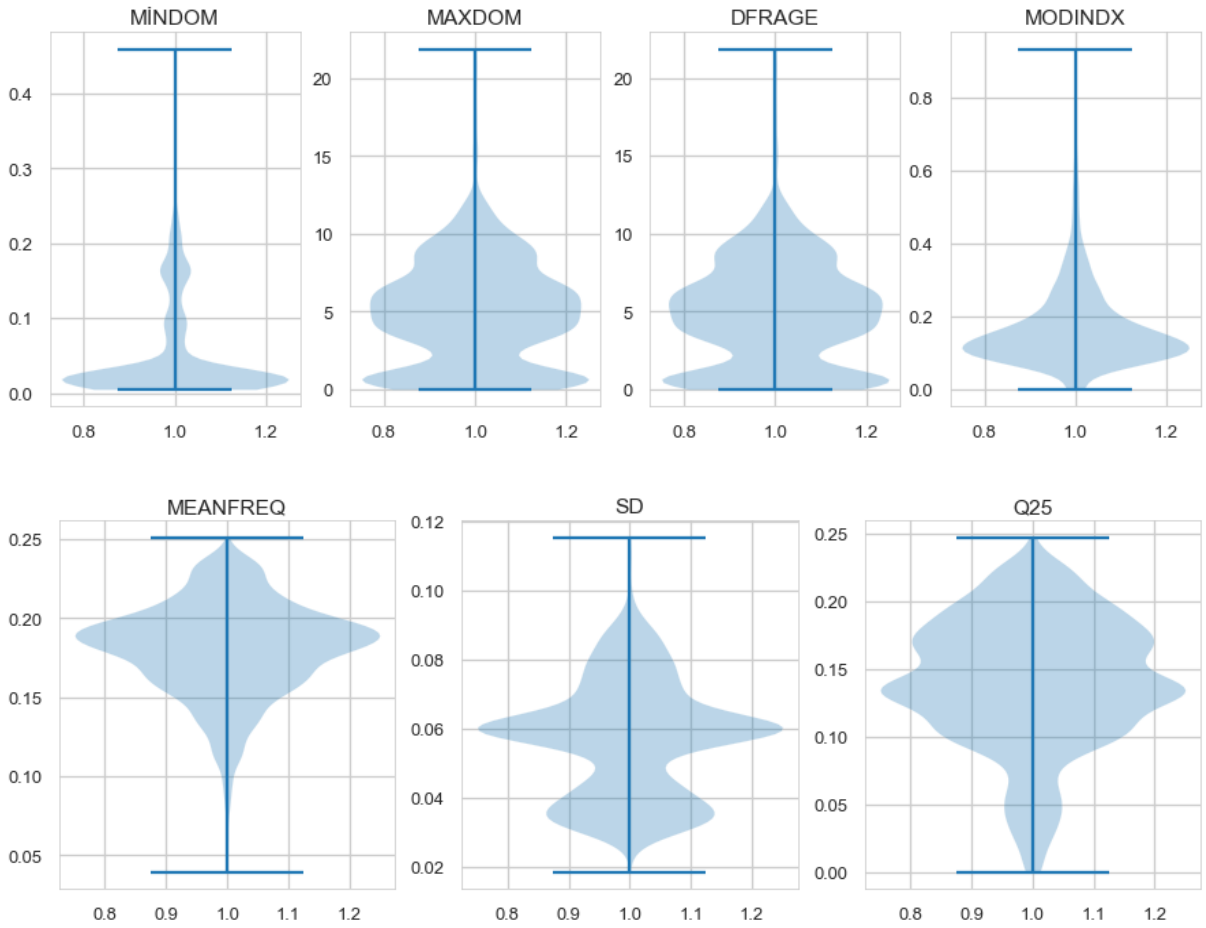

DFRANGE SÜTUN BİLGİSİ
Uyumsuzluk
Eksik
Min Değer
Max Değer
Ortalama Değer
Standart Sapma
0
0
0
21,8
4,99
3,52
dfrange



Parametrelerin Aldıkları Değerlerin Dağılımı (Violin Grafik)

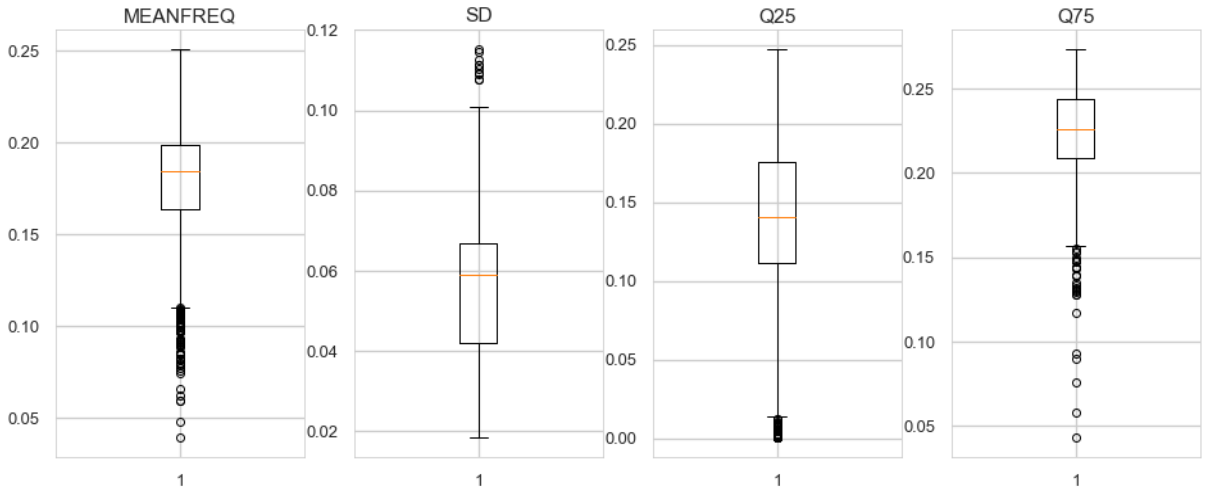


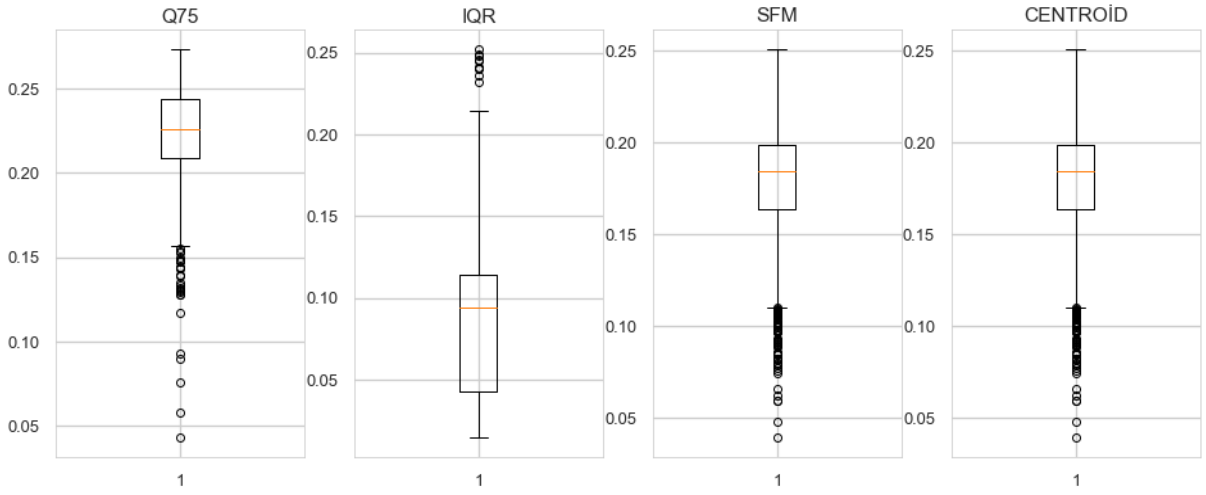


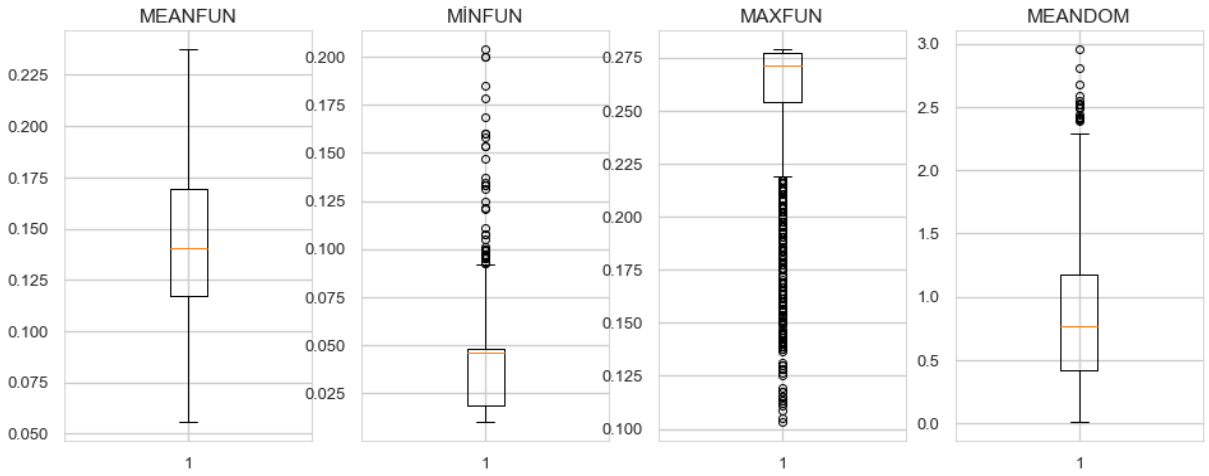


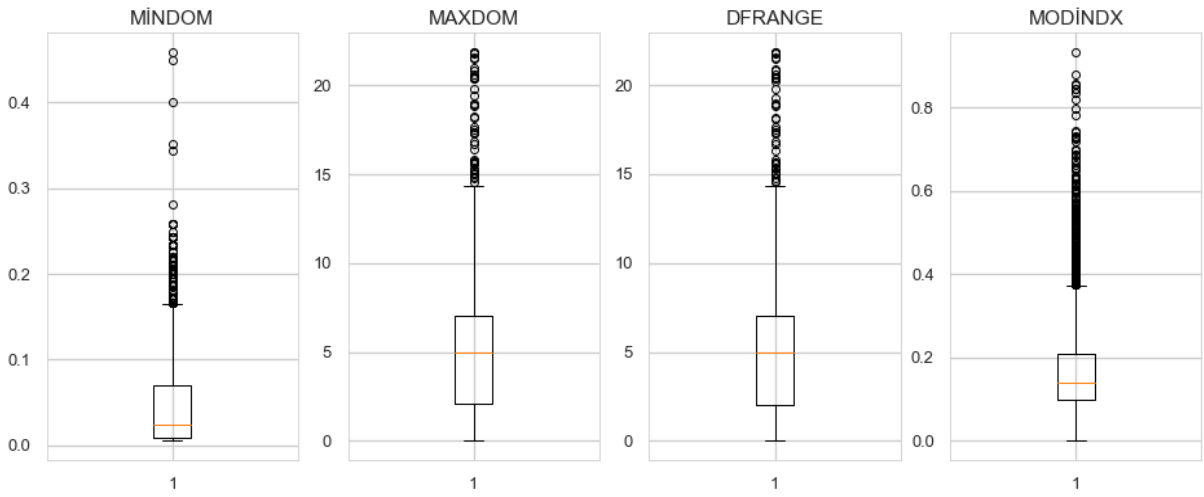
Yukarıdaki grafikleri incelediğimizde kullandığımız parametrelerin hangi değerler arasında yığılım yaptığı gösterilmiştir. Sığ olan dağılımlar dengeli olduğunu gösterirken sığ olmayan dağılımlar dengesizlik olduğunu gösterir. Sığ olmayan dağılımlar veri seti için olumsuz bir anlam ifade etmez çünkü her parametre kendi içerisinde farklı bir ölçüte göre hesaplanır. Sığ dağılımlarda dalgalanmanın birden azaldığı yerlerde dengeli dağılımın azaldığı gözlemlenir. Örneğin “Maxfun” parametresinde 2.5 değerinden sonra keskin bir azalış görülmüştür. Bu durum “Maxfun” parametresinin ölçümü sırasında aykırı sonuçlar elde edildiğini gösterir

Parametrelerin Aldıkları Değerlerin Aykırılık Durumu (Kutu Grafik)

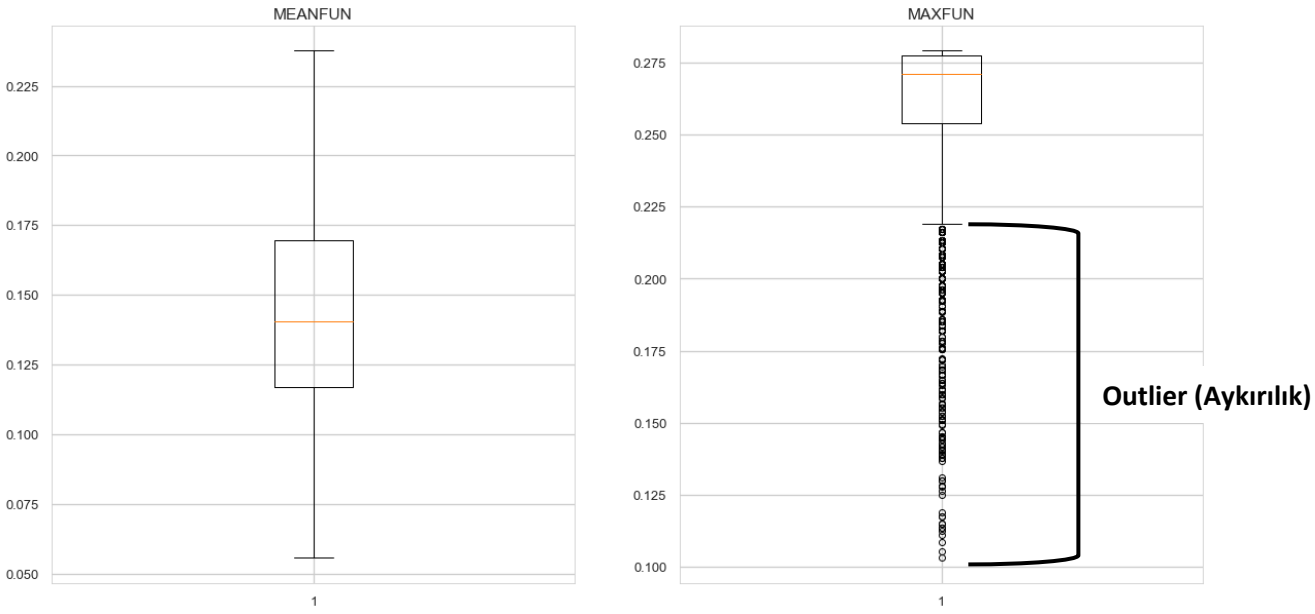








Yukarıdaki grafikler incelendiğinde parametreler ölçülürken ele alınan değerlerin aykırılık oranları gösterilmiştir.



Örneğin “Meanfun” parametresinin ölçümü yapılırken hiç aykırı değer elde edilmemişken “Maxfun” parametresi ölçümü yapılırken çok fazla aykırı değer elde edilmiştir. Aykırı veri oranı en düşük olan “Meanfun”,”Q25”,”Sd”,”IQR” giriş değerleri temel alınmıştır.

3.3.DEĞERLENDİRME KRİTERLERİNİN AÇIKLAMASI

Correctly Classified Instances

Veri setinde sınıflandırma yönteminde tahmin edilen doğruluk sayısıdır. Aynı zamanda yüzde olarakta sınıflandırma yönteminin başarı oranını göstermektedir.

Kappa Statistic

Cohen'in kappa katsayısı iki değerleyici arasındaki karşılaştırmalı uyuşmanın güvenirliğini ölçen bir istatistik yöntemidir. Cohen'in kappa ölçüsü her biri *N* tane maddeyi *C* tane birbirinden karşılıklı hariç olan kategoriye ayıran iki değerleyicinin arasında bulunan uyuşmayı ölçer. Cohen'in kappa ölçüsü bu uyuşmanın bir şans eseri olabileceğini de ele aldığı için basit yüzde orantı olarak bulunan uyuşmadan daha güçlü bir sonuç verdiği kabul edilir. Cohen'in kappa istatistiklerinin sonucunun yorumlamasının tablosunu alt kısımda görmekteyiz.

κ	Yorum
< 0	Hiç uyuşma olmaması
0.0 — 0.20	Önemsiz uyuşma olması
0.21 — 0.40	Orta derecede uyuşma olması
0.41 — 0.60	Ekseriyetle uyuşma olması
0.61 — 0.80	Önemli derecede uyuşma olması
0.81 — 1.00	Neredeyse mükemmel uyuşma olması

Mean Absolute Error

Ortalama mutlak hata iki sürekli değişken arasındaki farkın ölçüsüdür. MAE, yönlerini dikkate almadan bir dizi tahmindeki hataların ortalama büyüklüğünü ölçen, tüm tekil hataların ortalamada eşit olarak ağırlıklandırıldığı doğrusal bir skordur. MAE değeri 0'dan ∞'a kadar değişebilir. Negatif yönelimli puanlar yani daha düşük değerlere sahip tahminleyiciler daha iyi performans gösterir.

Root Mean Squared Error

Bir makine öğrenmesi modelinin, tahminleyicinin tahmin ettiği değerler ile gerçek değerleri arasındaki uzaklığın bulunmasında sıklıkla kullanılan, hatanın büyüklüğünü ölçen bir metriktir. Verilere en iyi uyan çizgi etrafında o verilerin ne kadar yoğun olduğunu söyler. RMSE değeri 0'dan ∞'a kadar değişebilir. Negatif yönelimli puanlar yani daha düşük değerlere sahip tahminleyiciler daha iyi performans gösterir. RMSE değerinin sıfır olması modelin hiç hata yapmadığı anlamına gelir.

Relative Absolute Error

Bir tahminde hesaplanan değeri ve gerçek değeri biliyorsak bağıl mutlak hata hesaplaması yapılır. Gerçek değer ile hesaplanan değer arasındaki farkın gerçek değere oranlanması ile elde edilir. Göreli mutlak hata 0'a yakın olduğunda başarı elde edilir ve negatif değer almaz.

Root Relative Squared Error

Gerçek değerlerin yalnızca ortalamasıdır. Böylece, göreceli kare hata toplamın karesi alınmış hatayı alır ve basit tahmin edicinin toplam kare hatası ile bölerek normalleştirir. Göreceli karesel hatanın karekökü alınarak, hata tahmin edilen miktarla aynı boyutlara indirgenir. RRSE değeri 0'dan ∞'a kadar değişebilir. Göreli hata karakök 0'a yakın olduğunda başarı elde edilir ve negatif değer almaz.

Total Number of Instances

Veri setinde sınıflandırma yönteminde kullanılan verilerin toplam sayısıdır.

Gerçek Pozitif Değerlerin Oranı (True Positive Rate)

Sınıflayıcının ne kadar gerçek pozitif değeri doğru tahmin ettiğinin bir ölçüsüdür. Hassasiyet, İsabet Oranı veya Hatırlama olarak da bilinir. TP Rate kısaca doğruya doğru demektir

Yanlış Pozitif Değerlerin Oranı (False Positive Rate)

Gerçek değeri 0 olmasına karşın 1 olarak tahmin edilenlerin oranıdır. FP Rate kısaca doğruya yanlış demektir.

Hassasiyet (Precision)

Tüm sınıflardan, doğru olarak ne kadar tahmin edildiğinin bir ölçüsüdür. Doğru olarak tahmin edilenlerin, toplama oranıdır. Doğru ne kadar tahmin edildiğinin bir ölçüsüdür. 0 ile 1 arasında değer alır, mümkün olduğunca yüksek olmalıdır.

Geri Çağırma (Recall)

Pozitif durumların ne kadar başarılı tahmin edildiğini gösterir. En iyi değer 1, en kötü değer 0'dır. Örnekle açıklarsak kullandığım veri setinde true pozitifler sesin kadın olduğu tahmin edilen ve gerçekte sesin kadın olduğu insanlar, false negativeler yani sesin erkek olduğu tahmin edilen ve gerçekte sesin erkek olmadığı insanlar. Recall'e bakılma sebebi tamamiyle paydadaki

false negativeler, yani sesin erkek olduğu tahmin edilen ve gerçekte sesin kadın olduğu insanlar. Bu verisetiyle ilgili bir çalışma yapıldığında kadın olan bir sesin erkek olarak algılanmasının maliyeti büyük olacağı için recall false negativelerin göz ardı edilemez durumlarda kullanılan önemli bir metriktir.

Roc-Area

ROC bize modelin true positive rate'iyle false positive rate'i cinsinden ne kadar iyi ayrım yapabildiğini açıklar. AREA ise ROC eğrisinin altında kalan alanı verir, 0'la 1 arasındadır, 0'sa bütün tahminler yanlıştır. True positive rate kısaca gerçekte durum pozitifse bunların kaçını pozitif tahmin ettiğimizi gösterir, false positive rate de gerçekte durum negatifken bunların kaçını pozitif olarak tahmin ettiğimizi (yanlış alarm da denir) gösterir.

F-Measure

Gerçek olumlu oranın (recall) ve hassasiyetin ağırlıklı ortalamasıdır. F-Measure 0 ile 1 arasında değer alır. Çıkan değer 1'e ne kadar yakınsa model o kadar başarılı olur.

MCC

İkili (iki sınıflı) sınıflandırmaların kalitesinin bir ölçüsü olarak kullanılır. Doğru ve yanlış pozitif ve negatifleri hesaba katar ve genellikle sınıflar çok farklı boyutlarda olsa bile kullanılabilecek dengeli bir ölçü olarak kabul edilir. 0 ile 1 arasında değer alır. Çıkan değer 1'e ne kadar yakınsa model o kadar başarılı olur.

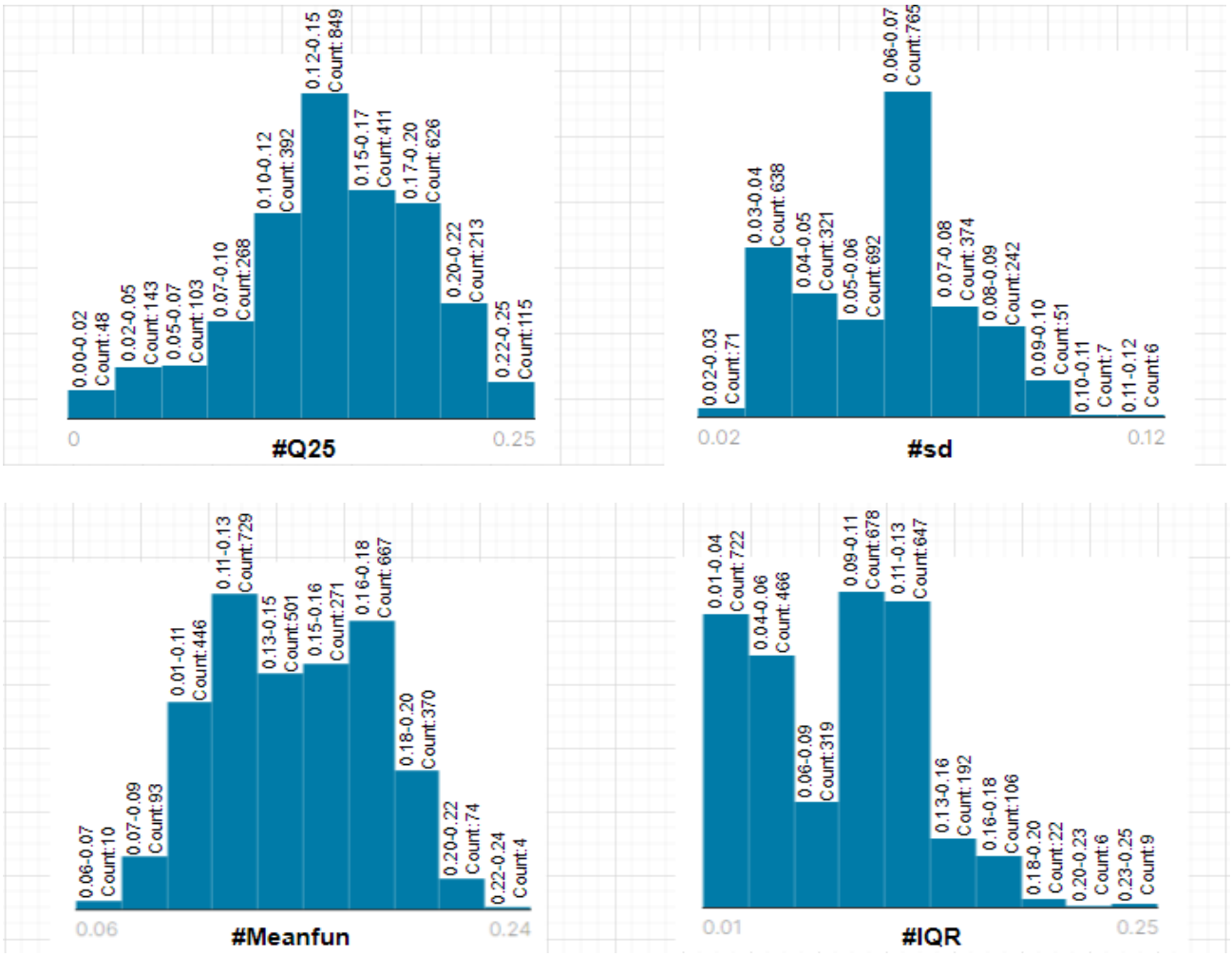
PRC Area

Bununla birlikte, dengesiz veri setlerine dayalı olarak ROC eğrilerinin görsel yorumu ve karşılaştırmaları yanıltıcı olabilir. ROC eğrisine bir alternatif, hassaslık-geri çağırma eğrisidir (PRC). ROC eğrilerinden daha az kullanılır, ancak göreceğimiz gibi PRC, dengesiz veri kümeleri için daha iyi bir seçim olabilir. 0 ile 1 arasında değer alır. Çıkan değer 1'e ne kadar yakınsa model o kadar başarılı olur.

3.4.KULLANILAN YÖNTEMLERİN FORMÜL HESAPLAMALARI

3.4.1.Ses ve Cinsiyet Tanıma Veri Setini Karar Ağacı Formülü İle Örnek Hesaplama

Girdi Değerlerinin Dağılımları



Yukarıdaki grafikler incelendiğinde temel alınan parametrelerde hangi aralıkta kaç adet verinin olduğu gösterilmiştir. Entropi hesaplaması için yukarıdaki dağılımlar temel alınmıştır.

Meanfun Sütunu Entropi Hesaplama

Aralık	*Toplam Adet*	*Male Sayı*	*Female Sayı*
0.06-0.07	10	10	0
0.07-0.09	93	93	0
0.09-0.11	446	442	4
0.11-0.13	729	713	16
0.13-0.15	501	308	193
0.15-0.16	271	4	267
0.16-0.18	667	14	653
0.18-0.20	370	0	370
0.20-0.22	74	0	74
0.22-0.24	4	0	7
Toplam	3168	1584	1584

Yukarıda gösterilen "Meanfun" kategorisi temel alınarak örnek entropi hesaplaması gösterilmiştir. Entropi değerinin hesaplanabilmesi için ele alınan parametrenin belirli aralıklara göre kadın ve erkek sayılarının kaç adet olduğu bilinmelidir. Bu bilgiye de spyder üzerinden analiz yapılarak ulaşılmıştır.

H(Label)= -(1584/3168 **x** $\log_2$1584/3168 + 1584/3168 **x** $\log_2$1584/3168) = **Sonuç1**

|Meanfun $_{\mathbf{0.05\text{-}0.07}}$|=10

|Meanfun $_{\mathbf{0.07\text{-}0.09}}$|=93

|Meanfun $_{\mathbf{0.09\text{-}0.11}}$|=446

|Meanfun $_{\mathbf{0.11\text{-}0.13}}$|=729

|Meanfun $_{\mathbf{0.13\text{-}0.15}}$|=501

|Meanfun $_{\mathbf{0.15\text{-}0.16}}$|=271

|Meanfun $_{\mathbf{0.16\text{-}0.18}}$|=667

|Meanfun $_{\mathbf{0.18\text{-}0.20}}$|=370

|Meanfun $_{\mathbf{0.20\text{-}0.22}}$|=74

|Meanfun $_{\mathbf{0.22\text{-}0.24}}$|=7

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|Sv|}{|S|} Entropy(S_v)$$

H(Meanfun, Label)= 10/3168 x H(Meanfun $_{\mathbf{0.05\text{-}0.07}}$) + 93/3168 x H(Meanfun $_{\mathbf{0.07\text{-}0.09}}$) + 446/3168 x H(Meanfun $_{\mathbf{0.09\text{-}0.11}}$) + 729/3168 x H(Meanfun $_{\mathbf{0.11\text{-}0.13}}$) + 501/3168 x H(Meanfun $_{\mathbf{0.13\text{-}0.15}}$) + 271/3168 x H(Meanfun $_{\mathbf{0.15\text{-}0.16}}$) + 667/3168 x H(Meanfun $_{\mathbf{0.16\text{-}0.18}}$) + 370/3168 x H(Meanfun $_{\mathbf{0.18\text{-}0.20}}$) + 74/3168 x H(Meanfun $_{\mathbf{0.20\text{-}0.22}}$) + 7/3168 x H(Meanfun $_{\mathbf{0.22\text{-}0.24}}$)

H(Meanfun $_{\mathbf{0.05\text{-}0.07}}$)= -(10/10 x $\log_2$10/10)= **s1**

H(Meanfun $_{\mathbf{0.07\text{-}0.09}}$)= -(93/93 x $\log_2$93/93)= **s2**

H(Meanfun $_{\mathbf{0.09\text{-}0.11}}$)= -(442/446 x $\log_2$442/446 + 4/446 x $\log_2$4/446)= **s3**

H(Meanfun $_{\mathbf{0.11\text{-}0.13}}$)= -(713/729 x $\log_2$713/729 + 16/729 x $\log_2$16/729)= **s4**

H(Meanfun $_{\mathbf{0.13\text{-}0.15}}$)= -(308/501 x $\log_2$308/501 + 193/501 x $\log_2$193/501)= **s5**

H(Meanfun $_{\mathbf{0.15\text{-}0.16}}$)= -(4/271 x $\log_2$4/271 + 267/271 x $\log_2$267/271)= **s6**

H(Meanfun $_{\mathbf{0.16\text{-}0.18}}$)= -(14/667 x $\log_2$14/667 + 653/667 x $\log_2$653/667)= **s7**

H(Meanfun $_{\mathbf{0.18\text{-}0.20}}$)= -(370/370 x $\log_2$370/370)= **s8**

H(Meanfun $_{\mathbf{0.20\text{-}0.22}}$)= -(74/74 x $\log_2$74/74)= **s9**

H(Meanfun $_{\mathbf{0.22\text{-}0.24)}}$= -(7/7 x $\log_2$7/7)= **s10**

H(Meanfun, Label)= 10/3168 x s1 + 93/3168 x s2 + 446/3168 x s3 + 729/3168 x s4 + 501/3168 x s5 + 271/3168 x s6 + 667/3168 x s7 + 370/3168 x s8 + 74/3168 x s9 + 7/3168 x s10= **Sonuç2**

Entropi → Kazanç(Meanfun, Label)= H(Label) – H(Meanfun, Label) → **Sonuç1 – Sonuç2**

Burada elde edilen Meanfun entropi değeri IQR, Q25, Sd parametrelerinin entropi değerinden daha yüksektir. Meanfun parametresinden sonra gelecek bir diğer koşul tekrar entropi değerleri hesaplanarak bulunur. Veri dağılımı sığ olan parametrelerin entropi değeri yüksek olduğunu göz önünde bulundurduğumuzda "Meanfun" parametresinden sonra gelecek 2. Koşulun "IQR" olacağı görülmektedir. 2. Koşulun "IQR" parametresi olduğunu karar ağacında da görmekteyiz. Bu formül uygulama başlığı altında karar ağacı sınıflandırma yöntemlerinde Örnek1'in hesaplamasıdır.

3.4.2.Ses ve Cinsiyet Tanıma Veri Setini Naive Bayes Formülü İle Örnek Hesaplama

Giriş Değerleri Aralıklarının Female ve Male Olarak Dağılımı

Meanfun Sütunu

Aralık	Male	Female	Toplam Adet
0.06/0.07	10	0	10/3168
0.07/0.09	93	0	93/3168
0.09/0.11	442	4	446/3168
0.11/0.13	713	16	729/3168
0.13/0.15	308	193	501/3168
0.15/0.16	4	267	271/3168
0.16/0.18	14	653	667/3168
0.18/0.20	0	370	370/3168
0.20/0.22	0	74	74/3168
0.22/0.24	0	7	7/3168
Toplam	1584/3168	1584/3168	

Q25 Sütunu

Aralık	Male	Female	Toplam Adet
0.00/0.02	26	22	48/3168
0.02/0.05	64	79	143/3168
0.05/0.07	49	54	103/3168
0.07/0.10	236	32	268/3168
0.10/0.12	375	17	392/3168
0.12/0.15	737	112	849/3168
0.15/0.17	77	334	411/3168
0.17/0.20	10	616	626/3168
0.20/0.22	6	207	213/3168

0.22/0.25	4	111	115/3168
Toplam	1584/3168	1584/3168	

SD Sütunu

Aralık	**Male**	**Female**	**Toplam Adet**
0.02-0.03	0	71	71/3168
0.03/0.04	0	638	638/3168
0.04/0.05	36	285	321/3168
0.05/0.06	506	186	692/3168
0.06/0.07	611	154	765/3168
0.07/0.08	275	99	374/3168
0.08/0.09	138	104	242/3168
0.09/0.10	18	33	51/3168
0.10/0.11	0	7	7/3168
0.11/0.12	0	6	6/3168
Toplam	1584/3168	1583/3168	

IQR Sütunu

Aralık	**Male**	**Female**	**Toplam Adet**
0.01/0.04	10	712	722/3168
0.04/0.06	9	457	466/3168
0.06/0.09	140	179	319/3168
0.09/0.11	635	43	678/3168
0.11/0.13	614	33	647/3168
0.13/0.16	132	60	192/3168
0.16/0.18	35	71	106/3168
0.18/0.20	9	13	22/3168

0.20/0.23	0	6	6/3168
0.23/0.25	0	9	9/3168
Toplam	1584/3168	1583/3168	

Yukarıdaki tablolarda giriş değerlerinin aralıkları ve bu aralıklarda kaç adet kadın ve erkek olduğu gösterilmiştir. Bu tablolar Naive Bayes hesaplama yöntemi gösterilmesi için oluşturulmuştur. Bu bilgilere spyder üzerinden analiz yapılarak ulaşılmıştır.

Örnek Hesaplama 1:

No.	1: sd	2: Q25	3: IQR	4: meanfun	5: **label**
	Numeric	Numeric	Numeric	Numeric	Nominal
1	0.06...	0.01...	0.07...	0.08427...	male
2	0.06...	0.01...	0.07...	0.10793...	male
3	0.08...	0.00...	0.12...	0.09870...	male
4	0.07...	0.09...	0.11...	0.08896...	male
5	0.07...	0.07...	0.12...	0.10639...	male
6	0.07...	0.06...	0.14...	0.11013...	male
7	0.07...	0.09...	0.11...	0.10594...	male
8	0.07...	0.11...	0.12...	0.09305...	male
9	0.07...	0.08...	0.12...	0.09672...	male
10	0.08...	0.07...	0.12...	0.10588...	male
11	0.07...	0.10...	0.11...	0.08889...	male
12	0.07...	0.08...	0.11...	0.10419...	male
13	0.08...	0.08...	0.12...	0.09264...	male
14	0.06...	0.12...	0.10...	0.13150...	
15	0.06...	0.12...	0.11...	0.10279...	male
16	0.06...	0.11...	0.11...	0.10204...	male

14. satırda bulunan sırasıyla 0.06, 0.12, 0.10, 0.13 olan giriş değerlerinin female veya male olma durumunu Naive Bayes hesaplama yöntemi ile inceleyelim.

1.Adım: Giriş değerlerinin hangi aralıkta oldukları yukarıdaki tablolara bakılarak belirlenir.

2.Adım: Giriş değerlerinin aralıkları sırasıyla 0.06-0.07, 0.12-0.15, 0.09-0.11, 0.13-0.15'dir. Bu giriş değeri aralıklarında kaç adet kadın ve erkek olduğu yukarıdaki tablolara bakılarak belirlenir.

3.Adım: Elde edilen female ve male oranları formüle yazılarak hesaplanır.

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

P(A|B) ; B olayı gerçekleştiği durumda A olayının meydana gelme olasılığıdır

P(B|A) ; A olayı gerçekleştiği durumda B olayının meydana gelme olasılığıdır

P(A) ve P(B) ; A ve B olaylarının önsel olasılıklarıdır.

4.Adım: Formül: $P(c|X)= P(X_1|c) \times P(X_2|c) \times \ldots\ldots\ldots \times P(X_n|c) \times P(c)$

5.Adım: Elde edilen sonuçlardan en büyük olan değer formülde kullanılan sınıfı 14. Satırda Label çıktısı olarak belirler.

Female Formül ve İşlem

$X^{female}=\{0.06\text{-}0.07_{sd}, 0.12\text{-}0.15_{Q25}, 0.09\text{-}0.11_{IQR}, 0.13\text{-}0.15_{meanfun}\}$

P (female| x) =P (0.06/0.07 |female) x P (0.12/0.15 |female) x P (0.09/0.11 |female) x P(0.13/0.15|female) x P (x |female)

$İşlem^{female}$: 154/1584 x 112/1584 x 43/1584 x 193/1584 x 1584/3168

$Sonuç^{female}$: 0.0000113687784552

Male Formül ve İşlem

$X^{male}=\{0.06\text{-}0.07_{sd}, 0.12\text{-}0.15_{Q25}, 0.09\text{-}0.11_{IQR}, 0.13\text{-}0.15_{meanfun}\}$

P (male| x) =P (0.06/0.07 |male) x P (0.12/0.15 |male) x P (0.09/0.11 |male) x P(0.13/0.15|male) x P (x |male)

$İşlem^{male}$: 611/1584 x 737/1584 x 635/1584 x 308/1584 x 1584/3168

$Sonuç^{male}$: 0.0069949148995887

İki sonucu incelediğimizde;

$Sonuç^{male} > Sonuç^{female}$ olduğu görülmektedir.

Bu durumda 14. Satırın çıktı değeri Male'dir.

No.	1: sd	2: Q25	3: IQR	4: meanfun	5: label
	Numeric	Numeric	Numeric	Numeric	Nominal
1	0.06...	0.01...	0.07...	0.08427...	male
2	0.06...	0.01...	0.07...	0.10793...	male
3	0.08...	0.00...	0.12...	0.09870...	male
4	0.07...	0.09...	0.11...	0.08896...	male
5	0.07...	0.07...	0.12...	0.10639...	male
6	0.07...	0.06...	0.14...	0.11013...	male
7	0.07...	0.09...	0.11...	0.10594...	male
8	0.07...	0.11...	0.12...	0.09305...	male
9	0.07...	0.08...	0.12...	0.09672...	male
10	0.08...	0.07...	0.12...	0.10588...	male
11	0.07...	0.10...	0.11...	0.08889...	male
12	0.07...	0.08...	0.11...	0.10419...	male
13	0.08...	0.08...	0.12...	0.09264...	male
14	0.06...	0.12...	0.11...	0.10279...	male
15	0.06...	0.11...	0.11...	0.10204...	male

Veri setinin 14. Satırını incelediğimizde Label parametresinin çıkış değerinin Male olduğu doğrulanmaktadır.

Örnek Hesaplama 2:

No.	1: sd	2: Q25	3: IQR	4: meanfun	5: label
	Numeric	Numeric	Numeric		Nominal
1	0.04...	0.19...	0.01...	0.19...	female
2	0.01...	0.20...	0.01...	0.19...	female
3	0.04...	0.18...	0.01...	0.19...	female
4	0.03...	0.18...	0.01...	0.18...	female
5	0.02...	0.17...	0.01...	0.15...	female
6	0.03...	0.15...	0.01...	0.14...	female
7	0.04...	0.18...	0.01...	0.19...	female
8	0.03...	0.18...	0.01...	0.18...	female
9	0.02...	0.20...	0.01...	0.18...	female
10	0.02...	0.22...	0.01...	0.20...	female
11	0.02...	0.17...	0.01...	0.16...	female
12	0.03...	0.21...	0.01...	0.20...	female
13	0.02...	0.22...	0.01...	0.20...	female
14	0.04...	0.19...	0.01...	0.19...	female
15	0.04...	0.20...	0.01...	0.21...	female
16	0.03...	0.22...	0.01...	0.19...	female
17	0.03...	0.22...	0.01...	0.21...	female
18	0.03...	0.17...	0.01...	0.17...	female
19	0.03...	0.22...	0.01...	0.19...	female
20	0.03...	0.20...	0.01...	0.17...	

20. satırda bulunan sırasıyla 0.03, 0.20, 0.01, 0.17 olan giriş değerlerinin female veya male olma durumunu Naive Bayes hesaplama yöntemi ile inceleyelim.

Female Formül ve İşlem

X^{female}={0.03-0.04$_{sd}$, 0.20-0.22$_{Q25}$, 0.01-0.04$_{IQR}$, 0.16-0.18$_{meanfun}$}

P (female| x) =P (0.03/0.04 |female) x P (0.20/0.22 |female) x P (0.01/0.04 |female) x P(0.16/0.18|female) x P (x |female)

İşlem^{female}: 638/1584 x 207/1584 x 712/1584 x 653/1584 x 1584/3168

Sonuç^{female}: 0.0048767837073852

Male Formül ve İşlem

X^{male}={0.03-0.04$_{sd}$, 0.20-0.22$_{Q25}$, 0.01-0.04$_{IQR}$, 0.16-0.18$_{meanfun}$}

P (male| x) =P (0.03/0.04 |male) x P (0.20/0.22 |male) x P (0.01/0.04 |male) x P(0.16/0.18|male) x P (x |male)

İşlem^{male}: 0/1584 x 6/1584 x 10/1584 x 14/1584 x 1584/3168

Sonuç^{male}: 0

İki sonucu incelediğimizde;

Sonuç^{female} > Sonuç^{male} olduğu görülmektedir.

Bu durumda 20. Satırın çıktı değeri Female'dir.

No.	1: sd	2: Q25	3: IQR	4:	5: **label**
	Numeric	Numeric	Numeric	meanfun	Nominal
1	0.04...	0.19...	0.01...	0.19...	female
2	0.01...	0.20...	0.01...	0.19...	female
3	0.04...	0.18...	0.01...	0.19...	female
4	0.03...	0.18...	0.01...	0.18...	female
5	0.02...	0.17...	0.01...	0.15...	female
6	0.03...	0.15...	0.01...	0.14...	female
7	0.04...	0.18...	0.01...	0.19...	female
8	0.03...	0.18...	0.01...	0.18...	female
9	0.02...	0.20...	0.01...	0.18...	female
10	0.02...	0.22...	0.01...	0.20...	female
11	0.02...	0.17...	0.01...	0.16...	female
12	0.03...	0.21...	0.01...	0.20...	female
13	0.02...	0.22...	0.01...	0.20...	female
14	0.04...	0.19...	0.01...	0.19...	female
15	0.04...	0.20...	0.01...	0.21...	female
16	0.03...	0.22...	0.01...	0.19...	female
17	0.03...	0.22...	0.01...	0.21...	female
18	0.03...	0.17...	0.01...	0.17...	female
19	0.03...	0.22...	0.01...	0.19...	female
20	0.03...	0.20...	0.01...	0.17...	female

Veri setinin 20. Satırını incelediğimizde Label parametresinin çıkış değerinin Female olduğu doğrulanmaktadır.

Örnek Hesaplama 3:

No.	1: sd Numeric	2: Q25 Numeric	3: IQR Numeric	4: meanfur	5: label Nominal
1	0.07...	0.11...	0.11...	0.05...	male
2	0.07...	0.08...	0.15...	0.05...	male
3	0.07...	0.08...	0.15...	0.06...	male
4	0.07...	0.11...	0.11...	0.06...	male
5	0.05...	0.14...	0.09...	0.06...	male
6	0.09...	0.08...	0.13...	0.06...	male
7	0.05...	0.15...	0.07...	0.06...	
8	0.09...	0.03...	0.16...	0.06...	male
9	0.07...	0.09...	0.12...	0.06...	male
10	0.07...	0.09...	0.14...	0.06...	male

7. satırda bulunan sırasıyla 0.05, 0.15, 0.07, 0.06 olan giriş değerlerinin female veya male olma durumunu Naive Bayes hesaplama yöntemi ile inceleyelim.

Female Formül ve İşlem

X^{female}={0.05-0.06$_{sd}$, 0.15-0.17$_{Q25}$, 0.06-0.09$_{IQR}$, 0.06-0.07$_{meanfun}$}

P (female| x) =P (0.05/0.06 |female) x P (0.15/0.17 |female) x P (0.06/0.09 |female) x P(0.06/0.07|female) x P (x |female)

İşlem^{female}: 186/1584 x 334/1584 x 179/1584 x 0/1584 x 1584/3168

Sonuç^{female}: 0

Male Formül ve İşlem

X^{male}={0.05-0.06$_{sd}$, 0.15-0.17$_{Q25}$, 0.06-0.09$_{IQR}$, 0.06-0.07$_{meanfun}$ }

P (male| x) =P (0.05/0.06 |male) x P (0.15/0.17 |male) x P (0.06/0.09 |male) x P(0.06/0.07|male) x P (x |male)

İşlem^{male}: 506/1584 x 77/1584 x 140/1584 x 10/1584 x 1584/3168

Sonuç^{male}: 0.0000043323005018

İki sonucu incelediğimizde;

Sonuç^{male} > Sonuç^{female} olduğu görülmektedir.

Bu durumda 7. Satırın çıktı değeri Male'dir.

No.	1: sd	2: Q25	3: IQR	4: meanfur	5: label
	Numeric	Numeric	Numeric		Nominal
1	0.07...	0.11...	0.11...	0.05...	male
2	0.07...	0.08...	0.15...	0.05...	male
3	0.07...	0.08...	0.15...	0.06...	male
4	0.07...	0.11...	0.11...	0.06...	male
5	0.05...	0.14...	0.09...	0.06...	male
6	0.09...	0.08...	0.13...	0.06...	male
7	0.05...	0.15...	0.07...	0.06...	male
8	0.09...	0.03...	0.16...	0.06...	male
9	0.07...	0.09...	0.12...	0.06...	male
10	0.07...	0.09...	0.14...	0.06...	male

Veri setinin 7. Satırını incelediğimizde Label parametresinin çıkış değerinin Male olduğu doğrulanmaktadır.

3.4.3.Ses ve Cinsiyet Tanıma Veri Setini K-NN Formülü İle Örnek Hesaplama

a) K'nın belirlenmesi: k=3 kabul edilir.

Sıra	SD	Q25	IQR	MEANFUN	LABEL
1	0.06	0.01	0.07	0.08	male
2	0.08	0.04	0.17	0.13	male
3	0.03	0.17	0.02	0.17	female
4	0.02	0.22	0.02	0.19	female
5	0.06	0.11	0.08	0.16	male
6	0.05	0.14	0.06	0.14	male
7	0.05	0.19	0.02	0.18	female
8	0.04	0.21	0.02	0.11	male
9	0.06	0.09	0.10	0.10	male
10	0.01	0.20	0.01	0.19	female
11	0.05	0.16	0.01	0.14	?

b) Uzaklıkların hesaplanması: (0.05,0.16,0.01,0.14) noktası ile gözlem değerlerinin her biri arasındaki uzaklıklar Öklid uzaklığına göre hesaplanır.

1.Satır $\sqrt{(0.06-0.05)^2+(0.01-0.16)^2+(0.07-0.01)^2+(0.08-0.14)^2}$=0.172

2.Satır $\sqrt{(0.08-0.05)^2+(0.04-0.16)^2+(0.17-0.01)^2+(0.13-0.14)^2}$=0.202

3.Satır $\sqrt{(0.03-0.05)^2+(0.17-0.16)^2+(0.02-0.01)^2+(0.17-0.14)^2}$=0.038

4.Satır $\sqrt{(0.02-0.05)^2+(0.22-0.16)^2+(0.02-0.01)^2+(0.19-0.14)^2}$=0.084

5.Satır $\sqrt{(0.06-0.05)^2+(0.11-0.16)^2+(0.08-0.01)^2+(0.16-0.14)^2}$=1.227

6.Satır $\sqrt{(0.05-0.05)^2+(0.14-0.16)^2+(0.06-0.01)^2+(0.14-0.14)^2}$=0.053

7.Satır $\sqrt{(0.05-0.05)^2+(0.19-0.16)^2+(0.02-0.01)^2+(0.18-0.14)^2}$=0.050

8.Satır $\sqrt{(0.04-0.05)^2+(0.21-0.16)^2+(0.02-0.01)^2+(0.11-0.14)^2}$=0.06

9.Satır $\sqrt{(0.06-0.05)^2+(0.09-0.16)^2+(0.10-0.01)^2+(0.10-0.14)^2}$=0.121

10.Satır $\sqrt{(0.01-0.05)^2+(0.20-0.16)^2+(0.01-0.01)^2+(0.19-0.14)^2}$=0.075

c) En küçük uzaklıkların belirlenmesi: Satırlar sıralanarak en küçük k=3 tanesi belirlenir. Bu üç nokta verilen (0.05,0.16,0.01,0.14) noktasına en yakın gözlem değerleridir.

SD	Q25	IQR	MEANFUN	Uzaklık	Sıralama
0.06	0.01	0.07	0.08	0.172626	
0.08	0.04	0.17	0.13	0.202484	
0.03	0.17	0.02	0.17	0.038729	1
0.02	0.22	0.02	0.19	0.084261	
0.06	0.11	0.08	0.16	1.227105	
0.05	0.14	0.06	0.14	0.053851	3
0.05	0.19	0.02	0.18	0.050990	2
0.04	0.21	0.02	0.11	0.06	4
0.06	0.09	0.10	0.10	0.121243	
0.01	0.20	0.01	0.19	0.075498	

d) Seçilen satırların ilişkin sınıfların belirlenmesi: (0.05,0.16,0.01,0.14) noktasına en yakın olan gözlem değerlerinin Label sınıfları göz önüne alınır ve içinde hangi değerin baskın olduğu araştırılır. Bu üç gözlem içinde 2 tane “Female” 1 tane “Male” sınıfı vardır.

SD	Q25	IQR	MEANFUN	Uzaklık	Sıralama	Label
0.06	0.01	0.07	0.08	0.172626		

0.08	0.04	0.17	0.13	0.202484		
0.03	0.17	0.02	0.17	0.038729	1	female
0.02	0.22	0.02	0.19	0.084261		
0.06	0.11	0.08	0.16	1.227105		
0.05	0.14	0.06	0.14	0.053851	3	male
0.05	0.19	0.02	0.18	0.050990	2	female
0.04	0.21	0.02	0.11	0.06		
0.06	0.09	0.10	0.10	0.121243		
0.01	0.20	0.01	0.19	0.075498		

e) Yeni gözlemin sınıfı: Female değerlerinin sayısı Male değerlerinin sayısından fazla olduğu için (0.05,0.16,0.01,0.14) noktasının sınıfı Female olarak belirlenir.

Yaptığımız hesaplamayı doğrulayalım;

No.	1: sd Numeric	2: Q25 Numeric	3: IQR Numeric	4: meanfur	5: **label** Nominal
1	0.04...	0.19...	0.01...	0.19...	female
2	0.01...	0.20...	0.01...	0.19...	female
3	0.04...	0.18...	0.01...	0.19...	female
4	0.03...	0.18...	0.01...	0.18...	female
5	0.02...	0.17...	0.01...	0.15...	female
6	0.03...	0.15...	0.01...	0.14...	female
7	0.04...	0.18...	0.01...	0.19...	female
8	0.03...	0.18...	0.01...	0.18...	female
9	0.02...	0.20...	0.01...	0.18...	female
10	0.02...	0.22...	0.01...	0.20...	female
11	0.02...	0.17...	0.01...	0.16...	female
12	0.03...	0.21...	0.01...	0.20...	female
13	0.02...	0.22...	0.01...	0.20...	female
14	0.04...	0.19...	0.01...	0.19...	female
15	0.04...	0.20...	0.01...	0.21...	female
16	0.03...	0.22...	0.01...	0.19...	female
17	0.03...	0.22...	0.01...	0.21...	female
18	0.03...	0.17...	0.01...	0.17...	female
19	0.03...	0.22...	0.01...	0.19...	female
20	0.03...	0.20...	0.01...	0.17...	female
21	0.04...	0.19...	0.01...	0.19...	female
22	0.03...	0.16...	0.01...	0.16...	female
23	0.03...	0.16...	0.01...	0.17...	female
24	0.04...	0.19...	0.01...	0.19...	female
25	0.03...	0.17...	0.01...	0.15...	female
26	0.05...	0.15...	0.01...	0.15...	female
27	0.05...	0.16...	0.01...	0.14...	female

Yaptığımız kümele hesaplamasını doğru tahmin ettiğimizi görmekteyiz.

Kümeyi 4 olarak belirlersek;

SD	Q25	IQR	MEANFUN	Uzaklık	Sıralama	Label

0.06	0.01	0.07	0.08	0.172626		
0.08	0.04	0.17	0.13	0.202484		
0.03	0.17	0.02	0.17	0.038729	1	female
0.02	0.22	0.02	0.19	0.084261		
0.06	0.11	0.08	0.16	1.227105		
0.05	0.14	0.06	0.14	0.053851	3	male
0.05	0.19	0.02	0.18	0.050990	2	female
0.04	0.21	0.02	0.11	0.06	4	male
0.06	0.09	0.10	0.10	0.121243		
0.01	0.20	0.01	0.19	0.075498		

Ağırlıklı Oylama Hesaplama

Formül: $d(i,j)' = \frac{1}{d(i,j)^2}$

1.Sıralama^{female} $d(0.038)' = \frac{1}{d(0.038)^2}$ = 693

2.Sıralama^{female} $d(0.050)' = \frac{1}{d(0.050)^2}$ = 400

3.Sıralama^{male} $d(0.053)' = \frac{1}{d(0.053)^2}$ = 356

4.Sıralama^{male} $d(0.06)' = \frac{1}{d(0.06)^2}$ = 278

SD	Q25	IQR	MEANFUN	Uzaklık	Sıralama	Label	Ağırlıklı Oylama
0.06	0.01	0.07	0.08	0.172626			
0.08	0.04	0.17	0.13	0.202484			
0.03	0.17	0.02	0.17	0.038729	1	female	693
0.02	0.22	0.02	0.19	0.084261			
0.06	0.11	0.08	0.16	1.227105			
0.05	0.14	0.06	0.14	0.053851	3	male	356
0.05	0.19	0.02	0.18	0.050990	2	female	400
0.04	0.21	0.02	0.11	0.06	4	male	278

0.06	0.09	0.10	0.10	0.121243			
0.01	0.20	0.01	0.19	0.075498			

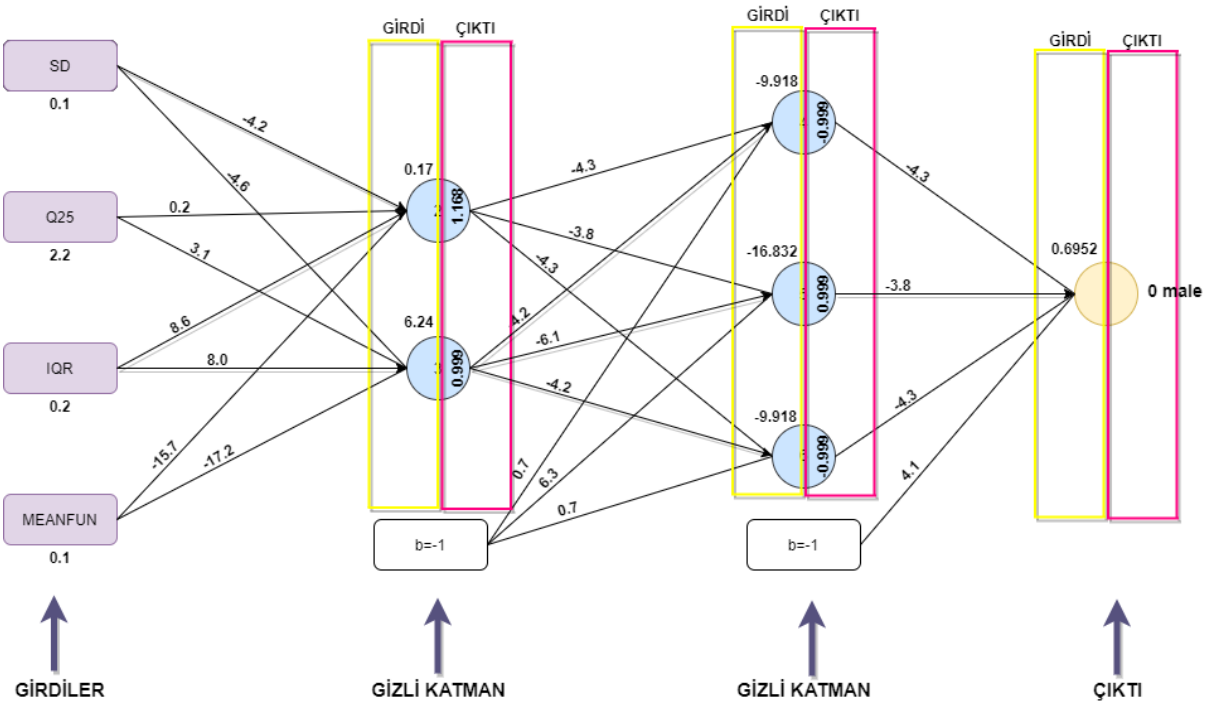
Toplam^{female}= 693+ 400 = 1093

Toplam^{male}= 356+ 278 = 634

Toplam^{female} > Toplam^{male}

Female ağırlıklı oylama değeri Male ağırlıklı oylama sınıfından yüksek olduğu için (0.05,0.16,0.01,0.14) noktasının sınıfı Female olarak belirlenir.

3.4.4.Ses ve Cinsiyet Tanıma Veri Setini Yapay Sinir Ağları Formülü İle Örnek Hesaplama



Not: Hazırladığımız örnekte bulunan ağırlık değerleri Weka yazılımından elde edilmiştir.

Nöron: Temel biyoloji teriminde nöronlar, aksonları boyunca dendritlerden bir uçtan diğer uca bir elektrik sinyali gönderir. Bu sinyaller daha sonra başka bir nörona geçirilir. Bu işlemler sinir sistemi boyunca iletilerek bilgilerin beyne iletilmesini sağlar. Benzer şekilde de yapay sinir ağlarındaki nöronlar elde ettiği bilgileri diğer nöronlara taşır. Böylece sistemin giriş değerlerine göre çıkış verilerini öğrenmesini sağlar.

Ağ: Nöronların birbirine bağlı olduğu graf yapılardır.

Katman: Farklı düzeylerde yer alan nöron gruplarıdır. Yapay sinir ağları üç ana katmandan oluşmaktadır. Bunlar sırasıyla;

- **Giriş Katmanı:** Sisteme giriş olarak gelen veriler bu katmanda yer alır. Bu katmanda giriş verileri üzerinde hiçbir değişiklik yapmadan bir sonraki katman olan hidden (gizli) katmana aktarır.
- **Gizli Katman:** Verinin transfer edildiği katmandır. Öğrenme bu katmanda olur.

Çıkış Katmanı: Sistemin giriş verilerine göre öğrenmesini istenilen çıkış değerleri burada yer alır. Sistem çıktısının alındığı yerdir.

1.Girdi İşlemi

W.X= $w_1x_1 + w_2x_2 + w_3x_3 + \dots + w_mx_m$

Node2= (-4.2 x 0.1) + (0.2 x 2.2) + (8.6 x 0.2) + (-15.7 x 0.1) = 0.17

Node3= (-4.6 x 0.1) + (3.1 x 2.2) + (8 x 0.2) + (-17.2 x 0.1) = 6.24

1.Çıktı İşlemi

Node2= $\frac{e^{(x)} - e^{-(x)}}{e^{(x)} + e^{-(x)}} = \frac{e^{(0.17)} - e^{-(0.17)}}{e^{(0.17)} + e^{-(0.17)}}$ = 0.168

Node3= $\frac{e^{(x)} - e^{-(x)}}{e^{(x)} + e^{-(x)}} = \frac{e^{(6.24)} - e^{-(6.24)}}{e^{(6.24)} + e^{-(6.24)}}$ = 0.999

2.Girdi İşlemi

Node4= (0.7 x -1) + (-4.3 x 1.168) + (-4.2 x 0.999) = -9.918

Node5= (6.3 x -1) + (-3.8 x 1.168) + (-6.1 x 0.999) = -16.832

Node6= (0.7 x -1) + (-4.3 x 1.168) + (-4.2 x 0.999) = -9.918

2.Çıktı İşlemi

Node4= $\frac{e^{(x)} - e^{-(x)}}{e^{(x)} + e^{-(x)}} = \frac{e^{(-9.918)} - e^{-(-9.918)}}{e^{(-9.918)} + e^{-(-9.918)}}$ = -0.999

Node5= $\frac{e^{(x)} - e^{-(x)}}{e^{(x)} + e^{-(x)}} = \frac{e^{(-16.832)} - e^{-(-16.832)}}{e^{(-16.832)} + e^{-(-16.832)}}$ = 0.999

Node6= $\frac{e^{(x)} - e^{-(x)}}{e^{(x)} + e^{-(x)}} = \frac{e^{(-9.918)} - e^{-(-9.918)}}{e^{(-9.918)} + e^{-(-9.918)}}$ = -0.999

3.Girdi İşlemi

Node0= (4.1 x -1) + (-0.999 x -4.3) + (-0.999 x -3.8) + (-0.999 x -4.3) = 0.695

3.Çıktı İşlemi

Aktivasyon Fonksiyonu: Nörona gelen bilginin bir sonraki nörona iletilip iletilmeyeceğine karar veren birimdir.

Aktivasyon fonksiyonu: $y = \begin{cases} 0, & x < \tau \\ 1, & x \geq \tau \end{cases}$

$H_0(x^{(i)})=f(0.695)=0$

SD	Q25	IQR	MEANFUN	LABEL
0.1	2.2	0.2	0.1	male(0)

Çıktı doğru bulunduğu için bir sonraki iterasyona gerek kalmaz eğer çıktı yanlış bulunsaydı iterasyon doğruyu bulana kadar tekrar ederdi. Çıktının 1 (female) olarak çıktığını farz edersek bu durumda ağın ürettiği hata hesaplanarak ağırlık değerleri güncellenecekti. Elde edilen yeni ağırlık değerleri ile bir alt satırdaki veriler tekrar hesaplanacaktı sonuç tekrar yanlış bulunduysa bu işlem doğruyu bulana kadar tekrar edecekti. Son iterasyona gelindiğinde doğru bulunan satırdaki ağırlık değerleri temel alınarak doğrusal modelin denklemi oluşturulur.

3.5.UYGULAMA

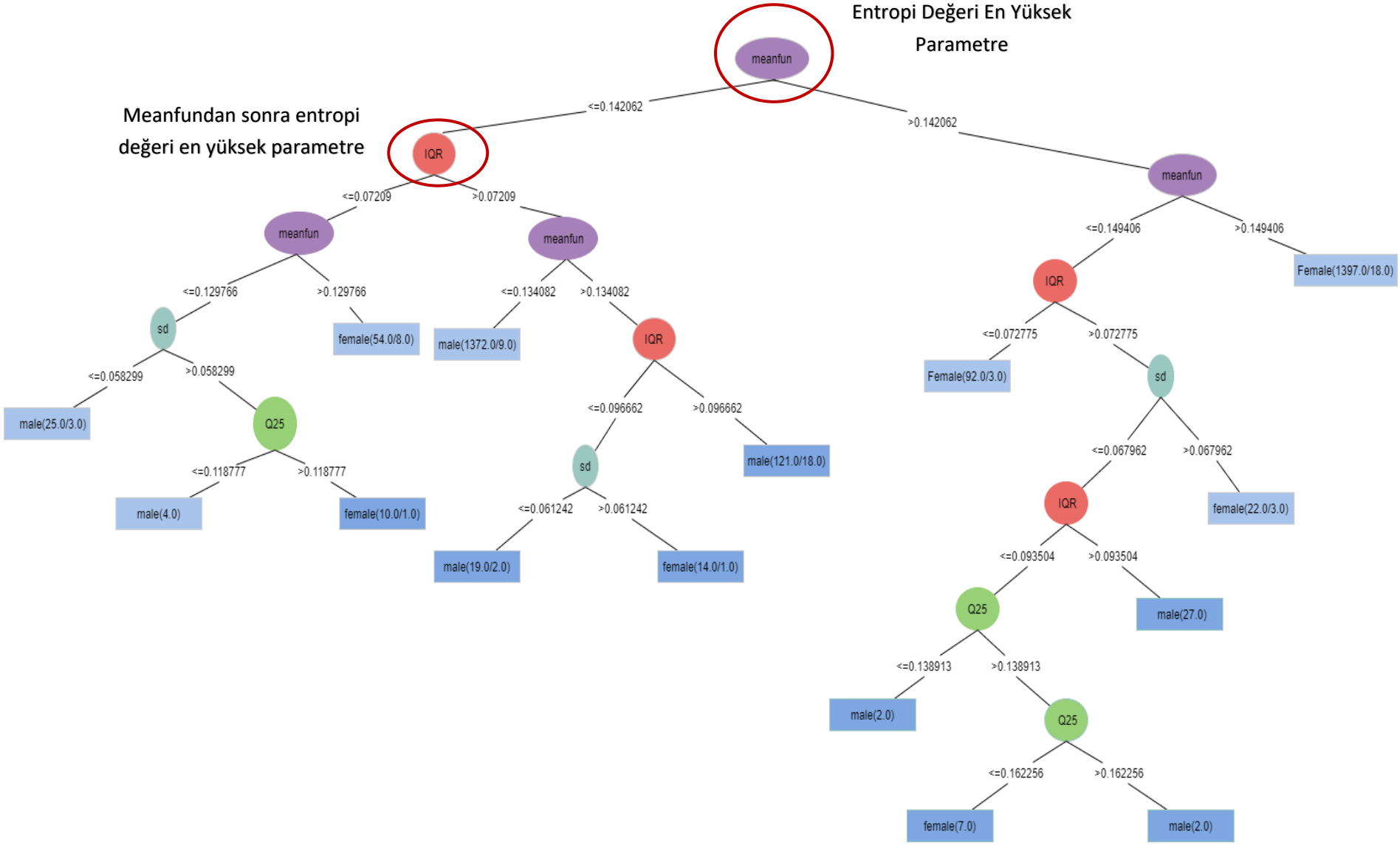
3.5.1.Karar Ağacı Sınıflandırma Yöntemi

Örnek1

Veri setimizin içerisinden meanfun, sd, Q25, IQR, label parametrelerini temel alarak ilerledik. Karar Ağacı algoritmasından makineye veri setinin sonucu bildiğimiz değerlerin %60'ını eğitim vererek %40 unu tahmin etmesini istedik. Label parametresi üzerinden de sınıflandırmasını yaptık. Yaprak başına düşen minimum obje (MinNumObj) sayısı 2 (varsayılan değer) olarak belirlenmiştir.

Entropi Değeri En Yüksek Parametre

Meanfundan sonra entropi değeri en yüksek parametre



Karar ağacını incelediğimizde dallanmanın ilk olarak “Meafun” parametresinden başlandığı görülmektedir. İlk koşulun “Meafun” parametresi olma sebebi entropi değerinin yüksek olmasıdır. Modellediğimiz karar ağacı 17 yaprak, 33 daldan oluşmaktadır.

Sonuç1

Başarı Oranı	***97.0008 %***
Doğru Sınıflandırılmış Örnekler	1229
Kappa İstatistiği	0.94
Ortalama Mutlak Hata	0.0446
Kök Ortalama Kare Hatası	0.1626
Göreli Mutlak Hata	8.9196 %
Kök Göreli Kare Hatası	32.9196 %
Toplam Örnek Sayısı	1267

Modeli oluşturmak için geçen süre: 0,03 saniyedir.

Correctly Classified Instances

Veri setinin %40’ı test kümesi olarak ele alındığı için 3168 satır veriden 1267’si üzerinde tahmin yapılmıştır. Bunun sonucunda 1267 adet veriden 1229’si doğru tahmin ederek %97.0008 oranında başarı elde edilmiştir. Modelimiz tahmin yaparken 38 adet veriyi yanlış bulmuştur.

Kappa Statistic

Kadın ve erkek değerlerinin arasındaki karşılaştırmalı uyuşmanın güvenilirlik oranı 0.94’dür. Elde edilen değer 0.81-1. 00 aralığında olduğu için kadın ve erkek değerleri arasında neredeyse mükemmel bir uyuşma olduğu görülmektedir.

Mean Absolute Error

Tahmin sonucunda elde edilen ortalama mutlak hata oranı 0.0446’dır. Bu değer 0’a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Mean Squared Error

Tahmin sonucunda elde edilen karekök ortalama hata oranı 0.1626’dır. Bu değer 0’a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Relative Absolute Error

Tahmin sonucunda gerçek değer ile hesaplanan değer arasındaki farkın gerçek değere oranlanması sonucunda 8.9196 değeri elde edilmiştir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Relative Squared Error

Tahmin sonucunda kök göreceli hata oranı 32,5157'dir. Bu değer 0'a çok yakın olmasa da diğer metriklerin 0'a yakın olmasından dolayı modelin başarı oranını çok etkilememiştir.

	TP Rate	***FP Rate***	***Presicion***	***Recall***	***F-Measure***	***MCC***	***ROC Area***	***PRC Area***	***Class***
	0.964	0.024	0.976	0.964	0.970	0.940	0.978	0.977	Male
	0.976	0.036	0.964	0.976	0.970	0.940	0.978	0.970	Female
Ağırlıklı Ortalama	0.970	0.030	0.970	0.970	0.970	0.940	0.978	0.974	

TP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini erkek sesi olarak tahmin oranı 0,964'dür. Sınıfı kadın olan verilerden kadın sesini kadın sesi olarak tahmin oranı 0,976'dir. Bu değerler 1'e yakın olduğu için iyi bir isabet oranı elde edildiği görülmektedir. Yapılan karar ağacı sınıflandırma yöntemi sonucunda verilerin tamamına yakın bir kısmını doğru tahmin ettiği görülmektedir. Aynı zamanda kadın ve erkek sınıflarında en çok doğru tahmini yapan kadın sınıfıdır.

FP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini kadın sesi olarak tahmin etme oranı 0,024'dür. Sınıfı kadın olan verilerden kadın sesini erkek sesi olarak tahmin etme oranı 0,036'dır. Bu değerler 0'a yakın olduğu için yapılan hatalı tahminin çok az olduğu görülmektedir. Aynı zamanda erkek ve kadın sınıfları arasında en çok hata yapan kadın sınıfıdır.

Precision

Tahmin sonucunda erkek sınıfı hassasiyet oranı 0,976 iken kadın sınıfı hassasiyet oranı 0,964'dür. Hassasiyet oranları 1'e yakın olduğu için sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Sınıfı erkek olan verilerin hassasiyet oranı daha yüksektir.

Recall

Tahmin sonucunda erkek sınıfı geri çağırma oranı 0,964 iken kadın sınıfı geri çağırma oranı 0,976'dir. Örneğin gerçekte sesin kadın olduğu durumda tahminin erkek sesi olarak yapılmasıdır. Bu hata 0'a yaklaştıkça artar 1'e yaklaştıkça azalır. Bu durumda sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Kadın sınıfına ait veriler daha doğru tahmin edilmiştir.

F-Measure

Tahmin sonucunda erkek ve kadın sınıfı F-Measure oranı 0,970 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

MCC

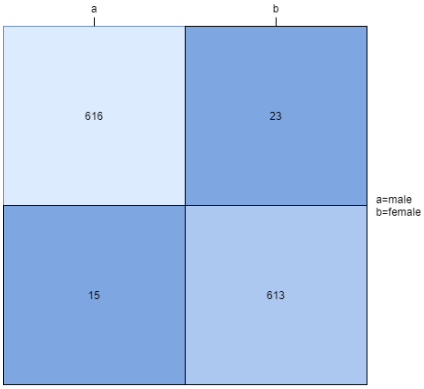
Tahmin sonucunda erkek ve kadın sınıfı MCC oranı 0,940 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

ROC Area

Tahmin sonucunda erkek ve kadın sınıfı ROC Area oranı 0,978 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

PRC Area

Tahmin sonucunda erkek sınıfı PRC Area oranı 0,977 iken kadın sınıfı oranı 0,970 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir. Aynı zamanda erkek sınıfının PRC Area oranı kadın sınıfından daha başarılıdır.

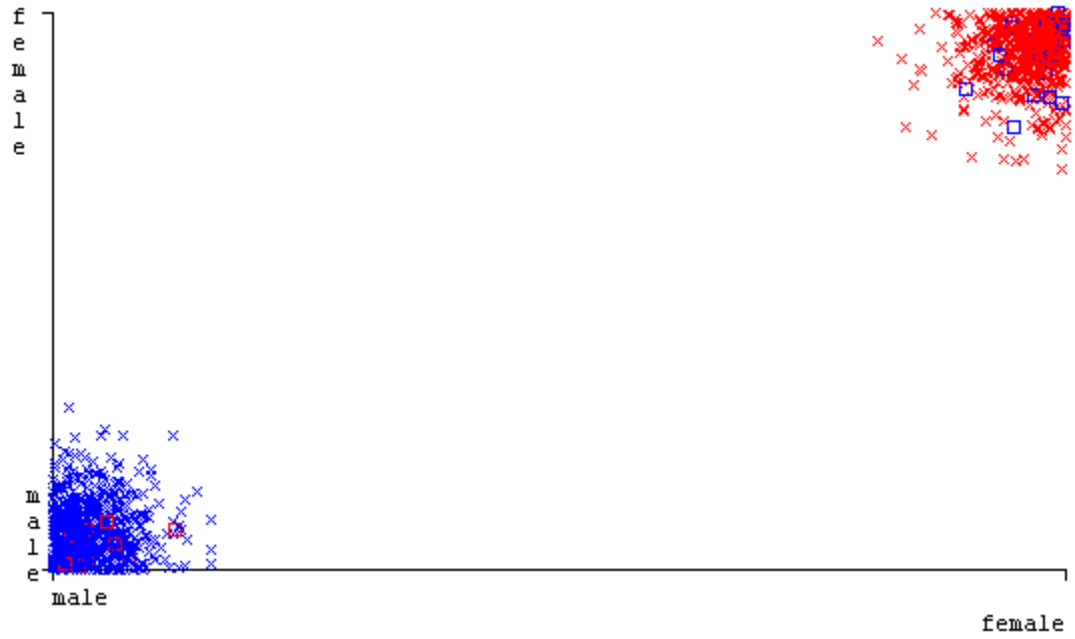


Confusion matrix kısmında oluşturduğumuz test ve eğitim kümeleri sonucunda label parametremizde bulunan male ve famele tahminlerindeki başarımızı görmekteyiz. Confusion matrixde male ifadesini a olarak, female ifadesini b olarak ele almaktayız. A olarak ele aldığımız Male olan 639 verinin 616 tanesini male olarak doğru tahmin ederken 23 tanesini female olarak yanlış tahmin etmiştir. B olarak ele aldığımız female olan 628 verinin 613 tanesini female olarak doğru tahmin ederken 15 tanesini male olarak yanlış tahmin etmiştir. Tüm veri setine baktığımız zamanda 1267 veriden 1229 veriyi doğru tahmin ederken 38 veriyi yanlış tahmin etmiştir.

Yanlış Tahmin Edilen Verilerin Grafik Olarak Gösterimi

Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin birleşimi gösterilmiştir. Sınıflandırıcı hata grafiğinde "female" sınıfı kırmızı, "male" sınıfı mavi renktedir. Grafikte ise "**x**" ile gösterilenler doğru sınıflandırılmış verileri "**kare**" şeklinde gösterilen yerler ise yanlış

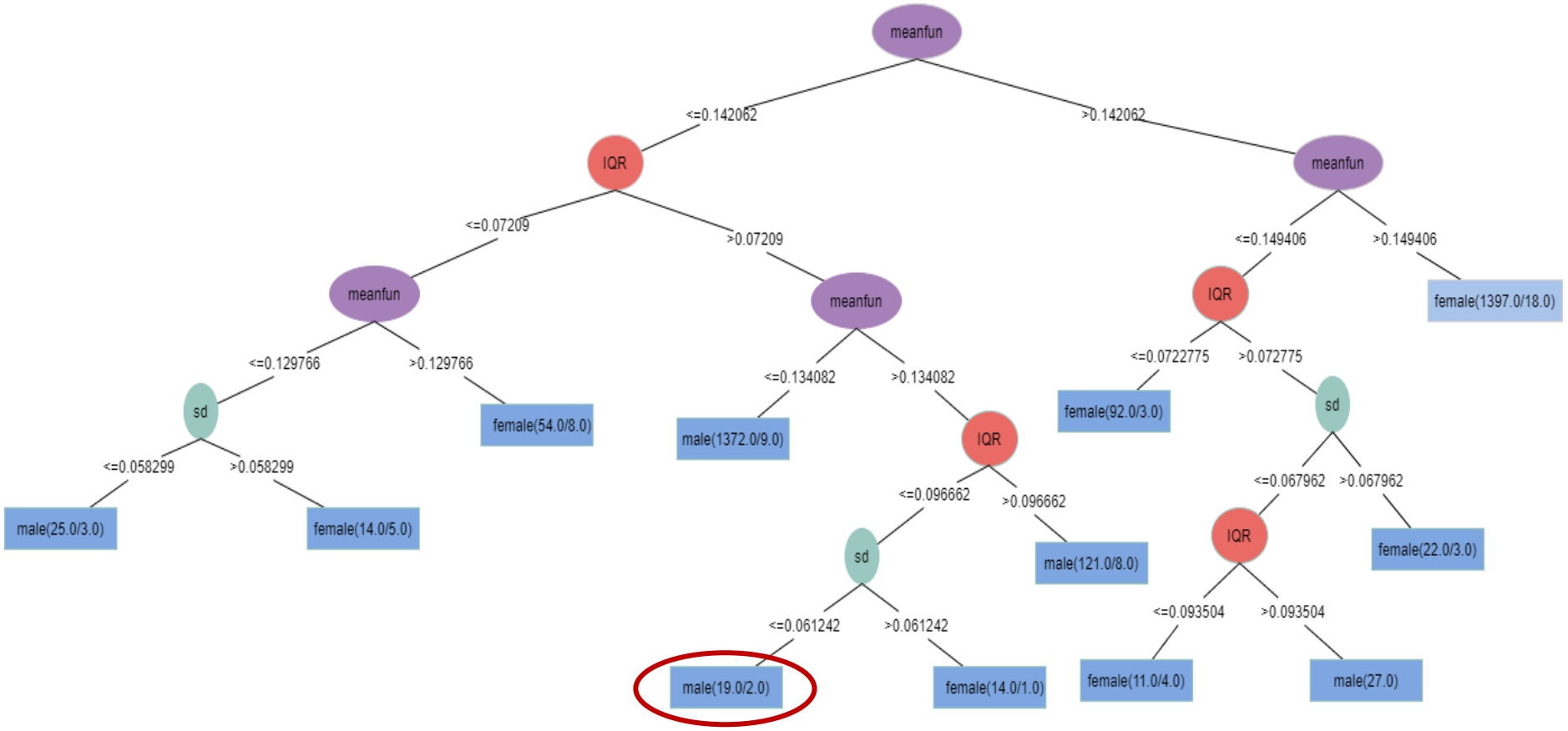
sınıflandırılmış verileri ifade ediyor. Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin birleşimi gösterilmiştir.



Örneğin **female** kısmında bulunan **mavi** bir kare bu değerin male sınıfına ait olduğunu ama yanlış şekilde sınıflandırılarak female sınıfına dâhil edildiğini gösteriyor. Tahmin modelimizin sınıflandırıcı hata grafiğine baktığımızda 23 male 15 female olmak üzere 38 hata yapıldığı görülmekte. 23 erkek çıktı kadın olarak tahmin edildiği ve 15 kadın çıktı ise erkek olarak tahmin edildiği görülmektedir. Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin kesişimi gösterilmiştir.

Örnek2

Weka programının Karar Ağacı sınıflandırma yönteminde karışık veri setlerinde kullanılan ve rastgele doğru ve yanlış çıkma olasılığını düşüren cross-validation seçeneğini 4 seçerek veri setimizi dörde bölüyoruz. Bu aşamada 729 adet veriden oluşan 4 adet veri seti oluşmaktadır. Daha sonra bu 4 adet veri setinden bir tanesi seçilerek test kümesi oluşturulur. Geriye kalan 2367 adet veri eğitim kümesi olarak belirlenmektedir. Oluşturulan eğitim kümesi ile model eğitilir ve 729 adet veri bulunan test kümesi ile test edilir. Kısaca cross-validation seçeneği 4 (%25 eğitim/ %75 test) seçilerek veri setinde temel alınan Meanfun, IQR, Sd, Q25, Label parametreleri üzerinde tahmin yapılmıştır. Model kurulurken yapraktaki minimum obje sayısı 10 olarak belirlenmiştir.


meanfun
<=0.142062
>0.142062
IQR
meanfun
<=0.07209
>0.07209
<=0.149406
>0.149406
meanfun
meanfun
IQR
female(1397.0/18.0)
<=0.129766
>0.129766
<=0.134082
>0.134082
<=0.0722775
>0.072775
sd
female(54.0/8.0)
male(1372.0/9.0)
IQR
female(92.0/3.0)
sd
<=0.058299
>0.058299
<=0.096662
>0.096662
<=0.067962
>0.067962
male(25.0/3.0)
female(14.0/5.0)
sd
male(121.0/8.0)
IQR
female(22.0/3.0)
<=0.061242
>0.061242
<=0.093504
>0.093504
male(19.0/2.0)
female(14.0/1.0)
female(11.0/4.0)
male(27.0)

Yaprak başına düşen minimum obje sayısı 10 olarak belirlendiği için dallanma sınırlanmıştır. Örneğin female 19 dan 2'sini yanlış bilmiş dallanma 10 ile sınıflanmasaydı 2 kez daha bölünebilirdi. Bir sonraki bölünme 2 olsaydı 9,5'da kalacaktı bu da belirtilen minimum obje sayısının altında kalacaktı. Bu sebeple dallanma 19 da kalmıştır. Bu yöntem karmaşık veri setleri için iyi bir alternatifken karmaşık olmayan setler için pek tercih edilmez. Karmaşık olmayan veri setlerinde minimum obje sayısı arttırıldıkça hata oranı artma eğilimi gösterebilir.

Sonuç2

Not: Sınıflandırma yönteminde aynı parametreler temel alındığı için entropi tekrar anlatılmamıştır.

Başarı Oranı	***96.6856 %***
Doğru Sınıflandırılmış Örnekler	3063
Kappa İstatistiği	0.9332
Ortalama Mutlak Hata	0.0477
Kök Ortalama Kare Hatası	0.1692
Göreli Mutlak Hata	9.5454 %
Kök Göreli Kare Hatası	33.8499 %
Toplam Örnek Sayısı	3168

Modeli oluşturmak için geçen süre: 0,01 saniyedir.

Correctly Classified Instances

Veri setinin %4'ü yani 792'si test kümesi kalan 2376'sı eğitim kümesi olarak ele alınarak tahmin yapılmıştır. Bunun sonucunda 3168 adet veriden 3063'ü doğru tahmin ederek %96.6856 oranında başarı elde edilmiştir. Modelimiz tahmin yaparken 105 adet veriyi yanlış bulmuştur.

Kappa Statistic

Kadın ve erkek değerlerinin arasındaki karşılaştırmalı uyuşmanın güvenilirlik oranı 0.9337'dir. Elde edilen değer 0.81-1.00 aralığında olduğu için kadın ve erkek değerleri arasında neredeyse mükemmel bir uyuşma olduğu görülmektedir.

Mean Absolute Error

Tahmin sonucunda elde edilen ortalama mutlak hata oranı 0.0477'dir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Mean Squared Error

Tahmin sonucunda elde edilen karekök ortalama hata oranı 0.1692'dır. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Relative Absolute Error

Tahmin sonucunda gerçek değer ile hesaplanan değer arasındaki farkın gerçek değere oranlanması sonucunda 9.5454 değeri elde edilmiştir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Relative Squared Error

Tahmin sonucunda kök göreceli hata oranı 33.8499'dur. Bu değer 0'a çok yakın olmasa da diğer metriklerin 0'a yakın olmasından dolayı modelin başarı oranını çok etkilememiştir.

	TP Rate	***FP Rate***	***Presicion***	***Recall***	***F-Measure***	***MCC***	***ROC Area***	***PRC Area***	***Class***
	0.961	0.027	0.973	0.961	0.967	0.934	0.987	0.985	Male
	0.973	0.039	0.961	0.973	0.967	0.934	0.987	0.981	Female
Ağırlıklı Ortalama	0.967	0.033	0.967	0.967	0.967	0.934	0.987	0.983	

TP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini erkek sesi olarak tahmin oranı 0,961'dür. Sınıfı kadın olan verilerden kadın sesini kadın sesi olarak tahmin oranı 0,973'dir. Bu değerler 1'e yakın olduğu için iyi bir isabet oranı elde edildiği görülmektedir. Yapılan Karar Ağacı sınıflandırma yöntemi sonucunda verilerin tamamına yakın bir kısmını doğru tahmin ettiği görülmektedir. Aynı zamanda kadın ve erkek sınıflarında en çok doğru tahmini yapan kadın sınıfıdır.

FP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini kadın sesi olarak tahmin etme oranı 0,027'dir. Sınıfı kadın olan verilerden kadın sesini erkek sesi olarak tahmin etme oranı 0,039' dur. Bu değerler 0'a yakın olduğu için yapılan hatalı tahminin çok az olduğu görülmektedir. Aynı zamanda erkek ve kadın sınıfları arasında en çok hata yapan kadın sınıfıdır.

Precision

Tahmin sonucunda erkek sınıfı hassasiyet oranı 0,973 iken kadın sınıfı hassasiyet oranı 0,961'dır. Hassasiyet oranları 1'e yakın olduğu için sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Sınıfı erkek olan verilerin hassasiyet oranı daha yüksektir.

Recall

Tahmin sonucunda erkek sınıfı geri çağırma oranı 0,961 iken kadın sınıfı geri çağırma oranı 0,973'dir. Örneğin gerçekte sesin kadın olduğu durumda tahminin erkek sesi olarak yapılmasıdır. Bu hata 0'a yaklaştıkça artar 1'e yaklaştıkça azalır. Bu durumda sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Kadın sınıfına ait veriler daha doğru tahmin edilmiştir.

F-Measure

Tahmin sonucunda erkek ve kadın sınıfı F-Measure oranı 0,967 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

MCC

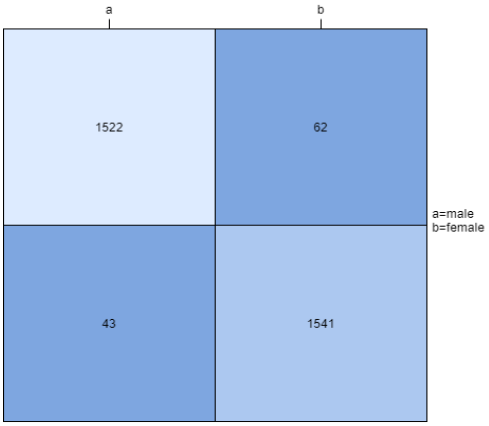
Tahmin sonucunda erkek ve kadın sınıfı MCC oranı 0,934 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

ROC Area

Tahmin sonucunda erkek ve kadın sınıfı ROC Area oranı 0,987 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

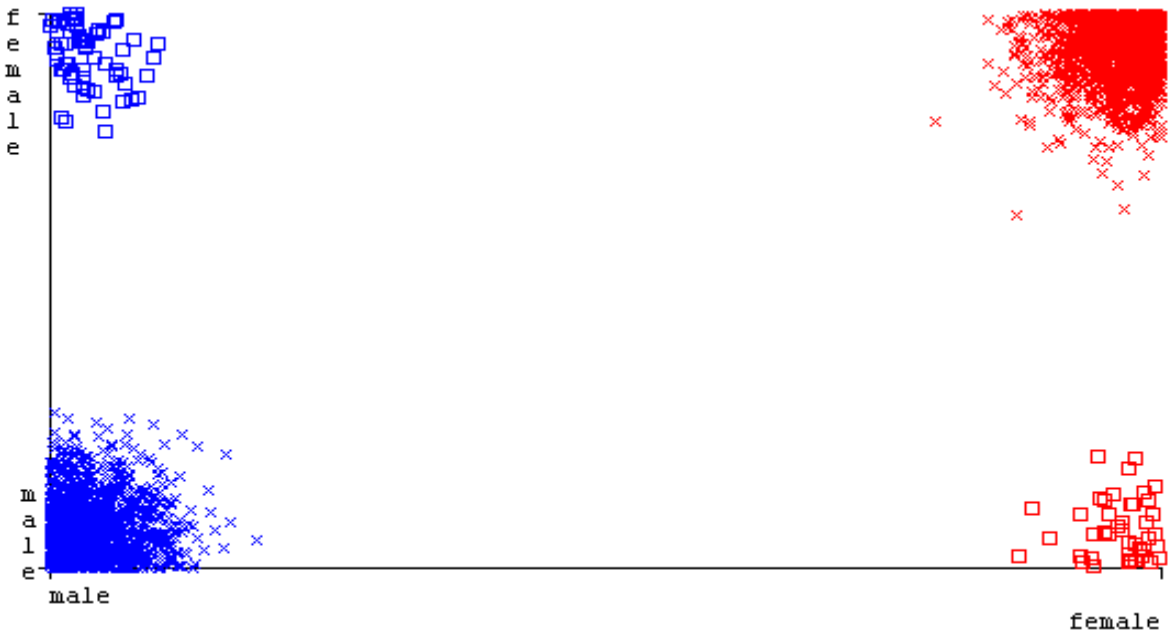
PRC Area

Tahmin sonucunda erkek ve kadın sınıfı PRC Area oranı 0,985 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.



Confusion matrix kısmında oluşturduğumuz test ve eğitim kümeleri sonucunda label parametremizde bulunan male ve famele tahminlerindeki başarımızı görmekteyiz. Confusion matrixde male ifadesini a olarak, female ifadesini b olarak ele almaktayız. A olarak ele aldığımız Male olan 1584 verinin 1522 tanesini male olarak doğru tahmin ederken 62 tanesini female olarak yanlış tahmin etmiştir. B olarak ele aldığımız female olan 1584 verinin 1541 tanesini female olarak doğru tahmin ederken 43 tanesini male olarak yanlış tahmin etmiştir. Tüm veri setine baktığımız zamanda 3168 veriden 3063 veriyi doğru tahmin ederken 105 veriyi yanlış tahmin etmiştir.

Yanlış Tahmin Edilen Verilerin Grafik Olarak Gösterimi

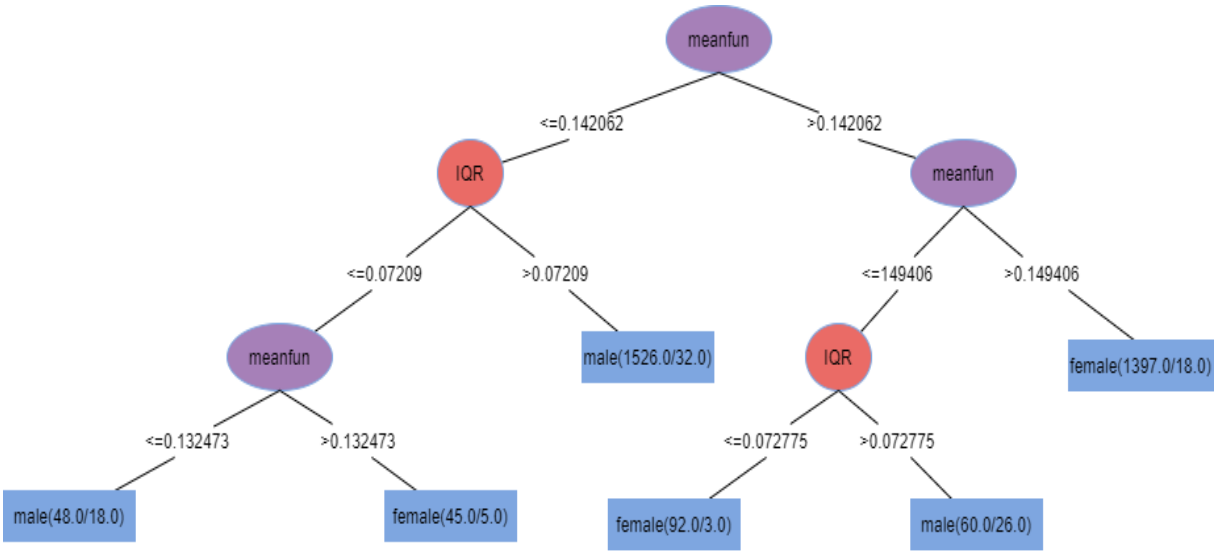


Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin birleşimi gösterilmiştir.

Tahmin modelimizin sınıflandırıcı hata grafiğine baktığımızda 62 male 43 female olmak üzere 105 hata yapıldığı görülmekte. 62 erkek çıktı kadın olarak tahmin edildiği ve 43 kadın çıktı ise erkek olarak tahmin edildiği görülmektedir. Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin kesişimi gösterilmiştir.

Örnek3

Veri setimizin içerisinden meanfun, sd, Q25, IQR, label parametrelerini temel alarak ilerledik. Karar Ağacı algoritmasından makineye veri setinin sonucu bildiğimiz değerlerin %11'ini eğitim vererek %89'unu tahmin etmesini istedik. Label parametresi üzerinden de sınıflandırmasını yaptık. Yaprak başına düşen minimum obje (MinNumObj) sayısı 45 olarak belirlenmiştir.



Sonuç 3

Not: Sınıflandırma yönteminde aynı parametreler temel alındığı için entropi tekrar anlatılmamıştır.

Başarı Oranı	***94.1135 %***
Doğru Sınıflandırılmış Örnekler	2654
Kappa İstatistiği	0.7932
Ortalama Mutlak Hata	0.0953
Kök Ortalama Kare Hatası	0.2333
Göreli Mutlak Hata	19.0553 %
Kök Göreli Kare Hatası	46.6636 %
Toplam Örnek Sayısı	2820

Modeli oluşturmak için geçen süre: 0 saniyedir.

Correctly Classified Instances

Veri setinin %89'ü yani 2820'si test kümesi kalan 348'i eğitim kümesi olarak ele alınarak tahmin yapılmıştır. Bunun sonucunda 2820 adet veriden 2654'ü doğru tahmin ederek %94.1135 oranında başarı elde edilmiştir. Modelimiz tahmin yaparken 166 adet veriyi yanlış bulmuştur.

Kappa Statistic

Kadın ve erkek değerlerinin arasındaki karşılaştırmalı uyuşmanın güvenilirlik oranı 0.7932'dir. Elde edilen değer 0.81-1.00 aralığında olduğu için kadın ve erkek değerleri arasında neredeyse **önemli derecede bir uyuşma** olduğu görülmektedir.

Mean Absolute Error

Tahmin sonucunda elde edilen ortalama mutlak hata oranı 0.0953'dür. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Mean Squared Error

Tahmin sonucunda elde edilen karekök ortalama hata oranı 0.2333'dür. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Relative Absolute Error

Tahmin sonucunda gerçek değer ile hesaplanan değer arasındaki farkın gerçek değere oranlanması sonucunda 19.0553 değeri elde edilmiştir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Relative Squared Error

Tahmin sonucunda kök göreceli hata oranı 46.6636'dur. Bu değer 0'a çok yakın olmasa da diğer metriklerin 0'a yakın olmasından dolayı modelin başarı oranını çok etkilememiştir.

	TP Rate	***FP Rate***	***Presicion***	***Recall***	***F-Measure***	***MCC***	***ROC Area***	***PRC Area***	***Class***
	0.981	0.099	0.909	0.981	0.943	0.885	0.941	0.901	Male
	0.901	0.019	0.979	0.901	0.939	0.885	0.941	0.932	Female
Ağırlıklı Ortalama	0.941	0.059	0.944	0.941	0.941	0.885	0.941	0.916	

TP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini erkek sesi olarak tahmin oranı 0,981'dür. Sınıfı kadın olan verilerden kadın sesini kadın sesi olarak tahmin oranı 0,901'dir. Bu değerler 1'e yakın olduğu için iyi bir isabet oranı elde edildiği görülmektedir. Yapılan Karar Ağacı sınıflandırma yöntemi sonucunda verilerin tamamına yakın bir kısmını doğru tahmin

ettiği görülmektedir. Aynı zamanda kadın ve erkek sınıflarında en çok doğru tahmini yapan erkek sınıfıdır.

FP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini kadın sesi olarak tahmin etme oranı 0,099'dur. Sınıfı kadın olan verilerden kadın sesini erkek sesi olarak tahmin etme oranı 0,019' dur. Bu değerler 0'a yakın olduğu için yapılan hatalı tahminin çok az olduğu görülmektedir. Aynı zamanda erkek ve kadın sınıfları arasında en çok hata yapan erkek sınıfıdır.

Precision

Tahmin sonucunda erkek sınıfı hassasiyet oranı 0,909 iken kadın sınıfı hassasiyet oranı 0,979'dur. Hassasiyet oranları 1'e yakın olduğu için sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Sınıfı erkek olan verilerin hassasiyet oranı daha yüksektir.

Recall

Tahmin sonucunda erkek sınıfı geri çağırma oranı 0,981 iken kadın sınıfı geri çağırma oranı 0,901'dir. Örneğin gerçekte sesin kadın olduğu durumda tahminin erkek sesi olarak yapılmasıdır. Bu hata 0'a yaklaştıkça artar 1'e yaklaştıkça azalır. Bu durumda sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Erkek sınıfına ait veriler daha doğru tahmin edilmiştir.

F-Measure

Tahmin sonucunda erkek sınıfı F-Measure oranı 0,943 iken kadın sınıfı oranı 0.939 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir. Erkek sınıfına ait verilerde daha fazla başarı elde edilmiştir.

MCC

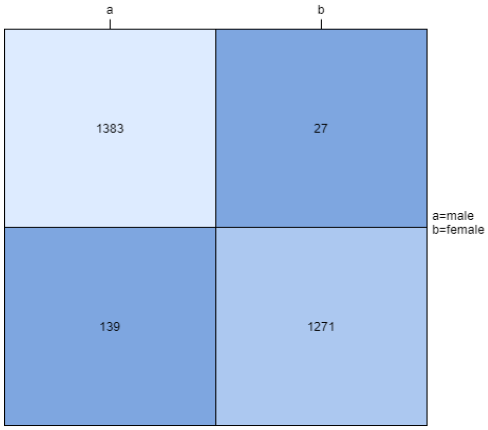
Tahmin sonucunda erkek ve kadın sınıfı MCC oranı 0,885 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

ROC Area

Tahmin sonucunda erkek ve kadın sınıfı ROC Area oranı 0,941 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

PRC Area

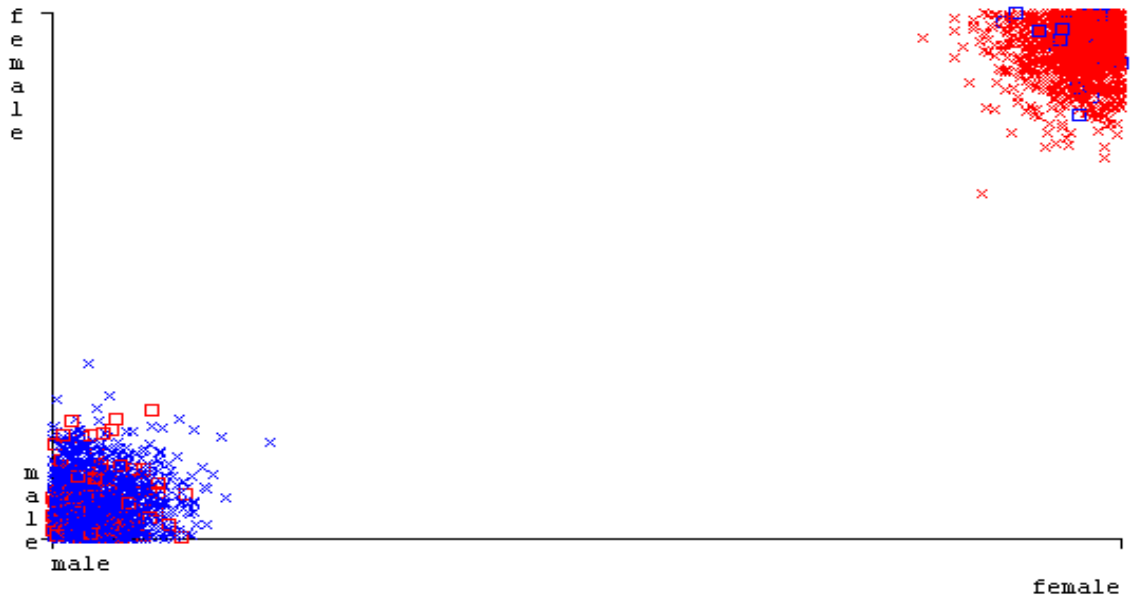
Tahmin sonucunda erkek sınıfı PRC Area oranı 0,901 iken kadın sınıfı oranı 0.932 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir. Kadın sınıfına ait verilerde daha fazla başarı elde edilmiştir.



Confusion matrix kısmında oluşturduğumuz test ve eğitim kümeleri sonucunda label parametremizde bulunan male ve famele tahminlerindeki başarımızı görmekteyiz. Confusion matrixde male ifadesini a olarak, female ifadesini b olarak ele almaktayız. A olarak ele aldığımız Male olan 1410 verinin 1383 tanesini male olarak doğru tahmin ederken 27 tanesini female olarak yanlış tahmin etmiştir. B olarak ele aldığımız female olan 1410 verinin 1271 tanesini female olarak doğru tahmin ederken 139 tanesini male olarak yanlış tahmin etmiştir. Tüm veri setine baktığımız zamanda 2820 veriden 2654 veriyi doğru tahmin ederken 166 veriyi yanlış tahmin etmiştir.

Yanlış Tahmin Edilen Verilerin Grafik Olarak Gösterimi

Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin birleşimi gösterilmiştir.



Tahmin modelimizin sınıflandırıcı hata grafiğine baktığımızda 27 male 139 female olmak üzere 166 hata yapıldığı görülmekte. 27 erkek çıktı kadın olarak tahmin edildiği ve 139 kadın çıktı ise erkek olarak tahmin edildiği görülmektedir. Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin kesişimi gösterilmiştir.

3.5.2.Naive Bayes Sınıflandırma Yöntemi

Örnek1:

Veri setimizin içerisinden meanfun, sd, Q25, IQR, label parametrelerini temel alarak ilerledik. Navie Bayes algoritmasından makineye veri setinin sonucu bildiğimiz değerlerin %60'ını eğitim vererek %40 unu tahmin etmesini istedik. Label parametresi üzerinden de sınıflandırmasını yaptık.

Sonuç1:

Başarı Oranı	***96.3694 %***
Doğru Sınıflandırılmış Örnekler	1221
Kappa İstatistiği	0.9274
Ortalama Mutlak Hata	0.0436
Kök Ortalama Kare Hatası	0.1701
Göreli Mutlak Hata	8.7107%
Kök Göreli Kare Hatası	34.0126%

Toplam Örnek Sayısı	1267

Modeli oluşturmak için geçen süre: 0 saniyedir.

Correctly Classified Instances

Veri setinin %40'ı yani 1267'si test kümesi kalan 1901'i eğitim kümesi olarak ele alınarak tahmin yapılmıştır. Bunun sonucunda 1267 adet veriden 1221'ini doğru tahmin ederek %96.3694 oranında başarı elde edilmiştir. Modelimiz tahmin yaparken 46 adet veriyi yanlış bulmuştur.

Kappa Statistic

Kadın ve erkek değerlerinin arasındaki karşılaştırmalı uyuşmanın güvenilirlik oranı 0.9274'dir. Elde edilen değer 0.81-1.00 aralığında olduğu için kadın ve erkek değerleri arasında neredeyse önemli derecede bir uyuşma olduğu görülmektedir.

Mean Absolute Error

Tahmin sonucunda elde edilen ortalama mutlak hata oranı 0.0436'dır. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Mean Squared Error

Tahmin sonucunda elde edilen karekök ortalama hata oranı 0.1701'dür. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Relative Absolute Error

Tahmin sonucunda gerçek değer ile hesaplanan değer arasındaki farkın gerçek değere oranlanması sonucunda 8.7107 değeri elde edilmiştir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Relative Squared Error

Tahmin sonucunda kök göreceli hata oranı 34.0126'dır. Bu değer 0'a çok yakın olmasa da diğer metriklerin 0'a yakın olmasından dolayı modelin başarı oranını çok etkilememiştir.

	TP Rate	*FP Rate*	*Presicion*	*Recall*	*F-Measure*	*MCC*	*ROC Area*	*PRC Area*	*Class*
	0.973	0.046	0.955	0.973	0.964	0.928	0.993	0.993	Male
	0.954	0.027	0.972	0.954	0.963	0.928	0.993	0.993	Female
Ağırlıklı Ortalama	0.964	0.036	0.964	0.964	0.964	0.928	0.993	0.993	

TP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini erkek sesi olarak tahmin oranı 0,973'dür. Sınıfı kadın olan verilerden kadın sesini kadın sesi olarak tahmin oranı 0,954'dür. Bu değerler 1'e yakın olduğu için iyi bir isabet oranı elde edildiği görülmektedir. Yapılan Navie Bayes sınıflandırma yöntemi sonucunda verilerin tamamına yakın bir kısmını doğru tahmin ettiği görülmektedir. Aynı zamanda kadın ve erkek sınıflarında en çok doğru tahmini yapan erkek sınıfıdır.

FP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini kadın sesi olarak tahmin etme oranı 0,046'dır. Sınıfı kadın olan verilerden kadın sesini erkek sesi olarak tahmin etme oranı 0,027' dir. Bu değerler 0'a yakın olduğu için yapılan hatalı tahminin çok az olduğu görülmektedir. Aynı zamanda erkek ve kadın sınıfları arasında en çok hata yapan erkek sınıfıdır.

Precision

Tahmin sonucunda erkek sınıfı hassasiyet oranı 0,955 iken kadın sınıfı hassasiyet oranı 0,972'dir. Hassasiyet oranları 1'e yakın olduğu için sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Sınıfı erkek olan verilerin hassasiyet oranı daha yüksektir.

Recall

Tahmin sonucunda erkek sınıfı geri çağırma oranı 0,973 iken kadın sınıfı geri çağırma oranı 0,954'dür. Örneğin gerçekte sesin kadın olduğu durumda tahminin erkek sesi olarak yapılmasıdır. Bu hata 0'a yaklaştıkça artar 1'e yaklaştıkça azalır. Bu durumda sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Erkek sınıfına ait veriler daha doğru tahmin edilmiştir.

F-Measure

Tahmin sonucunda erkek sınıfı F-Measure oranı 0,964 iken kadın sınıfı oranı 0.963 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir. Erkek sınıfına ait verilerde daha fazla başarı elde edilmiştir.

MCC

Tahmin sonucunda erkek ve kadın sınıfı MCC oranı 0,928 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

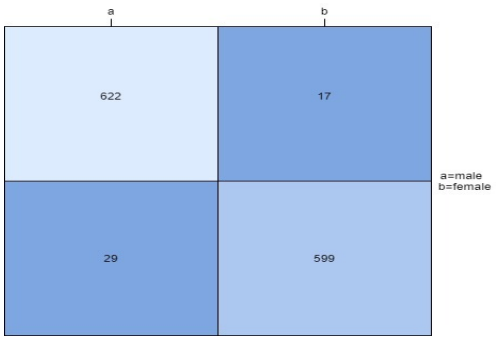
ROC Area

Tahmin sonucunda erkek ve kadın sınıfı ROC Area oranı 0,993 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

PRC Area

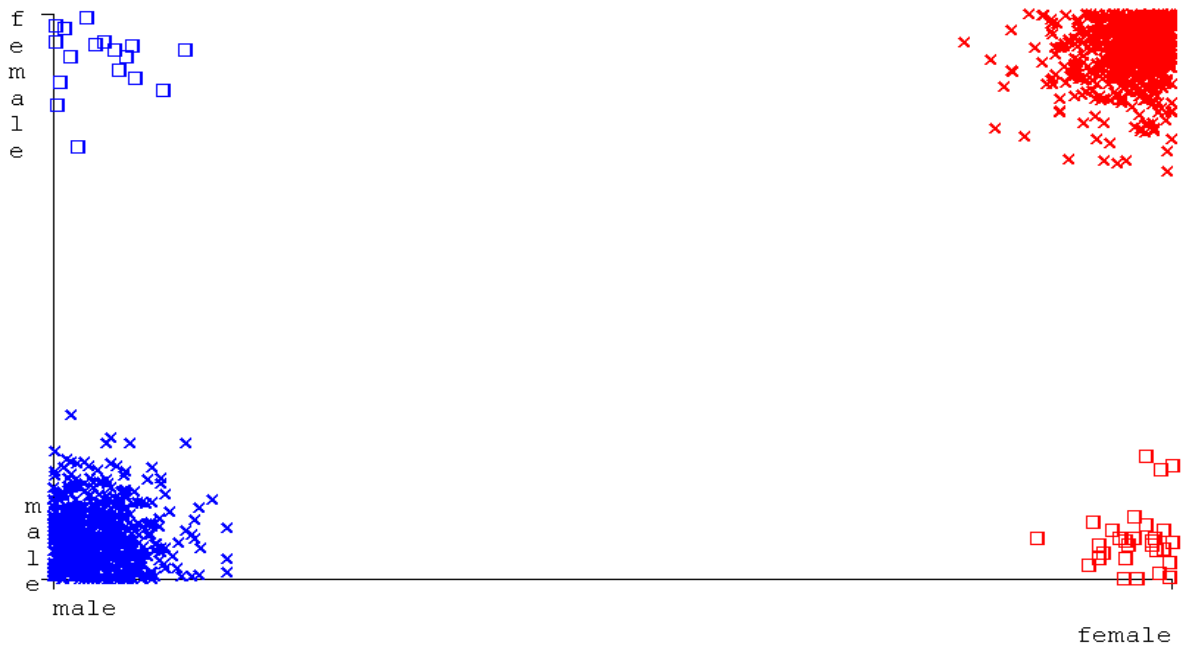
Tahmin sonucunda erkek ve kadın sınıfı PRC Area oranı 0,993 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir. Kadın sınıfına ait verilerde daha fazla başarı elde edilmiştir.

Confusion Matrix

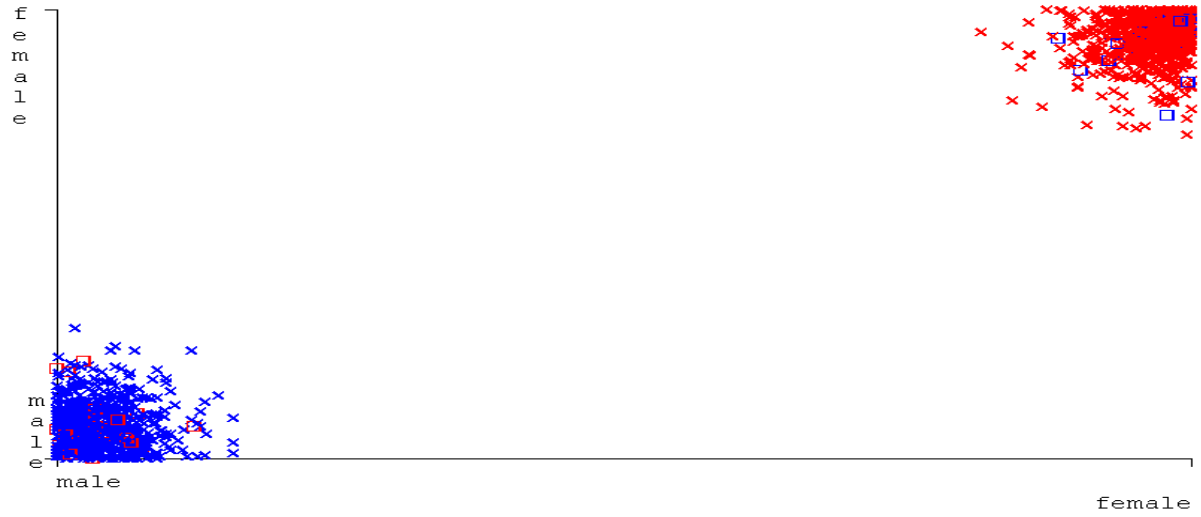


Confusion matrix kısmında oluşturduğumuz test ve eğitim kümeleri sonucunda label parametremizde bulunan male ve famele tahminlerindeki başarımızı görmekteyiz. Confusion matrixde male ifadesini a olarak, female ifadesini b olarak ele almaktayız. A olarak ele aldığımız Male olan 639 verinin 622 tanesini male olarak doğru tahmin ederken 17 tanesini female olarak yanlış tahmin etmiştir. B olarak ele aldığımız female olan 628 verinin 599 tanesini female olarak doğru tahmin ederken 29 tanesini male olarak yanlış tahmin etmiştir. Tüm veri setine baktığımız zamanda 1267 veriden 1221 veriyi doğru tahmin ederken 46 veriyi yanlış tahmin etmiştir.

Yanlış Tahmin Edilen Verilerin Grafik Olarak Gösterimi



Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin birleşimi gösterilmiştir.



Tahmin modelimizin sınıflandırıcı hata grafiğine baktığımızda 17 male 29 female olmak üzere 46 hata yapıldığı görülmekte. 17 erkek çıktı kadın olarak tahmin edildiği ve 29 kadın çıktı ise erkek olarak tahmin edildiği görülmektedir. Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin kesişimi gösterilmiştir.

Örnek2

Weka programının Navie Bayes sınıflandırma yönteminde karışık veri setlerinde kullanılan ve rastgele doğru ve yanlış çıkma olasılığını düşüren cross-validation seçeneğini 4 seçerek veri setimizi dörde bölüyoruz. Bu aşamada 729 adet veriden oluşan 4 adet veri seti oluşmaktadır. Daha sonra bu 4 adet veri setinden bir tanesi seçilerek test kümesi oluşturulur. Geriye kalan 2367 adet veri eğitim kümesi olarak belirlenmektedir. Oluşturulan eğitim kümesi ile model eğitilir ve 729 adet veri bulunan test kümesi ile test edilir. Kısaca cross-validation seçeneği 4 (%25 eğitim/ %75 test) seçilerek veri setinde temel alınan Meanfun, IQR, Sd, Q25, Label parametreleri üzerinde tahmin yapılmıştır. Model kurulurken yapraktaki minimum obje sayısı 10 olarak belirlenmiştir.

Sonuç2:

Başarı Oranı	***95.928 %***
Doğru Sınıflandırılmış Örnekler	3039
Kappa İstatistiği	0.9186
Ortalama Mutlak Hata	0.0485
Kök Ortalama Kare Hatası	0.1826
Göreli Mutlak Hata	9.6935%
Kök Göreli Kare Hatası	36.5224%
Toplam Örnek Sayısı	3168

Modeli oluşturmak için geçen süre: 0 saniyedir.

Correctly Classified Instances

Veri setinin %4'ü test kümesi olarak ele alındığı için 3168 satır veriden 3039'u üzerinde tahmin yapılmıştır. Bunun sonucunda 3168 adet veriden 3039 doğru tahmin ederek %95.928 oranında başarı elde edilmiştir. Modelimiz tahmin yaparken 129 adet veriyi yanlış bulmuştur.

Kappa Statistic

Kadın ve erkek değerlerinin arasındaki karşılaştırmalı uyuşmanın güvenilirlik oranı 0.9192'dir. Elde edilen değer 0.81-1. 00 aralığında olduğu için kadın ve erkek değerleri arasında neredeyse mükemmel bir uyuşma olduğu görülmektedir.

Mean Absolute Error

Tahmin sonucunda elde edilen ortalama mutlak hata oranı 0.0485'dir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Mean Squared Error

Tahmin sonucunda elde edilen karekök ortalama hata oranı 0.1826'dır. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Relative Absolute Error

Tahmin sonucunda gerçek değer ile hesaplanan değer arasındaki farkın gerçek değere oranlanması sonucunda 9.6935 değeri elde edilmiştir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Relative Squared Error

Tahmin sonucunda kök göreceli hata oranı 36.5224 dür. Bu değer 0'a çok yakın olmasa da diğer metriklerin 0'a yakın olmasından dolayı modelin başarı oranını çok etkilememiştir.

	TP Rate	***FP Rate***	***Presicion***	***Recall***	***F-Measure***	***MCC***	***ROC Area***	***PRC Area***	***Class***
	0.973	0.054	0.947	0.973	0.960	0.919	0.991	0.991	Male
	0.946	0.027	0.972	0.976	0.959	0.919	0.991	0.991	Female
Ağırlıklı Ortalama	0.959	0.041	0.960	0.970	0.959	0.919	0.991	0.991	

TP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini erkek sesi olarak tahmin oranı 0,973'dür. Sınıfı kadın olan verilerden kadın sesini kadın sesi olarak tahmin oranı 0,946'dır. Bu değerler 1'e yakın olduğu için iyi bir isabet oranı elde edildiği görülmektedir. Yapılan Navie Bayes sınıflandırma yöntemi sonucunda verilerin tamamına yakın bir kısmını doğru tahmin ettiği görülmektedir. Aynı zamanda kadın ve erkek sınıflarında en çok doğru tahmini yapan erkek sınıfıdır.

FP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini kadın sesi olarak tahmin etme oranı 0,054'dir. Sınıfı kadın olan verilerden kadın sesini erkek sesi olarak tahmin etme oranı 0,027'dir. Bu değerler 0'a yakın olduğu için yapılan hatalı tahminin çok az olduğu görülmektedir. Aynı zamanda erkek ve kadın sınıfları arasında en çok hata yapan erkek sınıfıdır.

Precision

Tahmin sonucunda erkek sınıfı hassasiyet oranı 0,947 iken kadın sınıfı hassasiyet oranı 0,972'dir. Hassasiyet oranları 1'e yakın olduğu için sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Sınıfı kadın olan verilerin hassasiyet oranı daha yüksektir.

Recall

Tahmin sonucunda erkek sınıfı geri çağırma oranı 0,973 iken kadın sınıfı geri çağırma oranı 0,976'dır. Örneğin gerçekte sesin erkek olduğu durumda tahminin kadın sesi olarak yapılmasıdır. Bu hata 0'a yaklaştıkça artar 1'e yaklaştıkça azalır. Bu durumda sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Kadın sınıfına ait veriler daha doğru tahmin edilmiştir.

F-Measure

Tahmin sonucunda erkek sınıfı F-Measure oranı 0,960, kadın sınıfı oranı 0,959 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

MCC

Tahmin sonucunda erkek ve kadın sınıfı MCC oranı 0,919 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

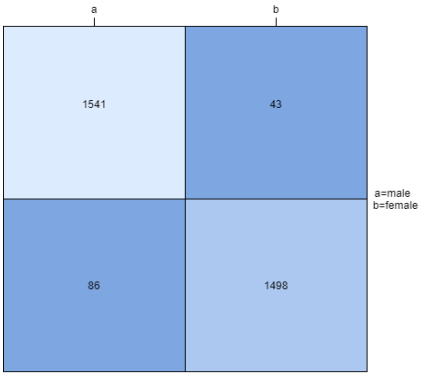
ROC Area

Tahmin sonucunda erkek ve kadın sınıfı ROC Area oranı 0,991 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

PRC Area

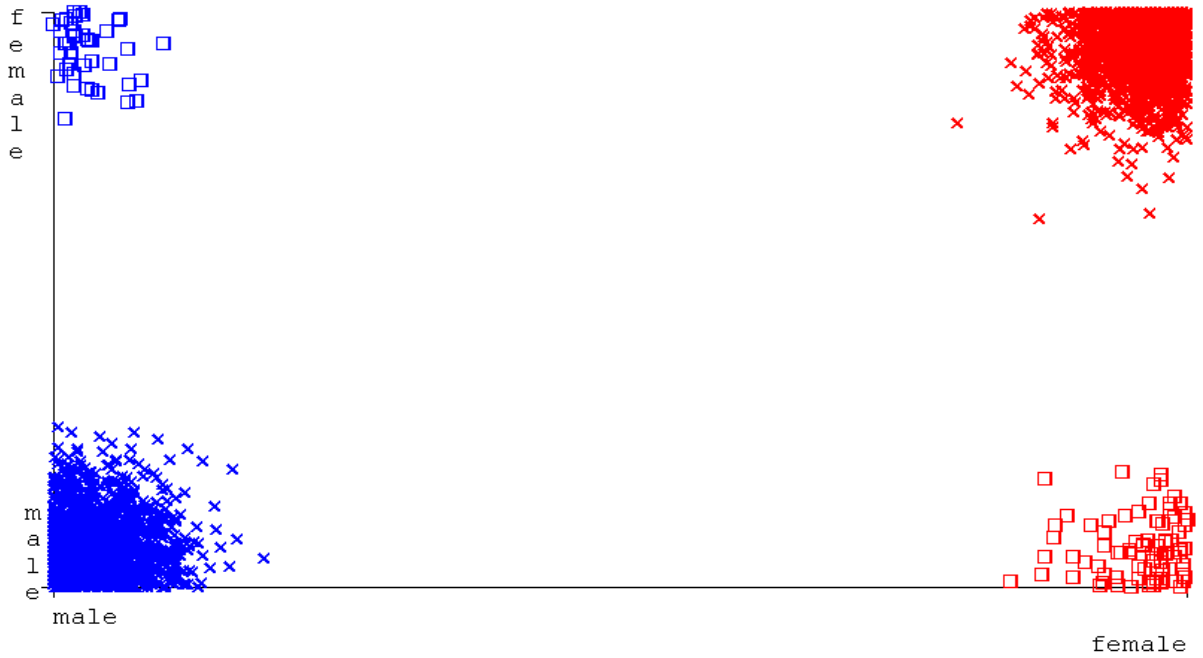
Tahmin sonucunda erkek ve kadın sınıfı PRC Area oranı 0,991 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

Confusion Matrix



Confusion matrix kısmında oluşturduğumuz test ve eğitim kümeleri sonucunda label parametremizde bulunan male ve female tahminlerindeki başarımızı görmekteyiz. Confusion matrixde male ifadesini a olarak, female ifadesini b olarak ele almaktayız. A olarak ele aldığımız Male olan 1584 verinin 1541 tanesini male olarak doğru tahmin ederken 43 tanesini female olarak yanlış tahmin etmiştir. B olarak ele aldığımız female olan 1584 verinin 1498 tanesini female olarak doğru tahmin ederken 86 tanesini male olarak yanlış tahmin etmiştir. Tüm veri setine baktığımız zamanda 3168 veriden 3039 veriyi doğru tahmin ederken 129 veriyi yanlış tahmin etmiştir.

Yanlış Tahmin Edilen Verilerin Grafik Olarak Gösterimi



Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin birleşimi gösterilmiştir.

Tahmin modelimizin sınıflandırıcı hata grafiğine baktığımızda 86 male 43 female olmak üzere 129 hata yapıldığı görülmekte. 86 erkek çıktı kadın olarak tahmin edildiği ve 43 kadın çıktı ise erkek olarak tahmin edildiği görülmektedir. Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin kesişimi gösterilmiştir.

Örnek3

Veri setimizin içerisinden meanfun, sd, Q25, IQR, label parametrelerini temel alarak ilerledik. Navie bayes algoritmasından makineye veri setinin sonucu bildiğimiz değerlerin %11'ini eğitim vererek %89'unu tahmin etmesini istedik. Aynı zamanda küme sayısını 6 olarak ele aldık. Label parametresi üzerinden de sınıflandırmasını yaptık.

SONUÇ 3:

Başarı Oranı	***62.5532 %***
Doğru Sınıflandırılmış Örnekler	1764
Kappa İstatistiği	0.2511
Ortalama Mutlak Hata	0.4201
Kök Ortalama Kare Hatası	0.4908
Göreli Mutlak Hata	84.0106 %
Kök Göreli Kare Hatası	98.1588 %
Toplam Örnek Sayısı	2820

Modeli oluşturmak için geçen süre: 0,13 saniyedir.

Correctly Classified Instances

Veri setinin %89'i test kümesi olarak ele alındığı için 3168 satır veriden 2820'si üzerinde tahmin yapılmıştır. Bunun sonucunda 2820 adet veriden 1764'ünü doğru tahmin ederek %62.5532 oranında başarı elde edilmiştir. Modelimiz tahmin yaparken 111 adet veriyi yanlış bulmuştur.

Kappa Statistic

Kadın ve erkek değerlerinin arasındaki karşılaştırmalı uyuşmanın güvenilirlik oranı 0.2511'dir. Elde edilen değer 0.81-1. 00 aralığında olduğu için kadın ve erkek değerleri arasında neredeyse mükemmel bir uyuşma olduğu görülmektedir.

Mean Absolute Error

Tahmin sonucunda elde edilen ortalama mutlak hata oranı 0.0479'dur. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Mean Squared Error

Tahmin sonucunda elde edilen karekök ortalama hata oranı 0.1795'dir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Relative Absolute Error

Tahmin sonucunda gerçek değer ile hesaplanan değer arasındaki farkın gerçek değere oranlanması sonucunda 9.5875 değeri elde edilmiştir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Relative Squared Error

Tahmin sonucunda kök göreceli hata oranı 35.9024'dür. Bu değer 0'a çok yakın olmasa da diğer metriklerin 0'a yakın olmasından dolayı modelin başarı oranını çok etkilememiştir.

	TP Rate	***FP Rate***	***Presicion***	***Recall***	***F-Measure***	***MCC***	***ROC Area***	***PRC Area***	***Class***
	0,972	0,051	0.950	0.972	0.961	0.922	0.991	0.991	Male
	0,949	0,028	0.972	0.949	0.960	0.922	0.991	0.991	Female
Ağırlıklı Ortalama	0,961	0.039	0.961	0.961	0.961	0.922	0.991	0.991	

TP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini erkek sesi olarak tahmin oranı 0,972 ’dir. Sınıfı kadın olan verilerden kadın sesini kadın sesi olarak tahmin oranı 0,949’dur. Bu değerler 1’e yakın olduğu için iyi bir isabet oranı elde edildiği görülmektedir. Yapılan Navie Bayes sınıflandırma yöntemi sonucunda verilerin tamamına yakın bir kısmını doğru tahmin ettiği görülmektedir. Aynı zamanda kadın ve erkek sınıflarında en çok doğru tahmini yapan kadın sınıfıdır.

FP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini kadın sesi olarak tahmin etme oranı 0,051’dür. Sınıfı kadın olan verilerden kadın sesini erkek sesi olarak tahmin etme oranı 0,028’dır. Bu değerler 0’a yakın olduğu için yapılan hatalı tahminin çok az olduğu görülmektedir. Aynı zamanda erkek ve kadın sınıfları arasında en çok hata yapan kadın sınıfıdır.

Precision

Tahmin sonucunda erkek sınıfı hassasiyet oranı 0,950 iken kadın sınıfı hassasiyet oranı 0,972’dür. Hassasiyet oranları 1’e yakın olduğu için sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Sınıfı erkek olan verilerin hassasiyet oranı daha yüksektir.

Recall

Tahmin sonucunda erkek sınıfı geri çağırma oranı 0,972 iken kadın sınıfı geri çağırma oranı 0,949’dir. Örneğin gerçekte sesin kadın olduğu durumda tahminin erkek sesi olarak yapılmasıdır. Bu hata 0’a yaklaştıkça artar 1’e yaklaştıkça azalır. Bu durumda sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Kadın sınıfına ait veriler daha doğru tahmin edilmiştir.

F-Measure

Tahmin sonucunda erkek F-Measure oranı 0,961 iken kadın sınıfı F-Measure oranı 0,960 çıkmıştır. Bu değerin 1’e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

MCC

Tahmin sonucunda erkek ve kadın sınıfı MCC oranı 0,922 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

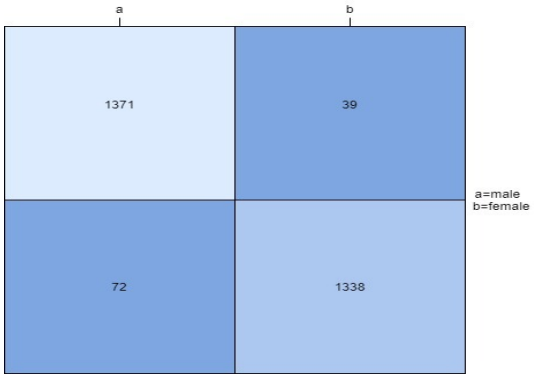
ROC Area

Tahmin sonucunda erkek ve kadın sınıfı ROC Area oranı 0,991 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

PRC Area

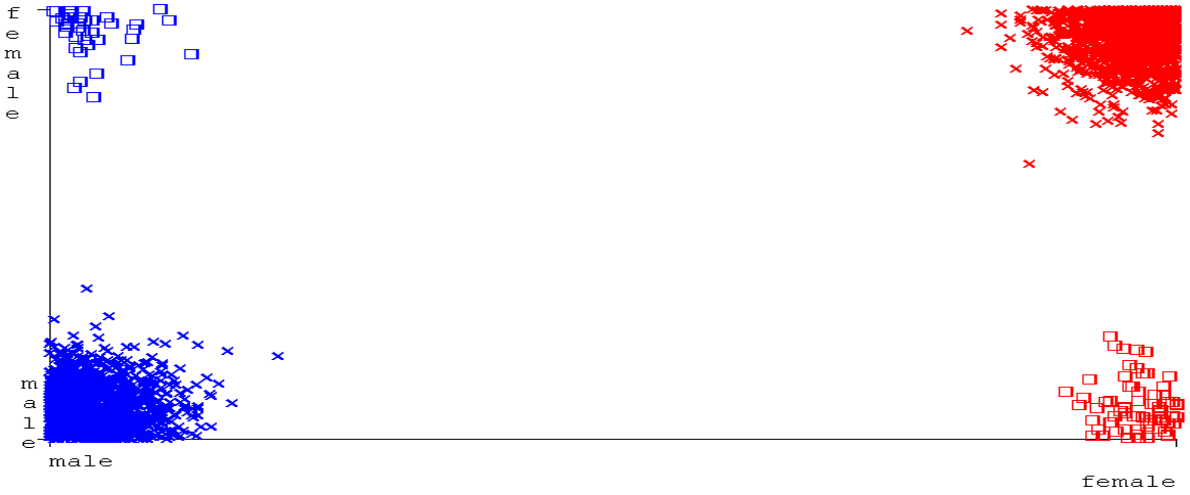
Tahmin sonucunda erkek ve kadın sınıfı PRC Area oranı 0,991 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir. Aynı zamanda erkek sınıfının PRC Area oranı kadın sınıfından daha başarılıdır.

Confusion Matrix



Confusion matrix kısmında oluşturduğumuz test ve eğitim kümeleri sonucunda label parametremizde bulunan male ve famele tahminlerindeki başarımızı görmekteyiz. Confusion matrixde male ifadesini a olarak, female ifadesini b olarak ele almaktayız. A olarak ele aldığımız Male olan 1410 verinin 1371 tanesini male olarak doğru tahmin ederken 39 tanesini female olarak yanlış tahmin etmiştir. B olarak ele aldığımız female olan 1410 verinin 1338 tanesini female olarak doğru tahmin ederken 72 tanesini male olarak yanlış tahmin etmiştir. Tüm veri setine baktığımız zamanda 2820 veriden 2709 veriyi doğru tahmin ederken 111 veriyi yanlış tahmin etmiştir.

Yanlış Tahmin Edilen Verilerin Grafik Olarak Gösterimi



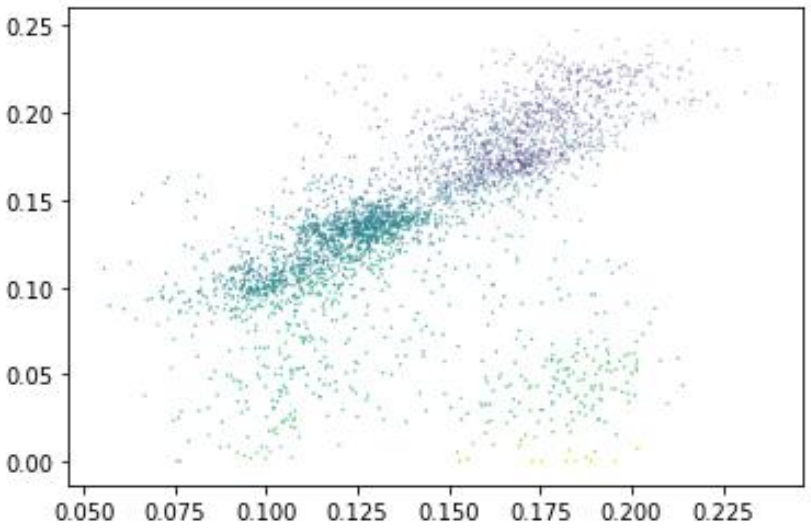
Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin birleşimi gösterilmiştir.

Tahmin modelimizin sınıflandırıcı hata grafiğine baktığımızda 39 male 72 female olmak üzere 111 hata yapıldığı görülmekte. 39 erkek çıktı kadın olarak tahmin edildiği ve 72 kadın çıktı ise erkek olarak tahmin edildiği görülmektedir. Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin kesişimi gösterilmiştir.

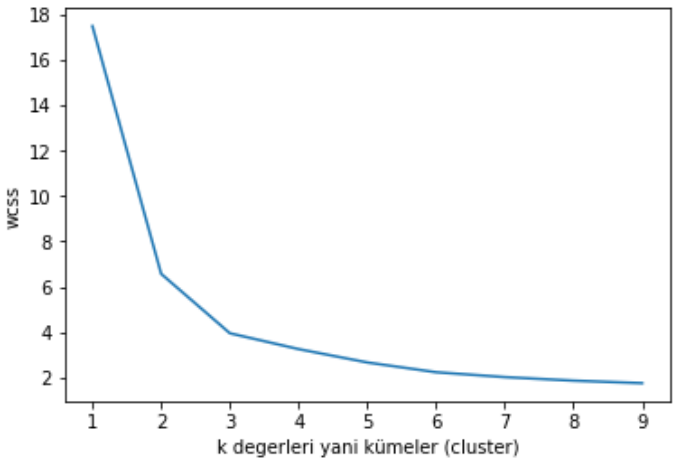
3.5.3.K-En Yakın Komşu Sınıflandırma Yöntemi

Örnek1

Veri setimizin içerisinden meanfun, sd, Q25, IQR, label parametrelerini temel alarak ilerledik. K-En Yakın Komşu algoritmasından makineye veri setinin sonucu bildiğimiz değerlerin %60'ını eğitim vererek %40 unu tahmin etmesini istedik. Küme sayısını 2 olarak ele alınmıştır. Label parametresi üzerinden de sınıflandırmasını yaptık.

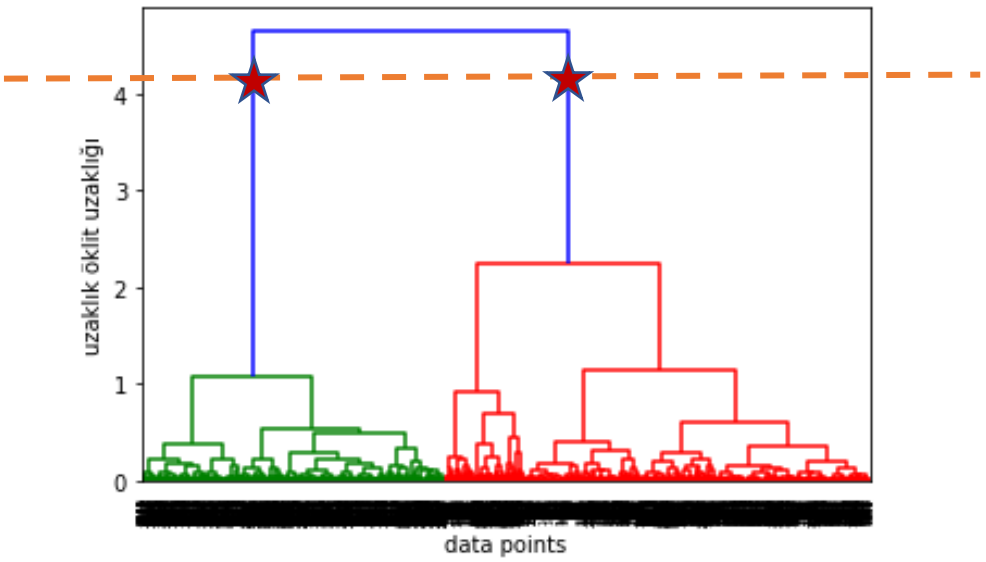


Yukarıdaki grafikte Meanfun (turkuaz), Q25 (mor), IQR (sarı) ve Sd(açık yeşil) verilerinin dağılımı görülmektedir. Meanfun ve Q25 verilerinin daha sığ olduğunu ve birbirine yakın olduğu görülmektedir.

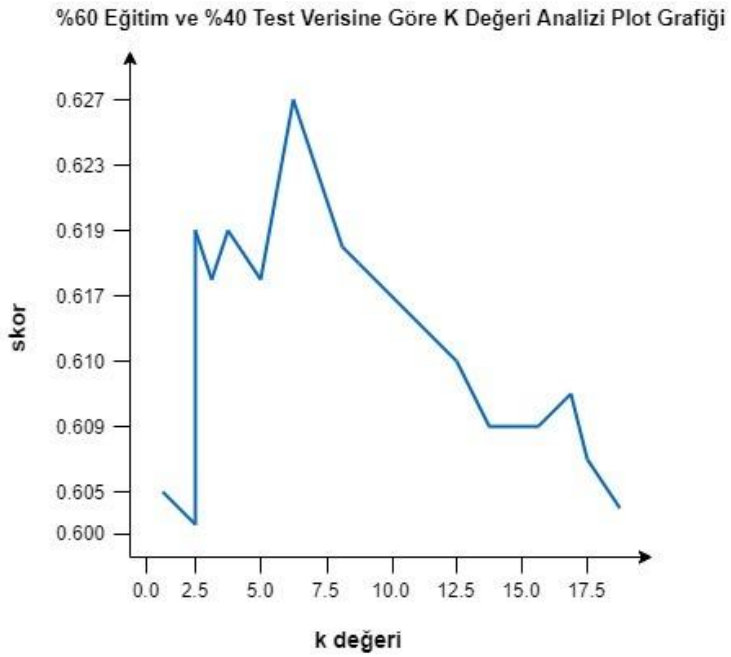


Yukarıdaki Meanfun, Q25, IQR ve Sd giriş değerleri temel alınarak çizdirilen kol grafiği incelendiğinde doğru 2 üzerinde kırılarak gittikçe düzleşmiştir. Bir diğer kırılma 3'te

gerçekleşmiştir. Bu kırılmaların amacı veri setinin 2 kümeden ya da 3 kümeden oluştuğunu göstermektir. Kol grafiğinde ilk keskin kırılma değeri ele alındığı için ilk kırılma 2 de gerçekleşmiştir. Bu bağlamda küme değeri 2 seçilirse modelin başarılı bir sonuç vereceğini göstermektedir.



Yukarıda çizdirilen dendogram grafiğinden küme sayısının kaç olması gerektiğini görebilmekteyiz. İlk olarak en uzun kenara bir doğru çizilir. Çizilen doğrunun kesişme noktaları ele alınır. Kesişme noktalarının sayısı KNN sınıflandırma yönteminde en yüksek başarı oranının hangi küme sayısında elde edileceğini göstermektedir. Çizilen doğruyu kesen 2 nokta bulunduğu için küme değeri 2 seçilir.



Yukarıdaki grafik veri setinin %60'ı eğitim ve %40'ı test verisi olduğunda seçilecek k değerlerinin çıkacak başarı değerleri gösterilmektedir. Spyder 'da grafik kodlanırken k değeri aralığı 1-20 olarak belirlenmiştir. Küme sayısı 2 seçilirse elde edilecek en yüksek başarı oranı %96 civarında olacağı 17 seçilirse en düşük başarı oranı %93 civarında olacağı görülmektedir. Bu grafik elde edildikten sonra oluşturduğumuz modelin kümesi 2 olarak belirlenmiştir.

Sonuç1:

Başarı Oranı	***96.764 %***
Doğru Sınıflandırılmış Örnekler	1226
Kappa İstatistiği	0.9353
Ortalama Mutlak Hata	0.0295
Kök Ortalama Kare Hatası	0.1513
Göreli Mutlak Hata	5.8898%
Kök Göreli Kare Hatası	30.2503%
Toplam Örnek Sayısı	1267
Küme Sayısı	2

Modeli test bölümünde test etmek için geçen süre: 0.44 saniyedir.

Correctly Classified Instances

Veri setinin %4'ü test kümesi olarak ele alındığı için 1267 satır veriden 1226'sı üzerinde tahmin yapılmıştır. Bunun sonucunda 1267 adet veriden 1226 doğru tahmin ederek %96.764 oranında başarı elde edilmiştir. Modelimiz tahmin yaparken 41 adet veriyi yanlış bulmuştur.

Kappa Statistic

Kadın ve erkek değerlerinin arasındaki karşılaştırmalı uyuşmanın güvenilirlik oranı 0.9353'dür. Elde edilen değer 0.81-1. 00 aralığında olduğu için kadın ve erkek değerleri arasında neredeyse mükemmel bir uyuşma olduğu görülmektedir.

Mean Absolute Error

Tahmin sonucunda elde edilen ortalama mutlak hata oranı 0.0295'dir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Mean Squared Error

Tahmin sonucunda elde edilen karekök ortalama hata oranı 0.1513'dür. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Relative Absolute Error

Tahmin sonucunda gerçek değer ile hesaplanan değer arasındaki farkın gerçek değere oranlanması sonucunda 5.8898 değeri elde edilmiştir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Relative Squared Error

Tahmin sonucunda kök göreceli hata oranı 30.2503 dür. Bu değer 0'a çok yakın olmasa da diğer metriklerin 0'a yakın olmasından dolayı modelin başarı oranını çok etkilememiştir.

	TP Rate	***FP Rate***	***Presicion***	***Recall***	***F-Measure***	***MCC***	***ROC Area***	***PRC Area***	***Class***
	0.980	0.045	0.957	0.980	0.968	0.936	0.983	0.977	Male
	0.955	0.020	0.979	0.955	0.967	0.936	0.983	0.972	Female
Ağırlıklı Ortalama	0.968	0.033	0.968	0.968	0.968	0.936	0.983	0.974	

TP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini erkek sesi olarak tahmin oranı 0,980'dir. Sınıfı kadın olan verilerden kadın sesini kadın sesi olarak tahmin oranı 0,955'dir. Bu değerler 1'e yakın olduğu için iyi bir isabet oranı elde edildiği görülmektedir. Yapılan K-Nearest Neighbor(KNN) sınıflandırma yöntemi sonucunda verilerin tamamına yakın bir kısmını doğru tahmin ettiği görülmektedir. Aynı zamanda kadın ve erkek sınıflarında en çok doğru tahmini yapan erkek sınıfıdır.

FP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini kadın sesi olarak tahmin etme oranı 0,045'dir. Sınıfı kadın olan verilerden kadın sesini erkek sesi olarak tahmin etme oranı

0,020’dir. Bu değerler 0’a yakın olduğu için yapılan hatalı tahminin çok az olduğu görülmektedir. Aynı zamanda erkek ve kadın sınıfları arasında en çok hata yapan erkek sınıfıdır.

Precision

Tahmin sonucunda erkek sınıfı hassasiyet oranı 0,957 iken kadın sınıfı hassasiyet oranı 0,979’dur. Hassasiyet oranları 1’e yakın olduğu için sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Sınıfı kadın olan verilerin hassasiyet oranı daha yüksektir.

Recall

Tahmin sonucunda erkek sınıfı geri çağırma oranı 0,980 iken kadın sınıfı geri çağırma oranı 0,955’dir. Örneğin gerçekte sesin erkek olduğu durumda tahminin kadın sesi olarak yapılmasıdır. Bu hata 0’a yaklaştıkça artar 1’e yaklaştıkça azalır. Bu durumda sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Erkek sınıfına ait veriler daha doğru tahmin edilmiştir.

F-Measure

Tahmin sonucunda erkek sınıfı F-Measure oranı 0,968, kadın sınıfı oranı 0,967 çıkmıştır. Bu değerin 1’e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

MCC

Tahmin sonucunda erkek ve kadın sınıfı MCC oranı 0,936 çıkmıştır. Bu değerin 1’e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

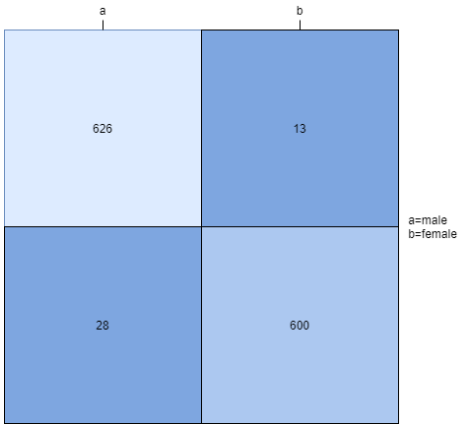
ROC Area

Tahmin sonucunda erkek ve kadın sınıfı ROC Area oranı 0,983 çıkmıştır. Bu değerin 1’e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

PRC Area

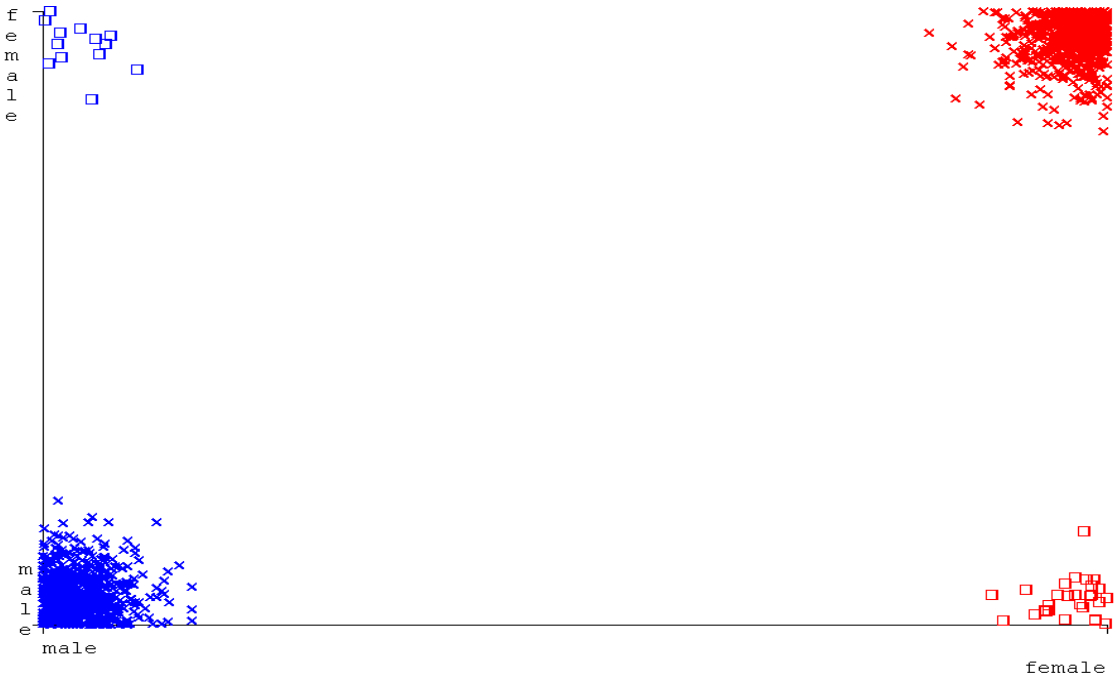
Tahmin sonucunda erkek sınıfının PRC Area oranı 0,977 ve kadın sınıfı PRC Area oranı 0,972 çıkmıştır. Bu değerin 1’e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

Confusion Matrix

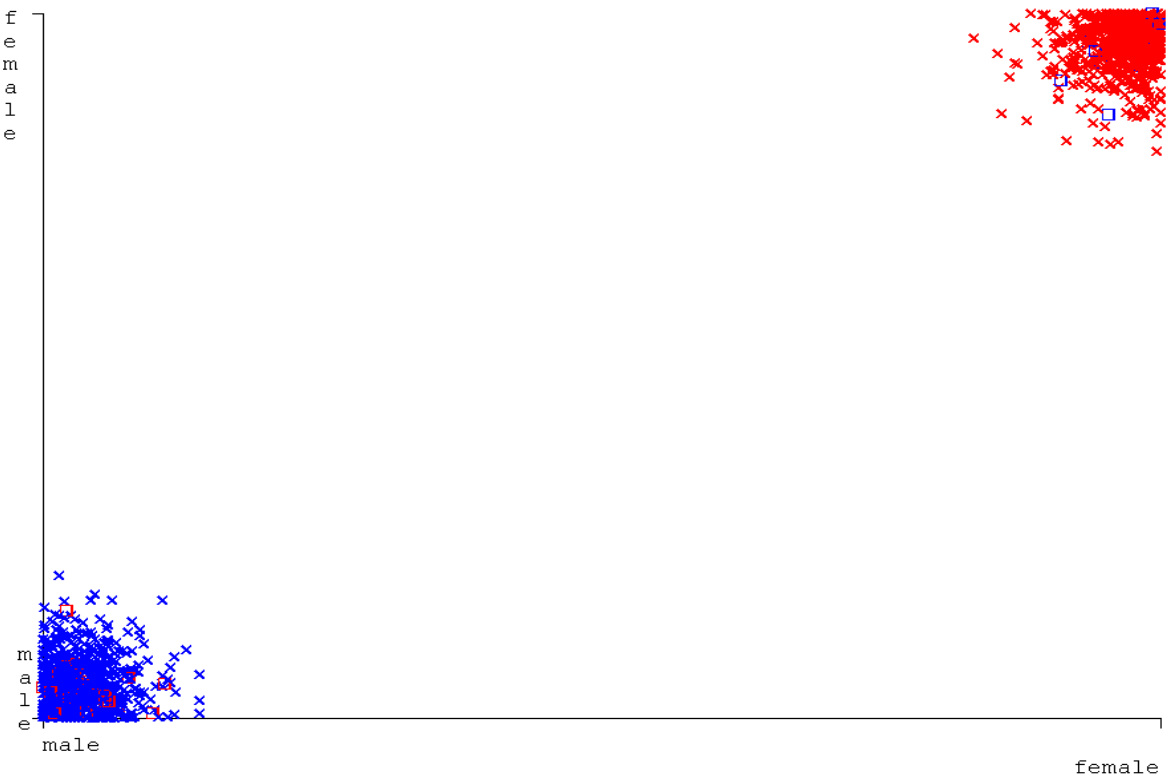


Confusion matrix kısmında oluşturduğumuz test ve eğitim kümeleri sonucunda label parametremizde bulunan male ve female tahminlerindeki başarımızı görmekteyiz. Confusion matrixde male ifadesini a olarak, female ifadesini b olarak ele almaktayız. A olarak ele aldığımız Male olan 639 verinin 626 tanesini male olarak doğru tahmin ederken 13 tanesini female olarak yanlış tahmin etmiştir. B olarak ele aldığımız female olan 628 verinin 600 tanesini female olarak doğru tahmin ederken 28 tanesini male olarak yanlış tahmin etmiştir. Tüm veri setine baktığımız zamanda 1267 veriden 1226 veriyi doğru tahmin ederken 41 veriyi yanlış tahmin etmiştir.

Yanlış Tahmin Edilen Verilerin Grafik Olarak Gösterimi



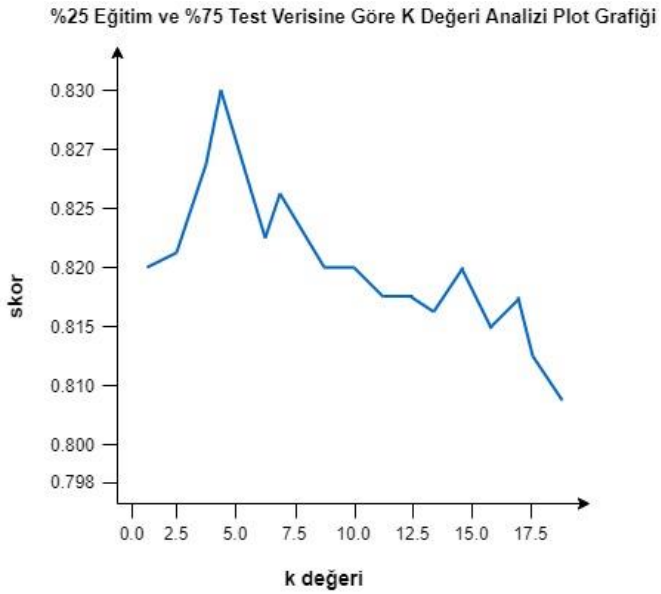
Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin birleşimi gösterilmiştir.



Tahmin modelimizin sınıflandırıcı hata grafiğine baktığımızda 28 male 13 female olmak üzere 41 hata yapıldığı görülmekte. 28 erkek çıktı kadın olarak tahmin edildiği ve 13 kadın çıktı ise erkek olarak tahmin edildiği görülmektedir. Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin kesişimi gösterilmiştir.

Örnek2

Weka programının K-En Yakın Komşu sınıflandırma yönteminde karışık veri setlerinde kullanılan ve rastgele doğru ve yanlış çıkma olasılığını düşüren cross-validation seçeneğini 4 seçerek veri setimizi dörde bölüyoruz. Bu aşamada 729 adet veriden oluşan 4 adet veri seti oluşmaktadır. Daha sonra bu 4 adet veri setinden bir tanesi seçilerek test kümesi oluşturulur. Geriye kalan 2367 adet veri eğitim kümesi olarak belirlenmektedir. Oluşturulan eğitim kümesi ile model eğitilir ve 729 adet veri bulunan test kümesi ile test edilir. Kısaca cross-validation seçeneği 4 (%25 eğitim/ %75 test) seçilerek veri setinde temel alınan Meanfun, IQR, Sd, Q25, Label parametreleri üzerinde tahmin yapılmıştır. Küme sayısını 3 olarak ele alınmıştır.



Yukarıdaki grafik veri setinin %25'i eğitim ve %75'u test verisi olduğunda seçilecek k değerlerinin çıkacak başarı değerleri gösterilmektedir. Spyder 'da grafik kodlanırken k değeri aralığı 1-20 olarak belirlenmiştir. Küme sayısı 3 seçilirse elde edilecek en yüksek başarı oranı %83 civarında olacağı 18 seçilirse en düşük başarı oranı %80 civarında olacağı görülmektedir. Bu grafik elde edildikten sonra oluşturduğumuz modelin kümesi 6 olarak belirlenmiştir.

Sonuç 2:

Başarı Oranı	***82.3864 %***
Doğru Sınıflandırılmış Örnekler	2610
Kappa İstatistiği	0.6477
Ortalama Mutlak Hata	0.1997
Kök Ortalama Kare Hatası	0.3622
Göreli Mutlak Hata	39.9369 %
Kök Göreli Kare Hatası	72.4404 %
Toplam Örnek Sayısı	3168
Küme Sayısı	3

Modeli oluşturmak için geçen süre: 0.01 saniyedir.

Correctly Classified Instances

Veri setinin %4'ü test kümesi olarak ele alındığı için 3168 satır veriden 3039'u üzerinde tahmin yapılmıştır. Bunun sonucunda 3168 adet veriden 3039 doğru tahmin ederek %82.3864 oranında başarı elde edilmiştir. Modelimiz tahmin yaparken 558 adet veriyi yanlış bulmuştur.

Kappa Statistic

Kadın ve erkek değerlerinin arasındaki karşılaştırmalı uyuşmanın güvenilirlik oranı 0.6477'dir. Elde edilen değer 0.61-0.80 aralığında olduğu için kadın ve erkek değerleri arasında önemli derecede uyuşma olduğu görülmektedir.

Mean Absolute Error

Tahmin sonucunda elde edilen ortalama mutlak hata oranı 0.1997'dir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Mean Squared Error

Tahmin sonucunda elde edilen karekök ortalama hata oranı 0.3622'dir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Relative Absolute Error

Tahmin sonucunda gerçek değer ile hesaplanan değer arasındaki farkın gerçek değere oranlanması sonucunda 39.9369 değeri elde edilmiştir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Relative Squared Error

Tahmin sonucunda kök göreceli hata oranı 72.4404 dür. Bu değer 0'a çok yakın olmasa da diğer metriklerin 0'a yakın olmasından dolayı modelin başarı oranını çok etkilememiştir.

	TP Rate	***FP Rate***	***Presicion***	***Recall***	***F-Measure***	***MCC***	***ROC Area***	***PRC Area***	***Class***
	0,850	0,202	0,808	0,850	0,828	0,649	0,887	0,836	Male
	0,798	0,150	0,842	0,798	0,819	0,649	0,887	0,867	Female
Ağırlıklı Ortalama	0,824	0,176	0,825	0,824	0,824	0,649	0,887	0,852	

TP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini erkek sesi olarak tahmin oranı 0,850'dir. Sınıfı kadın olan verilerden kadın sesini kadın sesi olarak tahmin oranı 0,798'dır. Bu değerler 1'e yakın olduğu için iyi bir isabet oranı elde edildiği görülmektedir. Yapılan K-Nearest Neighbor(KNN) sınıflandırma yöntemi sonucunda verilerin tamamına yakın bir kısmını doğru tahmin ettiği görülmektedir. Aynı zamanda kadın ve erkek sınıflarında en çok doğru tahmini yapan erkek sınıfıdır.

FP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini kadın sesi olarak tahmin etme oranı 0,202'dir. Sınıfı kadın olan verilerden kadın sesini erkek sesi olarak tahmin etme oranı 0,150'dir. Bu değerler 0'a yakın olduğu için yapılan hatalı tahminin çok az olduğu görülmektedir. Aynı zamanda erkek ve kadın sınıfları arasında en çok hata yapan erkek sınıfıdır.

Precision

Tahmin sonucunda erkek sınıfı hassasiyet oranı 0,808 iken kadın sınıfı hassasiyet oranı 0,842'dir. Hassasiyet oranları 1'e yakın olduğu için sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Sınıfı kadın olan verilerin hassasiyet oranı daha yüksektir.

Recall

Tahmin sonucunda erkek sınıfı geri çağırma oranı 0,850 iken kadın sınıfı geri çağırma oranı 0,798'dir. Örneğin gerçekte sesin erkek olduğu durumda tahminin kadın sesi olarak yapılmasıdır. Bu hata 0'a yaklaştıkça artar 1'e yaklaştıkça azalır. Bu durumda sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Kadın sınıfına ait veriler daha doğru tahmin edilmiştir.

F-Measure

Tahmin sonucunda erkek sınıfı F-Measure oranı 0,828, kadın sınıfı oranı 0,819 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

MCC

Tahmin sonucunda erkek ve kadın sınıfı MCC oranı 0,649 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

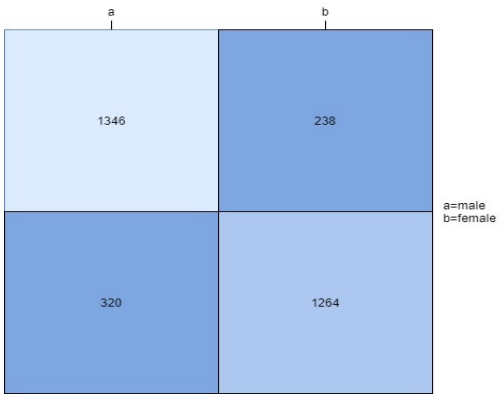
ROC Area

Tahmin sonucunda erkek ve kadın sınıfı ROC Area oranı 0,887 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

PRC Area

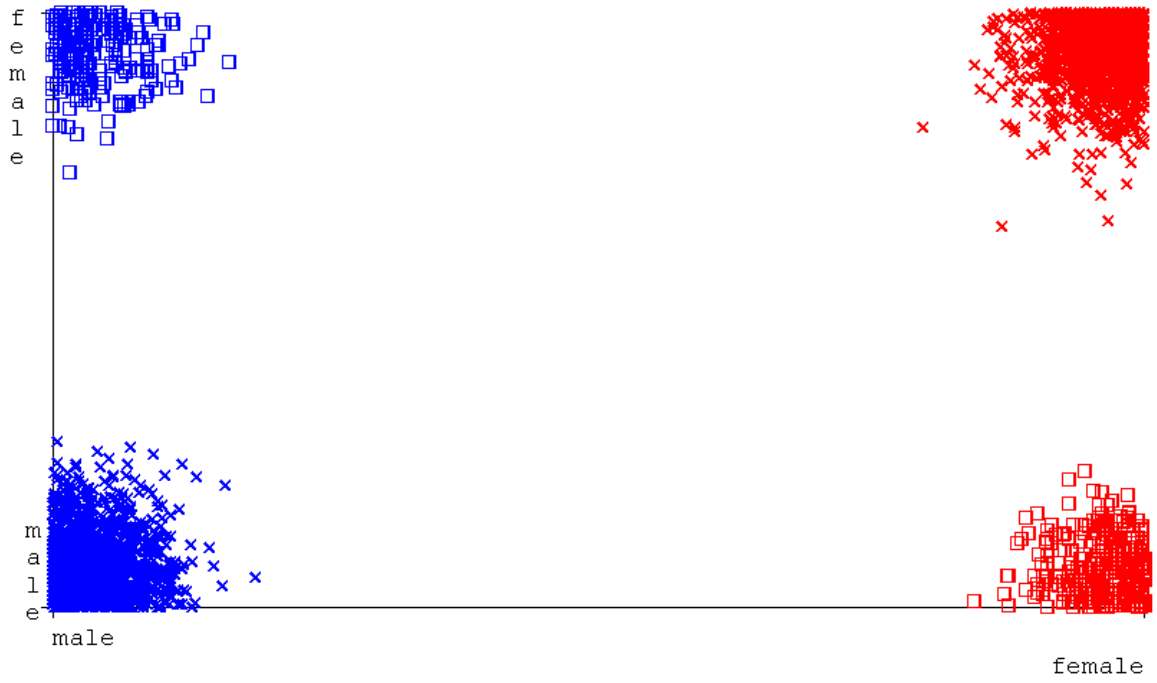
Tahmin sonucunda erkek sınıfında PRC Area oranı 0,836 iken kadın sınıfı 0,867 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

Confusion Matrix

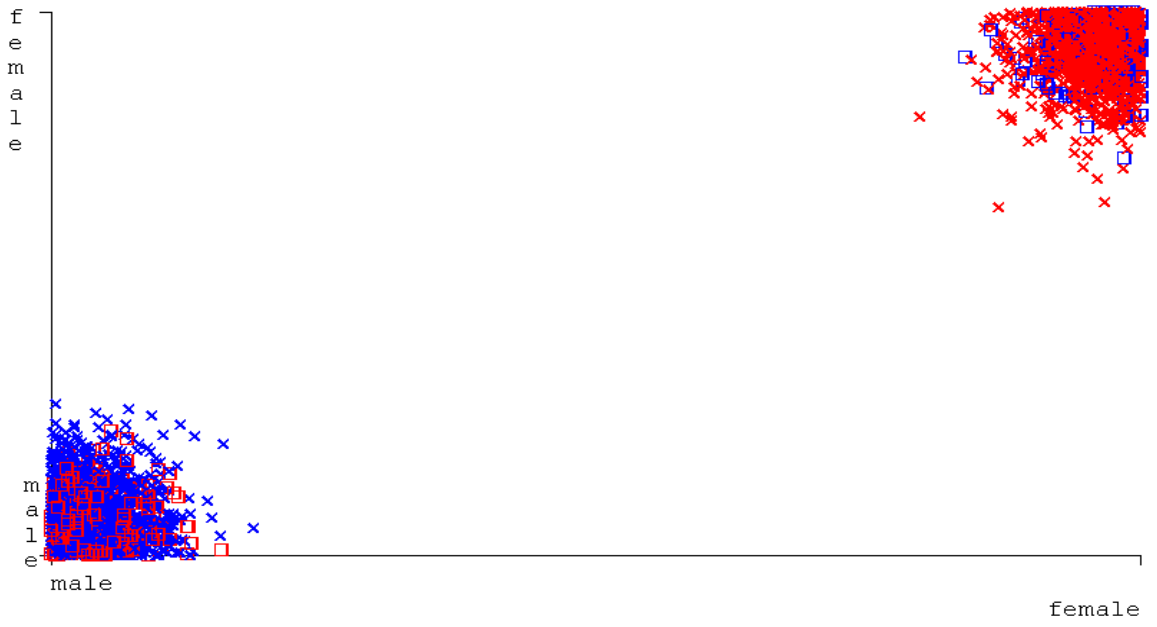


Confusion matrix kısmında oluşturduğumuz test ve eğitim kümeleri sonucunda label parametremizde bulunan male ve female tahminlerindeki başarımızı görmekteyiz. Confusion matrixde male ifadesini a olarak, female ifadesini b olarak ele almaktayız. A olarak ele aldığımız Male olan 1584 verinin 1346 tanesini male olarak doğru tahmin ederken 238 tanesini female olarak yanlış tahmin etmiştir. B olarak ele aldığımız female olan 1584 verinin 1264 tanesini female olarak doğru tahmin ederken 320 tanesini male olarak yanlış tahmin etmiştir. Tüm veri setine baktığımız zamanda 3168 veriden 2610 veriyi doğru tahmin ederken 558 veriyi yanlış tahmin etmiştir.

Yanlış Tahmin Edilen Verilerin Grafik Olarak Gösterimi



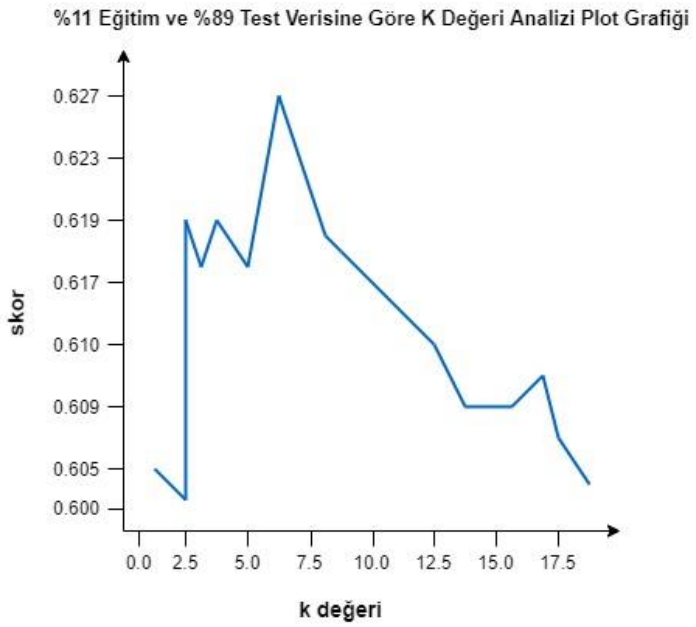
Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin birleşimi gösterilmiştir.



Tahmin modelimizin sınıflandırıcı hata grafiğine baktığımızda 238 male 320 female olmak üzere 558 hata yapıldığı görülmekte. 238 erkek çıktı kadın olarak tahmin edildiği ve 320 kadın çıktı ise erkek olarak tahmin edildiği görülmektedir. Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin kesişimi gösterilmiştir.

Örnek3

Veri setimizin içerisinden meanfun, sd, Q25, IQR, label parametrelerini temel alarak ilerledik. K-En Yakın Komşu algoritmasından makineye veri setinin sonucu bildiğimiz değerlerin %11'ini eğitim vererek %89'unu tahmin etmesini istedik. Label parametresi üzerinden de sınıflandırmasını yaptık. Küme sayısını 2 olarak ele alınmıştır.



Yukarıdaki grafik veri setinin %11'i eğitim ve %89'u test verisi olduğunda seçilecek k değerlerinin çıkacak başarı değerleri gösterilmektedir. Spyder 'da grafik kodlanırken k değeri aralığı 1-20 olarak belirlenmiştir. Küme sayısı 6 seçilirse elde edilecek en yüksek başarı oranı %62 civarında olacağı 18 seçilirse en düşük başarı oranı %60 civarında olacağı görülmektedir. Bu grafik elde edildikten sonra oluşturduğumuz modelin kümesi 6 olarak belirlenmiştir.

Sonuç3: ***62.5532 %***

Başarı Oranı

Doğru Sınıflandırılmış Örnekler	1764
Kappa İstatistiği	0.2511
Ortalama Mutlak Hata	0.4201
Kök Ortalama Kare Hatası	0.4908
Göreli Mutlak Hata	84.0106 %
Kök Göreli Kare Hatası	98.1588 %
Toplam Örnek Sayısı	2820
Küme Sayısı	6

Modeli oluşturmak için geçen süre: 0,01 saniyedir.

Correctly Classified Instances

Veri setinin %89'i test kümesi olarak ele alındığı için 3168 satır veriden 2820'si üzerinde tahmin yapılmıştır. Bunun sonucunda 3168 adet veriden 1764'ü doğru tahmin ederek %62.5532 oranında başarı elde edilmiştir. Modelimiz tahmin yaparken 1056 adet veriyi yanlış bulmuştur.

Kappa Statistic

Kadın ve erkek değerlerinin arasındaki karşılaştırmalı uyuşmanın güvenilirlik oranı 0.2511'dir. Elde edilen değer 0.21-0.40 aralığında olduğu için kadın ve erkek değerleri arasında orta derecede uyuşma olduğu görülmektedir.

Mean Absolute Error

Tahmin sonucunda elde edilen ortalama mutlak hata oranı 0.4201'dir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Mean Squared Error

Tahmin sonucunda elde edilen karekök ortalama hata oranı 0.4908'dir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Relative Absolute Error

Tahmin sonucunda gerçek değer ile hesaplanan değer arasındaki farkın gerçek değere oranlanması sonucunda 84.0106 değeri elde edilmiştir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Relative Squared Error

Tahmin sonucunda kök göreceli hata oranı 98.1588'dir. Bu değer 0'a çok yakın olmasa da diğer metriklerin 0'a yakın olmasından dolayı modelin başarı oranını çok etkilememiştir.

	TP Rate	***FP Rate***	***Presicion***	***Recall***	***F-Measure***	***MCC***	***ROC Area***	***PRC Area***	***Class***
	0,765	0,513	0,598	0,765	0,671	0,261	0,670	0,628	Male
	0,487	0,235	0,674	0,487	0,565	0,261	0,670	0,640	Female
Ağırlıklı Ortalama	0,626	0,374	0,636	0,626	0,618	0,261	0,670	0,634	

TP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini erkek sesi olarak tahmin oranı 0,765 ’dir. Sınıfı kadın olan verilerden kadın sesini kadın sesi olarak tahmin oranı 0,487’dür. Bu değerler 1’e yakın olduğu için iyi bir isabet oranı elde edildiği görülmektedir. Yapılan K-Nearest Neighbor(KNN) sınıflandırma yöntemi sonucunda verilerin tamamına yakın bir kısmını doğru tahmin ettiği görülmektedir. Aynı zamanda kadın ve erkek sınıflarında en çok doğru tahmini yapan kadın sınıfıdır.

FP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini kadın sesi olarak tahmin etme oranı 0,513’dür. Sınıfı kadın olan verilerden kadın sesini erkek sesi olarak tahmin etme oranı 0,235’dir. Bu değerler 0’a yakın olduğu için yapılan hatalı tahminin çok az olduğu görülmektedir. Aynı zamanda erkek ve kadın sınıfları arasında en çok hata yapan kadın sınıfıdır.

Precision

Tahmin sonucunda erkek sınıfı hassasiyet oranı 0,598 iken kadın sınıfı hassasiyet oranı 0,674’dir. Hassasiyet oranları 1’e yakın olduğu için sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Sınıfı erkek olan verilerin hassasiyet oranı daha yüksektir.

Recall

Tahmin sonucunda erkek sınıfı geri çağırma oranı 0,765 iken kadın sınıfı geri çağırma oranı 0,487’dir. Örneğin gerçekte sesin kadın olduğu durumda tahminin erkek sesi olarak yapılmasıdır. Bu hata 0’a yaklaştıkça artar 1’e yaklaştıkça azalır. Bu durumda sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Kadın sınıfına ait veriler daha doğru tahmin edilmiştir.

F-Measure

Tahmin sonucunda erkek F-Measure oranı 0,671iken kadın sınıfı F-Measure oranı 0,565 çıkmıştır. Bu değerin 1’e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

MCC

Tahmin sonucunda erkek ve kadın sınıfı MCC oranı 0,261 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

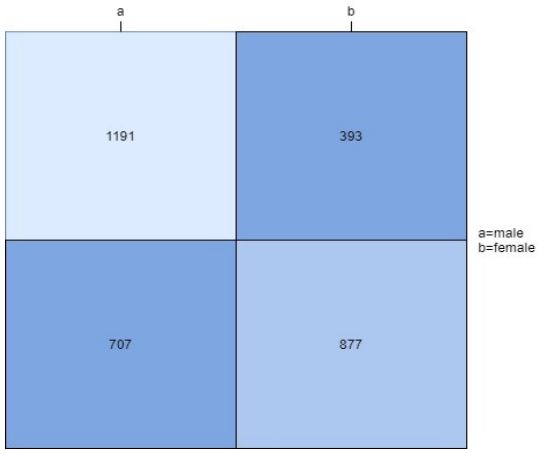
ROC Area

Tahmin sonucunda erkek ve kadın sınıfı ROC Area oranı 0,670 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

PRC Area

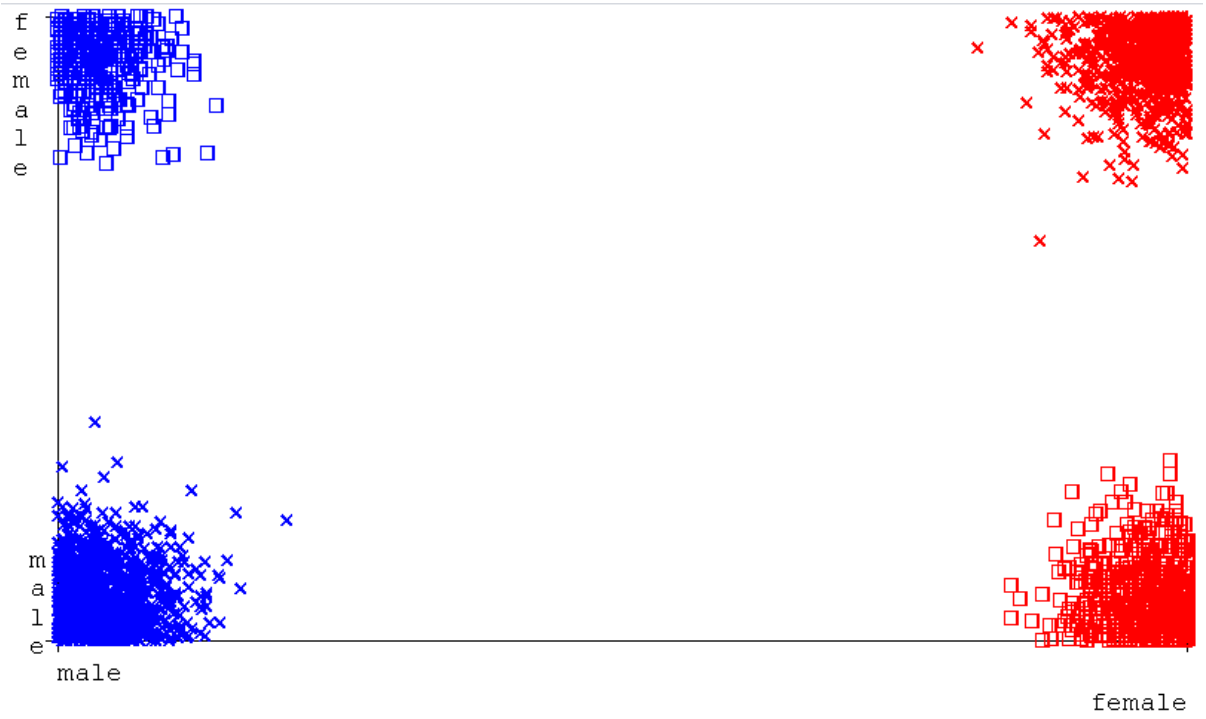
Tahmin sonucunda erkek PRC Area oranı 0,628iken kadın sınıfı PRC Area oranı 0,640 çıkmıştır. Bu değerlerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir. Aynı zamanda erkek sınıfının PRC Area oranı kadın sınıfından daha başarılıdır.

Confusion Matrix



Confusion matrix kısmında oluşturduğumuz test ve eğitim kümeleri sonucunda label parametremizde bulunan male ve famele tahminlerindeki başarımızı görmekteyiz. Confusion matrixde male ifadesini a olarak, female ifadesini b olarak ele almaktayız. A olarak ele aldığımız Male olan 1410 verinin 1078 tanesini male olarak doğru tahmin ederken 332 tanesini female olarak yanlış tahmin etmiştir. B olarak ele aldığımız female olan 1410 verinin 724 tanesini female olarak doğru tahmin ederken 686 tanesini male olarak yanlış tahmin etmiştir. Tüm veri setine baktığımız zamanda 2820 veriden 1764 veriyi doğru tahmin ederken 1056 veriyi yanlış tahmin etmiştir.

Yanlış Tahmin Edilen Verilerin Grafik Olarak Gösterimi



Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin birleşimi gösterilmiştir.

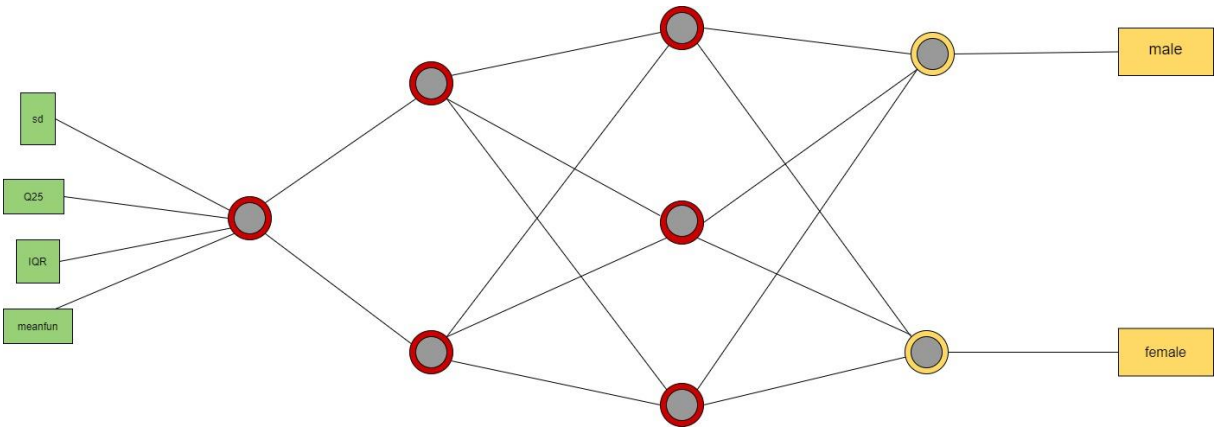
Tahmin modelimizin sınıflandırıcı hata grafiğine baktığımızda 393 male 707 female olmak üzere 1100 hata yapıldığı görülmekte. 393 erkek çıktı kadın olarak tahmin edildiği ve 707 kadın

çıktı ise erkek olarak tahmin edildiği görülmektedir. Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin kesişimi gösterilmiştir.

3.5.4.Yapay Sinir Ağları-Çok Katmanlı Algılayıcı Sınıflandırma Yöntemi

Örnek1:

Veri setimizin içerisinden meanfun, sd, Q25, IQR, label parametrelerini temel alarak ilerledik. Yapay Sinir Ağları Çok Katmanlı Algılayıcı algoritmasından makineye veri setinin sonucu bildiğimiz değerlerin %60'ını eğitim vererek %40 unu tahmin etmesini istedik. Label parametresi üzerinden de sınıflandırmasını yaptık. Örnekte momentum değeri 0.5, learningRate değeri 0.4, hiddenLayers değeri 1,2,3, trainingTime değeri 100 alınmıştır.



Sonuç1:

Başarı Oranı	***97.6322%***
Doğru Sınıflandırılmış Örnekler	1237
Kappa İstatistiği	0.9526
Ortalama Mutlak Hata	0.0475
Kök Ortalama Kare Hatası	0.1398
Göreli Mutlak Hata	9.5072%
Kök Göreli Kare Hatası	27.9628%
Toplam Örnek Sayısı	1267
Momentum	0.5
LearningRate	0.4

HiddenLayers	1,2,3
TrainingTime	100

Modeli oluşturmak için geçen süre: 0 saniyedir.

Correctly Classified Instances

Veri setinin %40'ı yani 1267'si test kümesi kalan 1901'i eğitim kümesi olarak ele alınarak tahmin yapılmıştır. Bunun sonucunda 1267 adet veriden 1237'sini doğru tahmin ederek %97.6322 oranında başarı elde edilmiştir. Modelimiz tahmin yaparken 30 adet veriyi yanlış bulmuştur.

Kappa Statistic

Kadın ve erkek değerlerinin arasındaki karşılaştırmalı uyuşmanın güvenilirlik oranı 0.9526'dır. Elde edilen değer 0.81-1.00 aralığında olduğu için kadın ve erkek değerleri arasında neredeyse önemli derecede bir uyuşma olduğu görülmektedir.

Mean Absolute Error

Tahmin sonucunda elde edilen ortalama mutlak hata oranı 0.0475'dir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Mean Squared Error

Tahmin sonucunda elde edilen karekök ortalama hata oranı 0.1398'dir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Relative Absolute Error

Tahmin sonucunda gerçek değer ile hesaplanan değer arasındaki farkın gerçek değere oranlanması sonucunda 9.5072 değeri elde edilmiştir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Relative Squared Error

Tahmin sonucunda kök göreceli hata oranı 27.9628'dir. Bu değer 0'a çok yakın olmasa da diğer metriklerin 0'a yakın olmasından dolayı modelin başarı oranını çok etkilememiştir.

	TP Rate	*FP Rate*	*Presicion*	*Recall*	*F-Measure*	*MCC*	*ROC Area*	*PRC Area*	*Class*
	0.972	0.019	0.981	0.972	0.976	0.953	0.993	0.994	Male
	0.981	0.028	0.972	0.981	0.976	0.953	0.993	0.989	Female
Ağırlıklı Ortalama	0.976	0.024	0.976	0.976	0.976	0.953	0.993	0.992	

TP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini erkek sesi olarak tahmin oranı 0,972'dir. Sınıfı kadın olan verilerden kadın sesini kadın sesi olarak tahmin oranı 0,981'dir. Bu değerler 1'e yakın olduğu için iyi bir isabet oranı elde edildiği görülmektedir. Yapılan Yapay Sinir Ağları Çok Katmanlı Algılayıcı yöntemi sonucunda verilerin tamamına yakın bir kısmını doğru tahmin ettiği görülmektedir. Aynı zamanda kadın ve erkek sınıflarında en çok doğru tahmini yapan erkek sınıfıdır.

FP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini kadın sesi olarak tahmin etme oranı 0,019'dur. Sınıfı kadın olan verilerden kadın sesini erkek sesi olarak tahmin etme oranı 0,028'dir. Bu değerler 0'a yakın olduğu için yapılan hatalı tahminin çok az olduğu görülmektedir. Aynı zamanda erkek ve kadın sınıfları arasında en çok hata yapan erkek sınıfıdır.

Precision

Tahmin sonucunda erkek sınıfı hassasiyet oranı 0,981 iken kadın sınıfı hassasiyet oranı 0,972'dir. Hassasiyet oranları 1'e yakın olduğu için sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Sınıfı erkek olan verilerin hassasiyet oranı daha yüksektir.

Recall

Tahmin sonucunda erkek sınıfı geri çağırma oranı 0,972 iken kadın sınıfı geri çağırma oranı 0,981'dir. Örneğin gerçekte sesin kadın olduğu durumda tahminin erkek sesi olarak yapılmasıdır. Bu hata 0'a yaklaştıkça artar 1'e yaklaştıkça azalır. Bu durumda sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Erkek sınıfına ait veriler daha doğru tahmin edilmiştir.

F-Measure

Tahmin sonucunda erkek ve kadın sınıfı F-Measure oranı 0,976 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir. Erkek sınıfına ait verilerde daha fazla başarı elde edilmiştir.

MCC

Tahmin sonucunda erkek ve kadın sınıfı MCC oranı 0,953 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

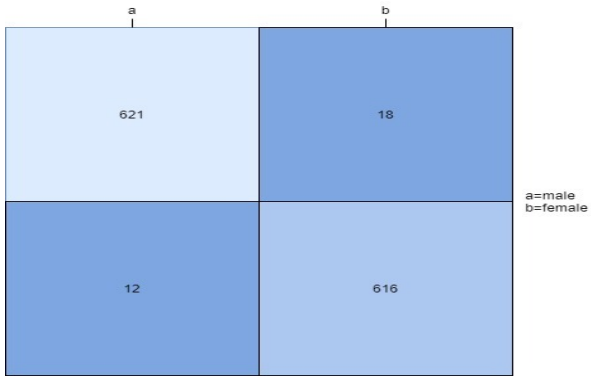
ROC Area

Tahmin sonucunda erkek ve kadın sınıfı ROC Area oranı 0,993 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

PRC Area

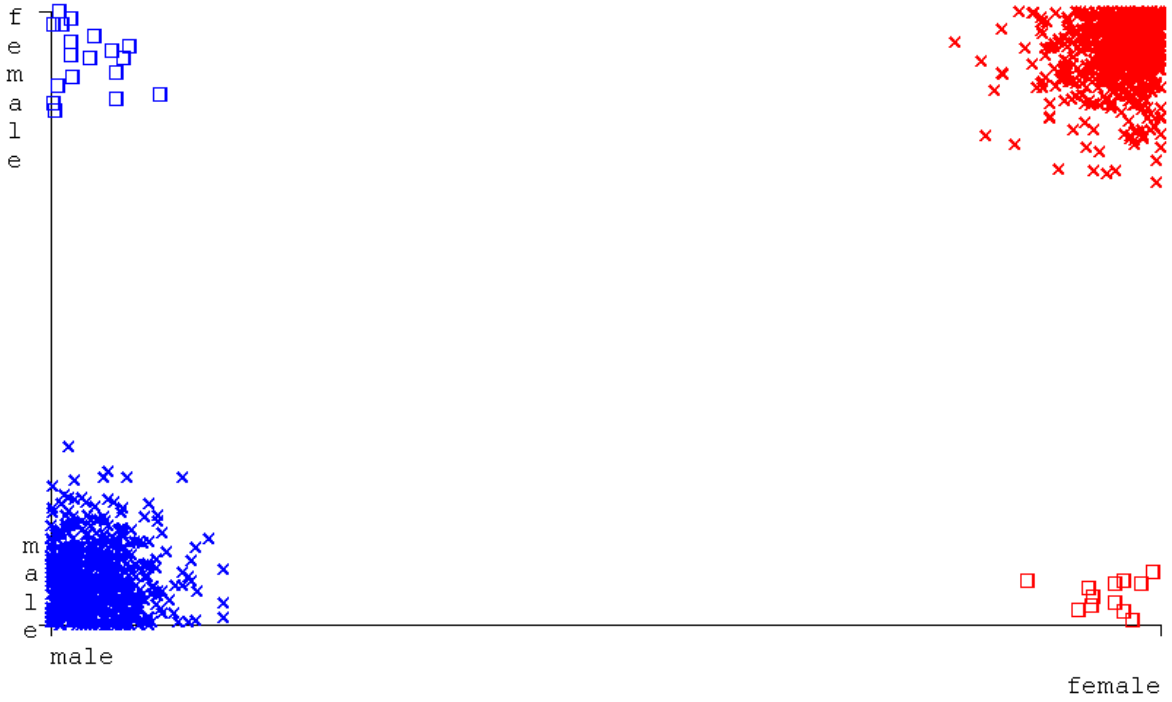
Tahmin sonucunda erkek sınıfı PRC Area oranı 0,984 iken kadın sınıfı PRC Area oranı 0.989 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir. Kadın sınıfına ait verilerde daha fazla başarı elde edilmiştir.

Confusion Matrix

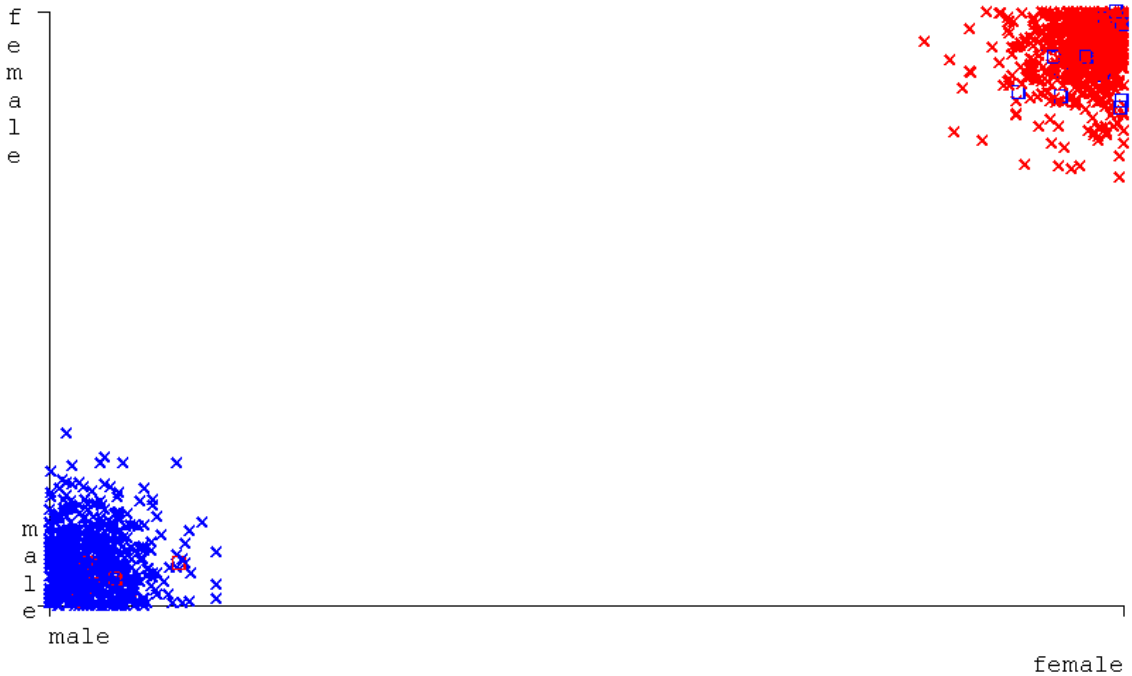


Confusion matrix kısmında oluşturduğumuz test ve eğitim kümeleri sonucunda label parametremizde bulunan male ve famele tahminlerindeki başarımızı görmekteyiz. Confusion matrixde male ifadesini a olarak, female ifadesini b olarak ele almaktayız. A olarak ele aldığımız Male olan 639 verinin 621 tanesini male olarak doğru tahmin ederken 18 tanesini female olarak yanlış tahmin etmiştir. B olarak ele aldığımız female olan 628 verinin 616 tanesini female olarak doğru tahmin ederken 12 tanesini male olarak yanlış tahmin etmiştir. Tüm veri setine baktığımız zamanda 1267 veriden 1237 veriyi doğru tahmin ederken 30 veriyi yanlış tahmin etmiştir.

Yanlış Tahmin Edilen Verilerin Grafik Olarak Gösterimi



Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin birleşimi gösterilmiştir.

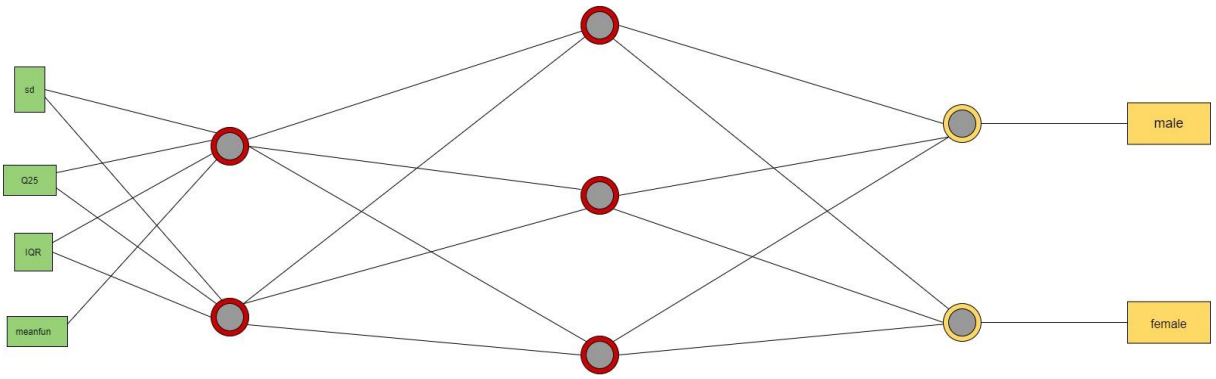


Tahmin modelimizin sınıflandırıcı hata grafiğine baktığımızda 18 male 12 female olmak üzere 30 hata yapıldığı görülmekte. 18 erkek çıktı kadın olarak tahmin edildiği ve 12 kadın çıktı ise

erkek olarak tahmin edildiği görülmektedir. Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin kesişimi gösterilmiştir.

Örnek2

Weka programının Yapay Sinir Ağları Çok Katmanlı Algılayıcı sınıflandırma yönteminde karışık veri setlerinde kullanılan ve rastgele doğru ve yanlış çıkma olasılığını düşüren cross-validation seçeneğini 4 seçerek veri setimizi dörde bölüyoruz. Bu aşamada 729 adet veriden oluşan 4 adet veri seti oluşmaktadır. Daha sonra bu 4 adet veri setinden bir tanesi seçilerek test kümesi oluşturulur. Geriye kalan 2367 adet veri eğitim kümesi olarak belirlenmektedir. Oluşturulan eğitim kümesi ile model eğitilir ve 729 adet veri bulunan test kümesi ile test edilir. Kısaca cross-validation seçeneği 4 (%25 eğitim/ %75 test) seçilerek veri setinde temel alınan Meanfun, IQR, Sd, Q25, Label parametreleri üzerinde tahmin yapılmıştır. Örnekte momentum değeri 1, learningRate değeri 0.5, hiddenLayers değeri 2,3 , trainingTime değeri 50 alınmıştır.



Sonuç2:

Başarı Oranı	***50 %***
Doğru Sınıflandırılmış Örnekler	1584
Kappa İstatistiği	0
Ortalama Mutlak Hata	0.5
Kök Ortalama Kare Hatası	0.7071
Göreli Mutlak Hata	100 %
Kök Göreli Kare Hatası	141.4214%
Toplam Örnek Sayısı	3168
Momentum	1
LearningRate	0.5

HiddenLayers	2,3
TrainingTime	50

Modeli oluşturmak için geçen süre: 0.01 saniyedir.

Correctly Classified Instances

Veri setinin %4'ü test kümesi olarak ele alındığı için 3168 satır veriden 3039'u üzerinde tahmin yapılmıştır. Bunun sonucunda 3168 adet veriden 3039 doğru tahmin ederek %50 oranında başarı elde edilmiştir. Modelimiz tahmin yaparken 1584 adet veriyi yanlış bulmuştur.

Kappa Statistic

Kadın ve erkek değerlerinin arasındaki karşılaştırmalı uyuşmanın güvenilirlik oranı 0' dır. Elde edilen değer 0.0-0.20 aralığında olduğu için kadın ve erkek değerleri arasında önemsiz uyuşma olduğu görülmektedir.

Mean Absolute Error

Tahmin sonucunda elde edilen ortalama mutlak hata oranı 0.5'dir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Mean Squared Error

Tahmin sonucunda elde edilen karekök ortalama hata oranı 0.7071'dir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Relative Absolute Error

Tahmin sonucunda gerçek değer ile hesaplanan değer arasındaki farkın gerçek değere oranlanması sonucunda 100 değeri elde edilmiştir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Relative Squared Error

Tahmin sonucunda kök göreceli hata oranı 141.4214'dür. Bu değer 0'a çok yakın olmasa da diğer metriklerin 0'a yakın olmasından dolayı modelin başarı oranını çok etkilememiştir.

	TP Rate	*FP Rate*	*Presicion*	*Recall*	*F-Measure*	*MCC*	*ROC Area*	*PRC Area*	*Class*
	0,750	0,750	0,500	0,750	0,600	0,000	0,500	0,500	Male
	0,250	0,250	0,500	0,250	0,333	0,000	0,500	0,500	Female
Ağırlıklı Ortalama	0,500	0,500	0,500	0,500	0,467	0,000	0,500	0,500	

TP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini erkek sesi olarak tahmin oranı 0,750'dir. Sınıfı kadın olan verilerden kadın sesini kadın sesi olarak tahmin oranı 0,250'dir. Bu değerler 1'e yakın olduğu için iyi bir isabet oranı elde edildiği görülmektedir. Yapılan Yapay Sinir Ağları Çok Katmanlı Algılayıcı sınıflandırma yöntemi sonucunda verilerin tamamına yakın bir kısmını doğru tahmin ettiği görülmektedir. Aynı zamanda kadın ve erkek sınıflarında en çok doğru tahmini yapan erkek sınıfıdır.

FP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini kadın sesi olarak tahmin etme oranı 0,750'dir. Sınıfı kadın olan verilerden kadın sesini erkek sesi olarak tahmin etme oranı 0,250'dir. Bu değerler 0'a yakın olduğu için yapılan hatalı tahminin çok az olduğu görülmektedir. Aynı zamanda erkek ve kadın sınıfları arasında en çok hata yapan erkek sınıfıdır.

Precision

Tahmin sonucunda erkek ve kadın sınıfı hassasiyet oranı 0,500'dür. Hassasiyet oranları 1'e yakın olduğu için sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Sınıfı kadın olan verilerin hassasiyet oranı daha yüksektir.

Recall

Tahmin sonucunda erkek sınıfı geri çağırma oranı 0,750 iken kadın sınıfı geri çağırma oranı 0,250'dir. Örneğin gerçekte sesin erkek olduğu durumda tahminin kadın sesi olarak yapılmasıdır. Bu hata 0'a yaklaştıkça artar 1'e yaklaştıkça azalır. Bu durumda sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Kadın sınıfına ait veriler daha doğru tahmin edilmiştir.

F-Measure

Tahmin sonucunda erkek sınıfı F-Measure oranı 0,600, kadın sınıfı oranı 0,333 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

MCC

Tahmin sonucunda erkek ve kadın sınıfı MCC oranı 0,000 çıkmıştır. Bu değerin 0'a yakın olması yapılan tahmin modelinin çok başarılı olmadığını göstermektedir.

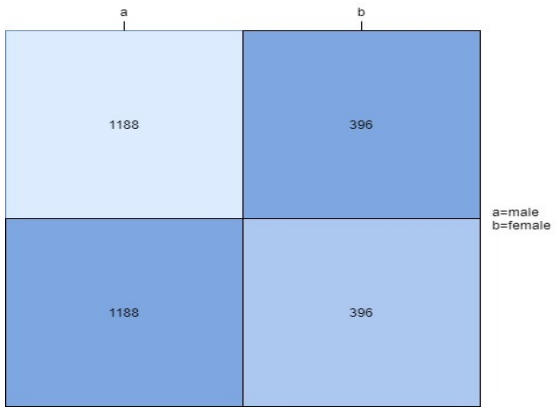
ROC Area

Tahmin sonucunda erkek ve kadın sınıfı ROC Area oranı 0,500 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

PRC Area

Tahmin sonucunda erkek ve kadın sınıfında PRC Area oranı 0,500 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

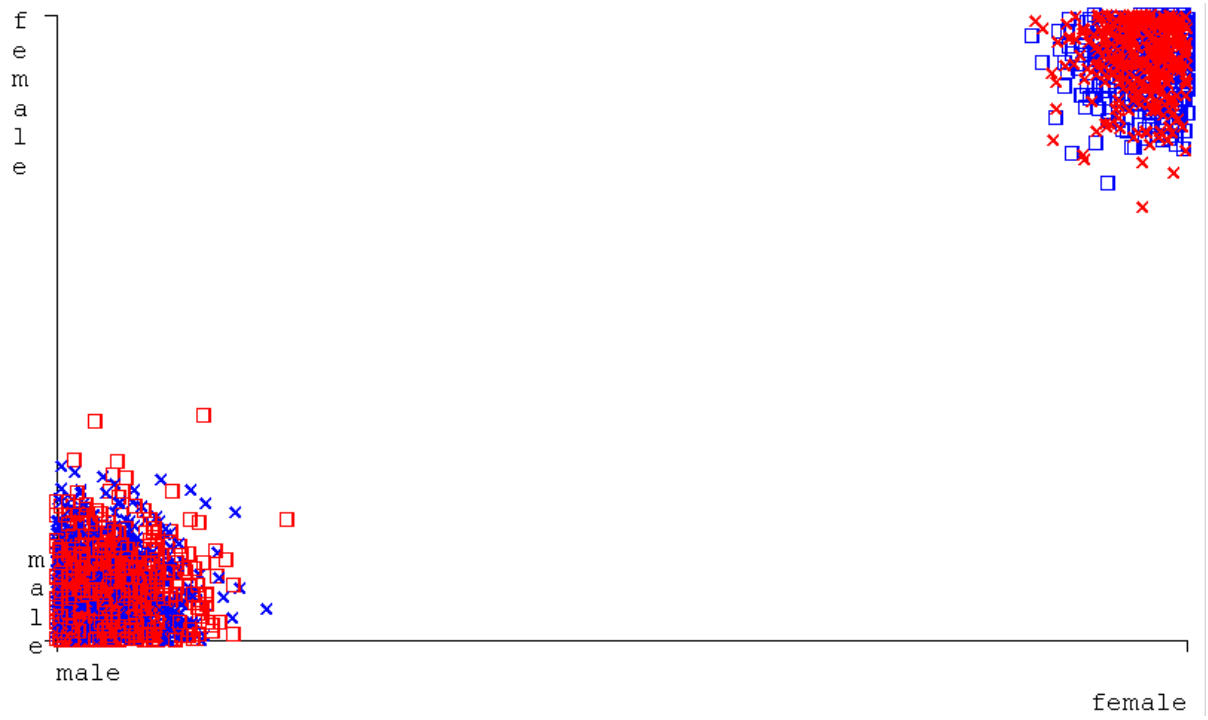
Confusion Matrix



Confusion matrix kısmında oluşturduğumuz test ve eğitim kümeleri sonucunda label parametremizde bulunan male ve female tahminlerindeki başarımızı görmekteyiz. Confusion matrixde male ifadesini a olarak, female ifadesini b olarak ele almaktayız. A olarak ele aldığımız Male olan 1584 verinin 1188 tanesini male olarak doğru tahmin ederken 396 tanesini female olarak yanlış tahmin etmiştir. B olarak ele aldığımız female olan 1584 verinin 1188 tanesini female olarak doğru tahmin ederken 396 tanesini male olarak yanlış tahmin etmiştir. Tüm veri setine baktığımız zamanda 3168 veriden 2376 veriyi doğru tahmin ederken 792 veriyi yanlış tahmin etmiştir.

Yanlış Tahmin Edilen Verilerin Grafik Olarak Gösterimi

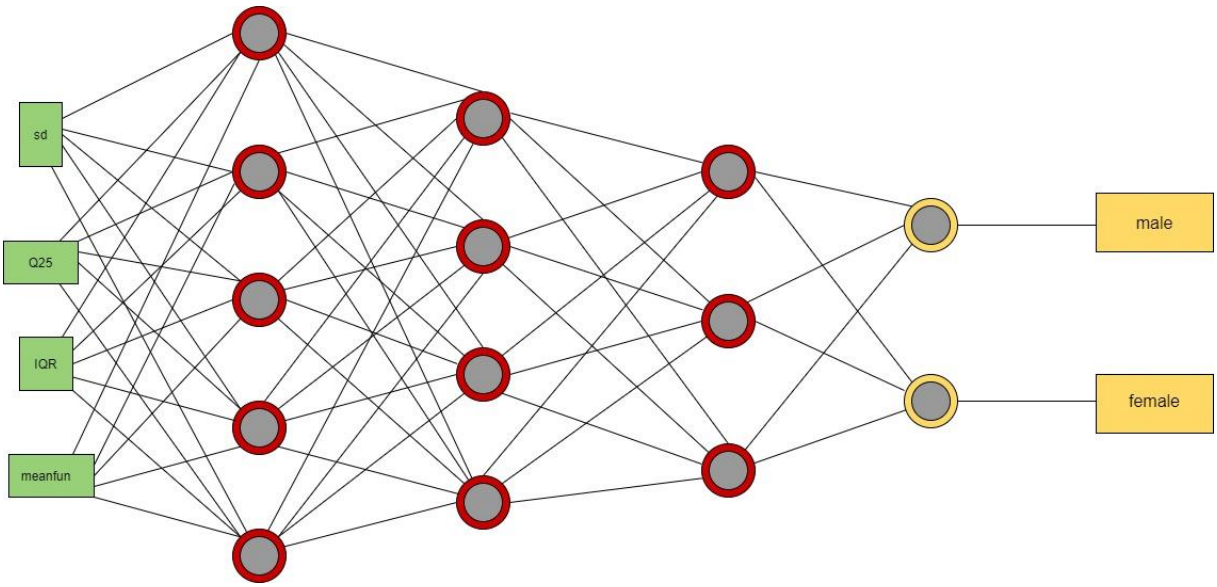
Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin birleşimi gösterilmiştir.



Tahmin modelimizin sınıflandırıcı hata grafiğine baktığımızda 396 male 396 female olmak üzere 792 hata yapıldığı görülmekte. 396 erkek çıktı kadın olarak tahmin edildiği ve 396 kadın çıktı ise erkek olarak tahmin edildiği görülmektedir. Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin kesişimi gösterilmiştir.

Örnek3

Veri setimizin içerisinden meanfun, sd, Q25, IQR, label parametrelerini temel alarak ilerledik. Yapay Sinir Ağları Çok Katmanlı Algılayıcı algoritmasından makineye veri setinin sonucu bildiğimiz değerlerin %11'ini eğitim vererek %89'unu tahmin etmesini istedik. Label parametresi üzerinden de sınıflandırmasını yaptık. Örnekte momentum değeri 0.7, learningRate değeri 0.4, hiddenLayers değeri 5,4,3 trainingTime değeri 35 alınmıştır.



Sonuç3:

Başarı Oranı	***88.7589 %***
Doğru Sınıflandırılmış Örnekler	2503
Kappa İstatistiği	0.7752
Ortalama Mutlak Hata	0.366
Kök Ortalama Kare Hatası	0.3798
Göreli Mutlak Hata	73.1974 %
Kök Göreli Kare Hatası	75.9644 %
Toplam Örnek Sayısı	2820
Momentum	0.7
LearningRate	0.4
HiddenLayers	5,4,3
TrainingTime	35

Modeli oluşturmak için geçen süre: 0,01 saniyedir.

Correctly Classified Instances

Veri setinin %89'i test kümesi olarak ele alındığı için 3168 satır veriden 2820'si üzerinde tahmin yapılmıştır. Bunun sonucunda 2820 adet veriden 2503'ünü doğru tahmin ederek % 88.7589 oranında başarı elde edilmiştir. Modelimiz tahmin yaparken 1056 adet veriyi yanlış bulmuştur.

Kappa Statistic

Kadın ve erkek değerlerinin arasındaki karşılaştırmalı uyuşmanın güvenilirlik oranı 0.7752'dir. Elde edilen değer 0.61-0.80 aralığında olduğu için kadın ve erkek değerleri arasında önemli derecede uyuşma olduğu görülmektedir.

Mean Absolute Error

Tahmin sonucunda elde edilen ortalama mutlak hata oranı 0.366'dır. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Mean Squared Error

Tahmin sonucunda elde edilen karekök ortalama hata oranı 0.3798'dir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Relative Absolute Error

Tahmin sonucunda gerçek değer ile hesaplanan değer arasındaki farkın gerçek değere oranlanması sonucunda 73.1974 değeri elde edilmiştir. Bu değer 0'a yakın olduğu için tahmin edilen modelde başarı elde edilmiştir.

Root Relative Squared Error

Tahmin sonucunda kök göreceli hata oranı 75.9644'dir. Bu değer 0'a çok yakın olmasa da diğer metriklerin 0'a yakın olmasından dolayı modelin başarı oranını çok etkilememiştir.

	TP Rate	***FP Rate***	***Presicion***	***Recall***	***F-Measure***	***MCC***	***ROC Area***	***PRC Area***	***Class***
	0,891	0,116	0,885	0,891	0,888	0,775	0,942	0,907	Male
	0,884	0,109	0,891	0,884	0,887	0,775	0,942	0,959	Female
Ağırlıklı Ortalama	0,888	0,112	0,888	0,888	0,888	0,775	0,942	0,933	

TP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini erkek sesi olarak tahmin oranı 0,891 ’dir. Sınıfı kadın olan verilerden kadın sesini kadın sesi olarak tahmin oranı 0,884’dür. Bu değerler 1’e yakın olduğu için iyi bir isabet oranı elde edildiği görülmektedir. Yapılan Yapay Sinir Ağları Çok Katmanlı Algılayıcı sınıflandırma yöntemi sonucunda verilerin tamamına yakın bir kısmını doğru tahmin ettiği görülmektedir. Aynı zamanda kadın ve erkek sınıflarında en çok doğru tahmini yapan kadın sınıfıdır.

FP Rate

Tahmin sonucunda sınıfı erkek olan verilerden erkek sesini kadın sesi olarak tahmin etme oranı 0116’dır. Sınıfı kadın olan verilerden kadın sesini erkek sesi olarak tahmin etme oranı 0,109’dur. Bu değerler 0’a yakın olduğu için yapılan hatalı tahminin çok az olduğu görülmektedir. Aynı zamanda erkek ve kadın sınıfları arasında en çok hata yapan kadın sınıfıdır.

Precision

Tahmin sonucunda erkek sınıfı hassasiyet oranı 0,885 iken kadın sınıfı hassasiyet oranı 0,891’dur. Hassasiyet oranları 1’e yakın olduğu için sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Sınıfı erkek olan verilerin hassasiyet oranı daha yüksektir.

Recall

Tahmin sonucunda erkek sınıfı geri çağırma oranı 0,891 iken kadın sınıfı geri çağırma oranı 0,884’dür. Örneğin gerçekte sesin kadın olduğu durumda tahminin erkek sesi olarak yapılmasıdır. Bu hata 0’a yaklaştıkça artar 1’e yaklaştıkça azalır. Bu durumda sınıflandırma modelinin doğru tahmin yaptığı görülmektedir. Kadın sınıfına ait veriler daha doğru tahmin edilmiştir.

F-Measure

Tahmin sonucunda erkek F-Measure oranı 0,888 iken kadın sınıfı F-Measure oranı 0,887 çıkmıştır. Bu değerin 1’e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

MCC

Tahmin sonucunda erkek ve kadın sınıfı MCC oranı 0,775 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

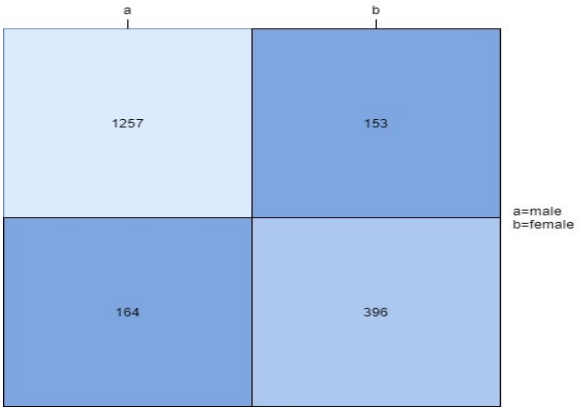
ROC Area

Tahmin sonucunda erkek ve kadın sınıfı ROC Area oranı 0,942 çıkmıştır. Bu değerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir.

PRC Area

Tahmin sonucunda erkek PRC Area oranı 0,907 iken kadın sınıfı PRC Area oranı 0,959 çıkmıştır. Bu değerlerin 1'e yakın olması yapılan tahmin modelinin başarılı olduğunu göstermektedir. Aynı zamanda erkek sınıfının PRC Area oranı kadın sınıfından daha başarılıdır.

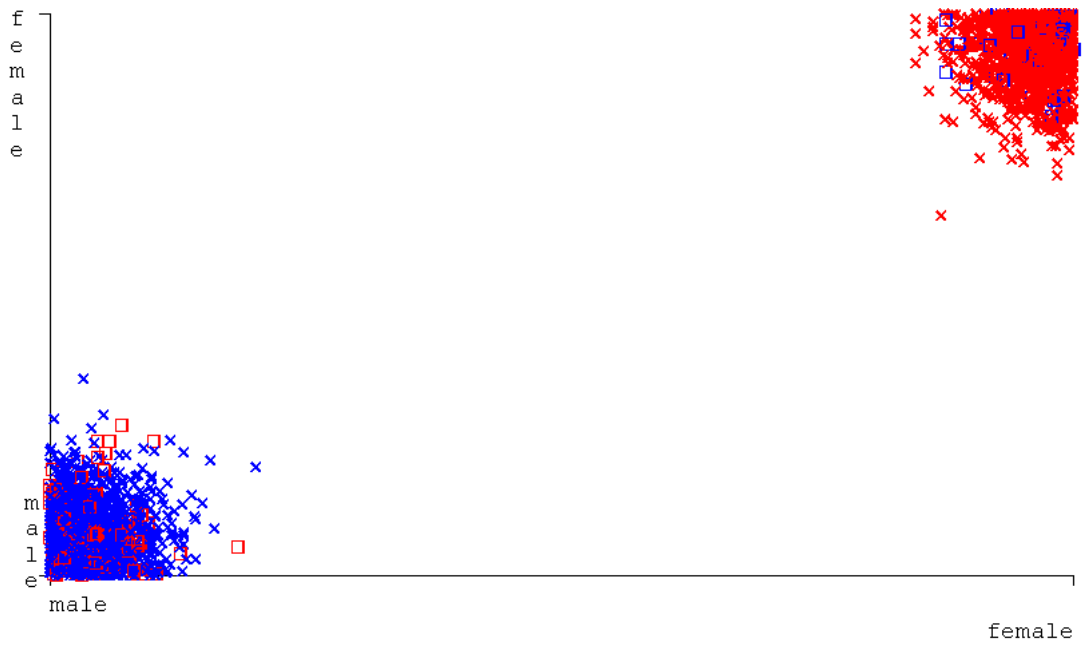
Confusion Matrix



Confusion matrix kısmında oluşturduğumuz test ve eğitim kümeleri sonucunda label parametremizde bulunan male ve famele tahminlerindeki başarımızı görmekteyiz. Confusion matrixde male ifadesini a olarak, female ifadesini b olarak ele almaktayız. A olarak ele aldığımız Male olan 1410 verinin 1257 tanesini male olarak doğru tahmin ederken 153 tanesini female olarak yanlış tahmin etmiştir. B olarak ele aldığımız female olan 1410 verinin 1246 tanesini female olarak doğru tahmin ederken 164 tanesini male olarak yanlış tahmin etmiştir. Tüm veri setine baktığımız zamanda 2820 veriden 2503 veriyi doğru tahmin ederken 317 veriyi yanlış tahmin etmiştir.

Yanlış Tahmin Edilen Verilerin Grafik Olarak Gösterimi

Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin birleşimi gösterilmiştir.



Tahmin modelimizin sınıflandırıcı hata grafiğine baktığımızda 153 male 164 female olmak üzere 111 hata yapıldığı görülmekte. 153 erkek çıktı kadın olarak tahmin edildiği ve 164 kadın çıktı ise erkek olarak tahmin edildiği görülmektedir. Yukarıdaki grafikte yapılan yanlış ve doğru tahminlerin kesişimi gösterilmiştir.

3.6.KODLAR

3.6.1.Karar Ağacı Sınıflandırma Yöntemi Spyder Kodları

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

data = pd.read_csv("ses.csv")

data.drop(["meanfreq","median","Q75","skew","kurt","sp.ent","sfm"
           ,"mode","centroid","minfun","maxfun","meandom",
           "mindom","maxdom","dfrange","modindx"],axis=1,inplace=True)
data.label=[1 if each=="male" else 0 for each in data.label]

y=data.label.values
x_data=data[["sd","meanfun","Q25","IQR"]]

#normalizasyon
x=(x_data-np.min(x_data))/(np.max(x_data)-np.min(x_data))

#train test split
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.6,random_state=1)

#knn
from sklearn.tree import DecisionTreeClassifier
dtc=DecisionTreeClassifier()
dtc.fit(x_train,y_train)

print(dtc.score(x_test,y_test))
prediction=dtc.predict(x_test)

from sklearn import tree
fn=["sd","meanfun","Q25","IQR"]
cn=['Male', 'Female']
fig, axes = plt.subplots(nrows = 1,ncols = 1,figsize = (4,4), dpi=500)
tree.plot_tree(dtc,feature_names = fn, class_names=cn,filled = True);
```

```
from sklearn import metrics
cnf_matrix = metrics.confusion_matrix(y_test,prediction)

import seaborn as sns
plt.figure(figsize=(9,9))
sns.heatmap(cnf_matrix, annot=True, fmt=".0f", linewidths=.5, square = True, cmap = 'Blues_r');
plt.ylabel('Gerçek');
plt.xlabel('Tahmin');
all_sample_title = 'Accuracy Score: {0}'.format(metrics.accuracy_score(y_test, prediction))
plt.title(all_sample_title, size = 15);

print("Doğruluk-Accuracy:",metrics.accuracy_score(y_test, prediction))
print("Hassasiyet-Precision:",metrics.precision_score(y_test, prediction))
print("Doğru tanımlama oranı-Recall:",metrics.recall_score(y_test, prediction))
```

Bu karar ağacı sınıflandırma yöntemi için yazılan kodlar ilk örneği temel alınarak yazılmıştır. Diğer iki örnek aynı kodlar üzerinde çalıştırılabilmektedir.

3.6.2.Karar Ağacı ve Naive Bayes Sınıflandırma Yöntemlerinin Aralık Değerlerine Göre Kadın-Erkek Dağılımını Gösteren Spyder Kodu

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
plt.style.use("seaborn-talk")

ses=pd.read_csv("ses.csv",sep=",")
#%%
veri1=ses[(ses['sd']>= 0.04) & (ses['sd']< 0.06)]
veri1
#%%

veri2=veri1[(veri1['label']=="male")]
mv=veri2.label.count()

veri4=veri1[(veri1['label']=="female")]
fv=veri4.label.count()
```

Bu işlem temel alınan girdi değerlerinde her aralığa uygulanmıştır.

3.6.3 Naive Bayes Sınıflandırma Yöntemi Spyder Kodları

```
import pandas as pd
import numpy as np


data = pd.read_csv("ses.csv")

data.drop(["meanfreq","median","Q75","skew","kurt","sp.ent","sfm"
           ,"mode","centroid","minfun","maxfun","meandom",
           "mindom","maxdom","dfrange","modindx"],axis=1,inplace=True)
data.label=[1 if each=="male" else 0 for each in data.label]

y=data.label.values
x_data=data[["sd","meanfun","Q25","IQR"]]

#normalizasyon
x=(x_data-np.min(x_data))/(np.max(x_data)-np.min(x_data))

#train test split
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.6,random_state=1)

#knn
from sklearn.naive_bayes import GaussianNB
nb=GaussianNB()
nb.fit(x_train,y_train)

prediction=nb.predict(x_test)

print(nb.score(x_test,y_test))


from sklearn import metrics
cnf_matrix = metrics.confusion_matrix(y_test,prediction)
```

```
import seaborn as sns
import matplotlib.pyplot as plt
plt.figure(figsize=(9,9))
sns.heatmap(cnf_matrix, annot=True, fmt=".0f", linewidths=.5, square = True, cmap = 'Blues_r');
plt.ylabel('Gerçek');
plt.xlabel('Tahmin');
all_sample_title = 'Accuracy Score: {0}'.format(metrics.accuracy_score(y_test, prediction))
plt.title(all_sample_title, size = 15);

print("Doğruluk-Accuracy:",metrics.accuracy_score(y_test, prediction))
print("Hassasiyet-Precision:",metrics.precision_score(y_test, prediction))
print("Doğru tanımlama oranı-Recall:",metrics.recall_score(y_test, prediction))
```

3.6.4 K-En Yakın Komşu Yöntemi Spyder Kodları

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

data = pd.read_csv("ses.csv")

data.drop(["meanfreq","median","Q75","skew","kurt","sp.ent","sfm"
           ,"mode","centroid","minfun","maxfun","meandom",
           "mindom","maxdom","dfrange","modindx"],axis=1,inplace=True)
data.label=[1 if each=="male" else 0 for each in data.label]
#%%
y=data.label.values
x_data=data.drop(["label"],axis=1)

x=(x_data-np.min(x_data))/(np.max(x_data)-np.min(x_data))

from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.6,random_state=1)

from sklearn.neighbors import KNeighborsClassifier
knn=KNeighborsClassifier(n_neighbors=2)
knn.fit(x_train,y_train)

prediction=knn.predict(x_test)

print("{} n-neighbors score: {}".format(3,knn.score(x_test,y_test)))

score_listesi=[]
for each in range(1,20):
    knn2=KNeighborsClassifier(n_neighbors=each)
    knn2.fit(x_train,y_train)
    score_listesi.append(knn2.score(x_test,y_test))
```

```
from sklearn import metrics
cnf_matrix = metrics.confusion_matrix(y_test,prediction)

import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(9,9))
sns.heatmap(cnf_matrix, annot=True, fmt=".0f", linewidths=.5, square = True, cmap = 'Blues_r');
plt.ylabel('Gerçek');
plt.xlabel('Tahmin');
all_sample_title = 'Accuracy Score: {0}'.format(metrics.accuracy_score(y_test, prediction))
plt.title(all_sample_title, size = 15);

print("Doğruluk-Accuracy:",metrics.accuracy_score(y_test, prediction))
print("Hassasiyet-Precision:",metrics.precision_score(y_test, prediction))
print("Doğru tanımlama oranı-Recall:",metrics.recall_score(y_test, prediction))
```

Yukarıdaki kodlar Naive Bayes Örnek1'i temel alınarak yazılmıştır. Kodlar test ve eğitim verisine göre hangi küme sayılarının en yüksek veya en düşük başarı oranı vereceğini gösteren grafik çizdirmek için kullanılmıştır. Aynı zamanda modelin başarı oranı da analiz edilmiştir.

```
from sklearn import datasets
import pandas as pd
import matplotlib.pyplot as plt

ses=pd.read_csv("ses.csv",sep=",")
data=ses.loc[:,["sd","meanfun","IQR","Q25"]]

def yeni(x):
    if (x=="Male"):
        return 1
    else:
        return 0

ses["gercek_deger"]=ses["label"].apply(yeni)
#%%

plt.scatter(data.sd,data.meanfun,data.Q25,data.IQR)
plt.show()

# %% KMEANS

from sklearn.cluster import KMeans
wcss = []

for k in range(1,10):
    kmeans = KMeans(n_clusters=k)
    kmeans.fit(data)
    wcss.append(kmeans.inertia_)

plt.plot(range(1,10),wcss)
plt.xlabel("k degerleri yani kümeler (cluster)")
plt.ylabel("wcss")
plt.show()
```

```
#%% k = 2 icin modelim

kmeans2 = KMeans(n_clusters=2)
clusters = kmeans2.fit_predict(data)


#%%dendogram

from scipy.cluster.hierarchy import linkage,dendrogram
merg=linkage(data,method="ward")
dendrogram(merg,leaf_rotation=90)
plt.xlabel("data points")
plt.ylabel("uzaklık öklit uzaklığı")
plt.show()
```

K-En Yakın Komşu sınıflandırma yönteminde saçılım (scatter) grafiği, dendogram ve kol grafiği yukarıdaki kodlar ile çizdirilmiştir.

3.6.5 Veri Dağılımının Violin ve Kutu Grafiği Spyder Kodları

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

ses=pd.read_csv("ses.csv",sep=",")

fig, (ax, ax2, ax3, ax4) = plt.subplots(ncols=4,figsize=(15, 5))

ax.violinplot(ses.sd)
ax.set_title("SD")

ax2.violinplot(ses.meanfun)
ax2.set_title("Menfun")

ax3.violinplot(ses.Q25)
ax3.set_title("Q25")

ax4.violinplot(ses.IQR)
ax4.set_title("IQR")

plt.show()
```

Yukarıdaki kodlar violin grafiği çizilirken kullanılmıştır.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

ses=pd.read_csv("ses.csv",sep=",")

fig, (ax, ax2, ax3, ax4) = plt.subplots(ncols=4,figsize=(15, 5))

ax.boxplot(ses.sd)
ax.set_title("SD")

ax2.boxplot(ses.meanfun)
ax2.set_title("Menfun")

ax3.boxplot(ses.Q25)
ax3.set_title("Q25")

ax4.boxplot(ses.IQR)
ax4.set_title("IQR")

plt.show()
```

Yukarıdaki kodlar kutu grafiği çizilirken kullanılmıştır.

3.7.UYGULAMA SONUÇLARININ KARŞILAŞTIRILMASI

3.7.1.Karar Ağacı Sınıflandırma Yöntemi Sonuçların Karşılaştırılması

	SONUÇ 1	*SONUÇ 2*	*SONUÇ 3*
EĞİTİM VE TEST ORANI	*%60-%40*	*%25-%75*	*%11-%89*
DOĞRU TAHMİN SAYISI	1229	3063	2654
YANLIŞ TAHMİN SAYISI	38	105	166
TOPLAM VERİ SAYISI	1267	3168	2820
YAPRAK VE DAL SAYISI	15-29	12-23	6-11
MODEL OLUŞTURMAK İÇİN HARCANAN ZAMAN	0.03 saniye	0.01 saniye	0.01 Saniye
YAPRAK BAŞINA DÜŞEN MİNİMUM OBJE SAYISI	2	10	45
BAŞARI ORANI	%97.0008	%96.6856	%94.1153
KAPPA İSTATİSTİK	0.94	0.9337	0.7932
ORTALAMA MUTLAK HATA	0.0446	0.0477	0.0953
KÖK ORTALAMA KARE HATASI	0.1626	0.1692	0.2333
GÖRELİ MUTLAK HATA	8.9196	9.5454	19.0553
KÖK GÖRELİ KARE HATASI	32.5157	33.8499	46.6636
ORTALAMA GERÇEK POZİTİF ORAN	0.970	0.967	0.941
ORTALAMA YANLIŞ POZİTİF ORAN	0.030	0.033	0.060
ORTALAMA HASSASİYET ORANI	0.970	0.967	0.944
ORTALAMA GERİ BİLDİRME	0.970	0.967	0.941
ORTALAMA F-ÖLÇÜSÜ	0.970	0.967	0.941
ORTALAMA MCC	0.940	0.934	0.885
ORTALAMA ROC AREA	0.978	0.987	0.941
ORTALAMA RPC AREA	0.974	0.983	0.916

Yapılan örneklerin sonuçları incelendiğinde eğitim seti oranı daha fazla olan yani örnek1'in başarı oranı daha fazla olduğu görülmektedir. Buradan şu çıkarımı yapabiliriz;

- Tahmin modelinin eğitim verisi oranını arttırdıkça elde edilecek başarı oranı da artar.

- Eğitim seti oranı düşük olan sınıflandırma modelinde hata oranı artar.
- Yaprak başına düşen minimum obje sayısı arttırıldıkça karışık veri setlerinde başarı oranı artarken karışık olmayan veri setlerinde bu oran düşer fakat model oluşturma süresi her iki veri setinde de düşer.
- Hata oranları arttıkça başarı oranı düşer.
- Gerçek pozitif oran, hassasiyet oranı, geri bildirme oranı, f-ölçüsü oranı, mcc oranı, roc area oranı ve rpc area oranı 1'e yaklaştıkça modelin başarısı artar. 0'a yaklaştıkça ise bu oran düşer.
- Yanlış pozitif oran 0'a yaklaştıkça modelin başarı oranı artarken 1'e yaklaştığında bu oran düşer.

3.7.2.Navie Bayes Sınıflandırma Yöntemi Sonuçların Karşılaştırılması

	SONUÇ 1	***SONUÇ 2***	***SONUÇ 3***
EĞİTİM VE TEST ORANI	*%60-%40*	*%25-%75*	*%11-%89*
DOĞRU TAHMİN SAYISI	1221	3039	2709
YANLIŞ TAHMİN SAYISI	46	129	111
TOPLAM VERİ SAYISI	1267	3168	2820
MODEL OLUŞTURMAK İÇİN HARCANAN ZAMAN	0.0 saniye	0.0 saniye	0.01 Saniye
BAŞARI ORANI	%96.3694	%95.928	%96.0638
KAPPA İSTATİSTİK	0.9274	0.9186	0.9213
ORTALAMA MUTLAK HATA	0.0436	0.0485	0.0479
KÖK ORTALAMA KARE HATASI	0.1701	0.1826	0.1795
GÖRELİ MUTLAK HATA	8.7107%	9.6935%	9.5875%
KÖK GÖRELİ KARE HATASI	34.0126%	36.5224%	35.9024%
ORTALAMA GERÇEK POZİTİF ORAN	0.964	0.959	0.961
ORTALAMA YANLIŞ POZİTİF ORAN	0.036	0.041	0.039
ORTALAMA HASSASİYET ORANI	0.964	0.960	0.961
ORTALAMA GERİ BİLDİRME	0.964	0.970	0.961
ORTALAMA F-ÖLÇÜSÜ	0.964	0.959	0.961

ORTALAMA MCC	0.928	0.919	0.922
ORTALAMA ROC AREA	0.993	0.991	0.991
ORTALAMA RPC AREA	0.993	0.991	0.991

Yapılan örneklerin sonuçları incelendiğinde eğitim seti oranı daha fazla olan yani örnek1'in başarı oranı daha fazla olduğu görülmektedir. Buradan şu çıkarımı yapabiliriz;

- Tahmin modelinin eğitim verisi oranını arttırdıkça elde edilecek başarı oranı da artar.
- Kappa istatistik oranı azaldıkça başarı oranı azalır.
- Ortalama mutlak hata, kök ortalama kare hata, göreli mutlak hata ve kök göreli kare hatası arttıkça başarı oranı azalır.
- Ortalama gerçek pozitif oranı ve ortalama hassasiyet oranı azaldıkça başarı oranı azalır.
- Ortalama yanlış pozitif oran arttıkça başarı oranı azalır.
- Hata oranları arttıkça başarı oranı düşer.
- Geri bildirme oranı, f-ölçüsü oranı ve mcc oranı 1'e yaklaştıkça modelin başarısı artar. 0'a yaklaştıkça ise bu oran düşer.

3.7.3.K-En Yakın Komşu Sınıflandırma Yöntemi Sonuçların Karşılaştırılması

	SONUÇ 1	***SONUÇ 2***	***SONUÇ 3***
EĞİTİM VE TEST ORANI	*%60-%40*	*%25-%75*	*%11-%89*
DOĞRU TAHMİN SAYISI	1226	2610	1764
YANLIŞ TAHMİN SAYISI	41	1008	1056
TOPLAM VERİ SAYISI	1267	3168	2820
MODEL OLUŞTURMAK İÇİN HARCANAN ZAMAN	0.44 saniye	0.1saniye	0.01 Saniye
BAŞARI ORANI	%96.764	%82.3864	%62.5532
KÜME SAYISI	2	3	6
KAPPA İSTATİSTİK	0.9353	0.6477	0.2511
ORTALAMA MUTLAK HATA	0.0295	0.1997	0.4201
KÖK ORTALAMA KARE HATASI	0.1513	0.3622	0.4908
GÖRELİ MUTLAK HATA	5.8898%	39.9369%	84.0106 %
KÖK GÖRELİ KARE HATASI	30.3503%	72.4404%	98.1588 %
ORTALAMA GERÇEK POZİTİF ORAN	0.968	0.824	0,626

ORTALAMA YANLIŞ POZİTİF ORAN	0.033	0.176	0,374
ORTALAMA HASSASİYET ORANI	0.968	0.825	0.636
ORTALAMA GERİ BİLDİRME	0.968	0.824	0.626
ORTALAMA F-ÖLÇÜSÜ	0.968	0.824	0.618
ORTALAMA MCC	0.936	0.649	0.261
ORTALAMA ROC AREA	0.983	0.887	0.670
ORTALAMA RPC AREA	0.974	0.852	0.634

Yapılan örneklerin sonuçları incelendiğinde eğitim seti oranı daha fazla olan yani örnek1'in başarı oranı daha fazla olduğu görülmektedir. Buradan şu çıkarımı yapabiliriz;

- Tahmin modelinin eğitim verisi oranını arttırdıkça elde edilecek başarı oranı da artar.
- Küme sayısı değerinin uygun değerde seçilmesi oldukça önem taşımaktadır. Küme sayısı arttıkça başarı oranı azalır.
- Küme sayısı azaldıkça model oluşturma süresi artar.
- Küme sayısı arttıkça yanlış tahmin sayısı artar.
- Küme sayısı arttıkça başarı oranı ve kappa istatistik oranı azalır.
- Kök göreceli kare hatası arttıkça başarı oranı azalır.
- Göreli mutlak hata arttıkça başarı oranı azalır.
- Ortalama gerçek pozitif oranı ve ortalama yanlış pozitif oran arttıkça başarı oranı azalır.
- Ortalama hassasiyet oranı, ortalama geri bildirme, ortalama f-ölçüsü, ortalama mcc, ortalama roc area, ortalama rpc area oranı azaldıkça başarı oranı da azalır.

3.7.4.Yapay Sinir Ağları-Çok Katmanlı Algılayıcı Sınıflandırma Yöntemi Sonuçların Karşılaştırılması

	SONUÇ 1	***SONUÇ 2***	***SONUÇ 3***
EĞİTİM VE TEST ORANI	*%60-%40*	*%25-%75*	*%11-%89*
DOĞRU TAHMİN SAYISI	1237	1584	2503
YANLIŞ TAHMİN SAYISI	30	1584	317
TOPLAM VERİ SAYISI	1267	3168	2820
MOMENTUM	0.5	1	0.7
LEARNİNGRATE	0.4	0.5	0.4
HİDDENLAYERS	1,2,3	2,3	5,4,3
TRAİNİNGTİME	100	50	35

MODEL OLUŞTURMAK İÇİN HARCANAN ZAMAN	0.00 saniye	5.7 saniye	2.57 Saniye
BAŞARI ORANI	%97.6322	%50	%88.7589
KAPPA İSTATİSTİK	0.9526	0	0.7752
ORTALAMA MUTLAK HATA	0.0475	0.5	0.366
KÖK ORTALAMA KARE HATASI	0.1398	0.7071	0.3798
GÖRELİ MUTLAK HATA	9.5072%	100%	73.1974 %
KÖK GÖRELİ KARE HATASI	27.9628%	141.4214%	75.9644%
ORTALAMA GERÇEK POZİTİF ORAN	0.976	0.500	0,888
ORTALAMA YANLIŞ POZİTİF ORAN	0.024	0.500	0,112
ORTALAMA HASSASİYET ORANI	0.976	0.500	0.888
ORTALAMA GERİ BİLDİRME	0.976	0.500	0.888
ORTALAMA F-ÖLÇÜSÜ	0.976	0.467	0.888
ORTALAMA MCC	0.953	0.000	0.775
ORTALAMA ROC AREA	0.993	0.500	0.942
ORTALAMA RPC AREA	0.992	0.500	0.933

Yapılan örneklerin sonuçları incelendiğinde eğitim seti oranı daha fazla olan yani örnek1'in başarı oranı daha fazla olduğu görülmektedir. Buradan şu çıkarımı yapabiliriz;

- Tahmin modelinin eğitim verisi oranını arttırdıkça elde edilecek başarı oranı da artar.
- Trainingtime değeri, kappa istatistik ve kök ortalama kare hatası oranı azaldıkça başarı oranı azalır.
- Katman sayısı arttırıldığında adım sayısı düşük olduğunda model düşük başarı oranı verir.
- Momentum sayısı yüksek olduğunda yerel çözümler kabul edilebilir hata düzeyinin altına düşmez bunun için modelin başarı oranı düşer.

- Momentum katsayısı ile öğrenme katsayısı arasındaki fark arttıkça modelin öğrenme süresi de artar.

- Öğrenme katsayısı gereğinden yüksek olduğunda problem uzayında rasgele gezinme olur ve bu da başarı oranını düşürür.
- Kök göreli kare hatası arttıkça başarı oranı azalır.
- Ortalama gerçek pozitif oran azaldıkça başarı oranı azalır.
- Kök göreceli kare hatası, göreli mutlak hata oranı ve ortalama yanlış pozitif oran arttıkça başarı oranı azalır.
- Ortalama gerçek pozitif oranı ve ortalama hassasiyet, ortalama geri bildirme, ortalama f-ölçüsü, ortalama mcc, ortalama roc area, ortalama rpc area oranı azaldıkça başarı oranı azalır.

4.SONUÇ VE ÖNERİ

Bu projede Veri Madenciliği konusunda bir altyapı oluşturmak ve ses tanıma alanlarında Veri Madenciliği'nin kullanımı ile ilgili örnekler sunarak karar verme süreçleri açısından yeni bir bakış açısı kazandırmak amaçlanmıştır. Veri Madenciliği'nin ses tanıma uygulama alanlarında kullanımını bu örnekler ile sınırlamak mümkün değildir. Kaggle üzerinden eriştiğimiz veri setine WEKA veri madenciliği ve Spyder yazılımı ile çeşitli algoritmalar uygulanmıştır. Bu kapsamda personelin sisteme yanlış veya uç değer olarak girdiği verilerin teker teker kontrol edilip yanlış veya uç verilerin en az olduğu giriş değerleri temel alınır. Bütün önişleme süreçlerinin tamamlanmasından sonra açık kaynak kodlu veri madenciliği yazılı olan WEKA ve Spyder ile veri seti üzerinde çeşitli algoritmalar uygulanarak modeller oluşturulmuş ve buna göre en başarılı algoritma olarak Yapay Sinir Ağları-Çok Katmanlı Algılayıcı algoritması bulunmuştur. Kullanılan algoritmaların eğitim-test ve başarı oranları göz önüne alındığında Yapay Sinir Ağları-Çok Katmanlı Algılayıcının %60 eğitim - %40 test verisi olarak model oluşturulduğunda %97.6322 en yüksek başarı oranı elde edilmiştir. İkinci en iyi başarı değeri Karar Ağacı algoritmasında %60 eğitim-%40 test verisi ile %97.0008 olarak elde edilmiştir. Yapay Sinir Ağları-Çok Katmanlı Algılayıcı %25 eğitim-%75 test verisi olarak model oluşturulduğunda %50 en düşük başarı oranı elde edilmiştir. Çok Katmanlı Algılayıcı algoritmasında başarı değerlerinin oranları sadece test-eğitim setinin farklı olmasından değil katman sayısı, momentum katsayısı, öğrenme katsayısı ve adım sayısının farklı olmasından değişiklik göstermiştir. Çok Katmanlı Algılayıcı algoritması gibi kullandığımız diğer algoritmalarda kendi içerisinde birçok özelliğe göre sınıflandırılmıştır. Model oluşturmada kullanılan algoritmaların karşılaştırılması sonucunda en başarılı olan Çok Katmanlı Algılayıcı algoritması ve bu veri seti kullanılarak

ilerleyen çalışmalarda cinsiyete göre ses tanıma çalışma alanlarına yönelik bir uygulama geliştirilmesi ve cinsiyet tespit sürelerinin kısalması çalışanlara fikir vermesi olabilir.

Çalışma kapsamında cinsiyete göre ses tanıma veri setinde 21 adet farklı parametreden 4 parametre girdi ve 1 parametre çıktı değişkeni olarak belirlenmiştir. Yapılacak diğer çalışmalarda farklı parametreler girdi değişkeni olarak veri madenciliği modelinde kullanılabilir ve elde edilecek sonuçlar incelenebilir. Çalışmada kullanılacak girdi değerleri uç veriler barındırmamalıdır.

Sınıflandırma algoritmaları kullanılmadan önce veri seti iyi analiz edilmeli ve aykırı veriler belirlenmelidir. Belirlenen aykırı veriler temizlenmeli veya kullanacağınız girdi değerlerinde aykırı veri oranı düşük olmalıdır. Verilerin düzenli dağılım göstermesi başarı oranını arttıracaktır.

Çalışma kapsamında kullanılan her sınıflandırma algoritması kendi içerisinde 3 farklı şekilde analiz yaparak sonuçlar üretmiştir. Yapılacak çalışmalarda veri setinde çok daha fazla analiz yaparak daha iyi sonuçlar üretilebilir.

Çalışmada kullanılan veri seti 4 yıl öncesine aittir yapılacak çalışmalarda hazır veri seti kullanıldığı durumlarda veri setinin güncel olmasına dikkat edilmelidir.

Çalışma kapsamında veri madenciliği modelleri içerisinde yer alan tahmin edici modellerden olan sınıflandırma tekniklerinden Karar Ağacı, K-En Yakın Komşu, Naive Bayes ve Yapay Sinir Ağları-Çok Katmanlı Algılayıcı kullanılarak cinsiyet tahmin edilmeye çalışılmıştır. Yapılacak diğer çalışmalarda diğer sınıflandırma teknikleri (Regresyon, Zaman Serisi Analizleri, Genetik Algoritmalar, Kestirim) kullanılarak öğrenci başarıları tahmin edilmeye çalışılabilir.

Kullanılacak sınıflandırma teknikleri veri setinin büyüklüğüne, karmaşıklığına, girdi sayısına, çıktı sayısına ve konusu gibi birçok özelliğe göre değişiklik gösterir. Bu durumda veri setinin çok iyi analiz edilmesi ve doğru sınıflandırma yöntemlerinin seçilmesi gerekir.

Veri madenciliği sürecinin herkes tarafından daha kolay ve hızlı gerçekleştirilebilmesi için ses tanıma ile ilgili özellikler geliştirilerek kaydedilen verilerin ses tanıma üzerinde veri madenciliği sürecine tabi tutulması sağlanabilir. Bankamatiklerde ses ile müşteri bilgileri ilişkilendirilerek işlem yapılabilir. Bankamatiklerde ses ile giriş ve işlem yapabilme kullanıcıların hem işlem yapmasını kolaylaştırır ve güvenlik artar. Bankamatiklerde gerçekleştirilecek işlemlerin

bankada gerçekleştirilmesi banka yoğunluğunu arttırmakta ve kullanıcıların diğer işlemlerini sekteye uğratmaktadır. Bu alanda yapılacak çalışmalar etkin ve verimli olabilir.

5.KAYNAKÇA

1. Alpaydın, E. (2013). Yapay Öğrenme. İstanbul: MIT, BÜTEK
2. Pacheco, P. (2011). An Introduction to Parallel Programming. Burlington: Morgan Kaufmann Publishers is an imprint of Elsevier.
3. Şeyda DEMİREL, Selay GİRAY YAKUT "KARAR AĞACI ALGORİTMALARI VE ÇOCUK İŞÇİLİĞİ ÜZERİNE BİR UYGULAMA" Sosyal Bilimler Araştırma Dergisi Yıl 2019, Cilt 8 , Sayı 4, Sayfalar 52 – 65
4. Neslihan KÖSE, Filiz ERSÖZ "VERİ MADENCİLİĞİNDE KARAR AĞACI ALGORİTMALARI İLE DEMİR ÇELİK ENDÜSTRİSİNDE İŞ KAZALARI ÜZERİNE BİR UYGULAMA" Avrupa Bilim ve Teknoloji Dergisi Yıl 2020, Cilt , Sayı , Sayfalar 397 – 407
5. Ersan OKATAN , Ali Hakan IŞIK "SAĞLIK HARCAMALARININ TAHMİNİNDE KARAR AĞACININ KULLANIMI" Mehmet Akif Ersoy Üniversitesi Fen Bilimleri Enstitüsü Dergisi Yıl 2020, Cilt 11 , Sayı 1, Sayfalar 86 – 94
6. Nihat Barış SEBİK, Halil İbrahim BÜLBÜL "VERİ MADENCİLİĞİ MODELLERİNİN AKCİĞER KANSERİ VERİ SETİ ÜZERİNDE BAŞARILARININ İNCELENMESİ" TÜBAV Bilim Dergisi Yıl 2018, Cilt 11, Sayı 3, Sayfalar 1 – 7
7. U. Tuğba GÜRSOY, Şafiye BİLGİN "BANKA MÜŞTERİLERİNİN İNTERNET BANKACILIĞINA İLİŞKİN YAKLAŞIMLARININ VERİ MADENCİLİĞİ TEKNİKLERİ İLE İNCELENMESİ" Kafkas Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi Yıl 2016, Cilt 7, Sayı 14, Sayfalar 421 – 442
8. Muhammed Resul AYDIN "YAPAY SİNİR AĞLARI İLE TALEP TAHMİNİ: PERAKENDE SEKTÖRÜNDE BİR UYGULAMA" İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi Yıl 2019, Cilt 18, Sayı 35, Sayfalar 43 – 55
9. Pakize ERDOĞMUŞ, Buket ÇOLA, Zehra DURDAĞ "K-MEANS ALGORİTMASI İLE OTOMATİK KÜMELEME" El-Cezeri Journal of Science and Engineering Yıl 2016, Cilt 3, Sayı 2 Sayfalar 0-0
10. Saadet Aytaç ARPACI, Oya KALIPSIZ "YAZILIM HATA SINIFLANDIRMASINDA FARKLI NAİVE BAYES TEKNİKLERİN KIYASLANMASI" Niğde Ömer

Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi Yıl 2018, Cilt 7, Sayı 1, Sayfalar 1 – 13

11. Emre GÜL, Mete KALYONCU "AĞIR VASITA HAVA KOMPRESÖRÜ PİSTON SEGMANI AŞINMASI DURUMLARINDA K-EN YAKIN KOMŞU ALGORİTMASININ SINIFLANDIRMA PERFORMANSININ İNCELENMESİ" Avrupa Bilim ve Teknoloji Dergisi Yıl 2020, Cilt, Sayı, Sayfalar 78 – 90
12. Banu AKKUŞ, Metin ZONTUL "VERI MADENCILIĞI YÖNTEMLERI ILE ÜLKELERI GELIŞMIŞLIK ÖLÇÜTLERINE GÖRE KÜMELEME ÜZERINE BIR UYGULAMA" AURUM Mühendislik Sistemleri ve Mimarlık Dergisi Yıl 2019, Cilt 3, Sayı 1, Sayfalar 51 – 64
13. Wu, D. (2009) "Supplier Selection: A Hybrid Model Using DEA, Decision Tree and Neural Network", Expert Systems with Applications, 36(1): 9105-9112.
14. Zhao Han, S. ve Bing Xiang, L. (2005) "Research Method of Customer Churn Crisis Based on Decision Tree", Journal of Management Sciences in China, 8(2): 20-25.
15. Akpınar, H. (2000) "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", İstanbul Üniversitesi İşletme Fakültesi Dergisi, 29(1): 1-22.
16. Wang, J., (2006). Encyclopedia of Data Warehousing and Mining. Information Science Reference, Volume: 49, ss: 140.
17. Mitchell, T., "Machine Learning", McGraw Hill, New York, (1997).
18. Han, J. and Kamber, M., "Data mining: concepts and techniques", Morgan Kaufmann Publishers, Burlington, (2006).
19. Kabalcı, E. (2014). Yapay Sinir Ağları. Ders Notları https://ekblc.files.wordpress.com/2013/09/ysa.pdf
20. Ağyar, Z. (2015). Yapay Sinir Ağlarının Kullanım Alanları ve Bir Uygulama. Mühendis ve Makine 56(662), 22-23.
21. Kaya, Ü., Oğuz, Y., & Şenol, Ü. (2018). An Assessment of Energy Production Capacity of Amasra Town Using Artificial Neural Networks. Turkish Journal of Electromechanics and Energy, 3(1), 22- 26.
22. Tektaş, M., Akbaş, A., Topuz, V. (2006). Yapay Zekâ Tekniklerinin Trafik Kontrolünde Kullanılması Üzerine Bir İnceleme. İstanbul: Marmara Üniversitesi.
23. Koyuncugil, A., & Özgülbaş, N. (2009). Veri madenciliği: Tıp ve sağlık hizmetlerinde kullanımı ve uygulamaları. Bilişim Teknolojileri Dergisi, 2(2).

24. Karahan, M. (2011). İstatistiksel tahmin yöntemleri: Yapay sinir ağları ile ürün talep tahmini uygulaması. Doctoral dissertation: Selçuk Üniversitesi Sosyal Bilimler Enstitüsü.
25. Partal, T., Kahya, E., & Cığızoğlu, K. (2011). Yağış verilerinin yapay sinir ağları ve dalgacık dönüşümü yöntemleri ile tahmini. İTÜ Dergisi/d, 7(3).
26. Engin, O., & Döyen, A. (2004). Artificial immune systems and applications in industrial problems. Gazi University Journal of Science, 17(1), 71-84.
27. Adepoju, G. A., Ogunjuyigbe, S. O. A., & Alawode, K. O. (2007). Application of neural network to load forecasting in Nigerian electrical power system. The Pacific Journal of Science and Technology, 8(1), 68-72
28. Kabalcı,Ersan. "Jeoloji Mühendisliği A.B.D. Esnek Hesaplama Yöntemleri- I"(Nisan 2019)