T.C.
NECMETTİN ERBAKAN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

NORMALİZASYON TEKNİKLERİNİN BİYOMEDİKAL VERİLERDE SINIFLAMA BAŞARISINA ETKİSİ

Hakan YÜCE

YÜKSEK LİSANS TEZİ

Elektrik Elektronik Mühendisliği Anabilim Dalı

Haziran-2021
KONYA

TEZ KABUL VE ONAYI

Hakan YÜCE tarafından hazırlanan "**Normalizasyon tekniklerinin biyomedikal verilerde sınıflama başarısına etkisi**" adlı tez çalışması 15/06/2021 tarihinde aşağıdaki jüri tarafından oy birliği / ~~oy çokluğu~~ ile Necmettin Erbakan Üniversitesi Fen Bilimleri Enstitüsü Elektrik Elektronik Mühendisliği Anabilim Dalı'nda YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Jüri Üyeleri	**İmza**
Başkan Doç. Dr. Bayram AKDEMİR	……..
Danışman Dr. Öğr. Üyesi Ali Osman ÖZKAN	……..
Üye Dr. Öğr. Üyesi Sabri ALTUNKAYA	……..

Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 28/05/2021 gün ve 2021/22-11 sayılı kararıyla onaylanmıştır.

Prof. Dr. İbrahim KALAYCI
FBE Müdürü

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Hakan YÜCE

Haziran 2021

ÖZET

YÜKSEK LİSANS TEZİ

NORMALİZASYON TEKNİKLERİNİN BİYOMEDİKAL VERİLERDE SINIFLAMA BAŞARISINA ETKİSİ

Hakan YÜCE

Necmettin Erbakan Üniversitesi Fen Bilimleri Enstitüsü
Elektrik Elektronik Mühendisliği Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Ali Osman ÖZKAN

2021, 92 Sayfa

Jüri
Dr. Öğr. Üyesi Ali Osman ÖZKAN
Doç. Dr. Bayram AKDEMİR
Dr. Öğr. Üyesi Sabri ALTUNKAYA

Son zamanlarda yapay zekâ uygulamaları askeri, ekonomi, tıp, v.b. gibi birçok alanda etkin olarak kullanılmaktadır. Özellikle sağlık sektöründe bilgisayarlarda saklanan hastalara ait verilerden hastaya ait teşhisi tahmin etme yapay zekâ uygulamalarından bir tanesidir. Fakat bilindiği gibi bu saklanan veriler çok büyük boyutlara sahip olup eşit derecede incelenmesi sonucu en doğru şekilde tahmin etmemize olanak sağlayacaktır. Bu verilerin daha etkin kullanılması için normalizasyon yöntemleri kullanılmaktadır. Bu çalışmada, diyabet hastalığı veri seti, göğüs kanseri hastalığı veri seti, karaciğer hastalığı veri seti ve kalp hastalığı veri setine minimum-maksimum (min-mak) normalizasyon yöntemi, ondalık ölçekleme normalizasyon yöntemi, z-skor normalizasyon yöntemi ve norm normalizasyon yöntemi uygulanmış ayrıca bu veri setleri normalize edilmeden de değerlendirilmiştir. Daha sonra normalize edilmiş ve ham verilere, 4 farklı k-kat çaprazlama (2,5,10,20) kriterinde yapay sinir ağları (YSA), karar ağacı (KA), destek vektör metodu (DVM), k en yakın komşu (k-NN) ve Naive Bayes gibi çeşitli sınıflandırma algoritmalarıyla ORANGE programı kullanılarak sınıflandırma işlemi yapılmış ve sınıflama doğrulukları değerlendirilmiştir. Sonuçlar istatiksel olarak incelenmiş ve normalizasyon yöntemlerinin yapay zekâ sınıflandırma yöntemlerinin performansını artırabileceği gözlenmiştir.

Anahtar Kelimeler: Yapay zekâ, min-mak normalizasyon yöntemi, ondalık ölçekleme normalizasyon yöntemi, z-skor normalizasyon yöntemi, norm normalizasyon yöntemi, YSA, DVM, KA, k-NN, Naive Bayes, ORANGE programı

ABSTRACT

MS THESIS

EFFECT OF NORMALIZATION TECHNIQUES ON CLASSIFICATION SUCCESS IN BIOMEDICAL DATA

Hakan YÜCE

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE OF NECMETTİN ERBAKAN UNIVERSITY
THE DEGREE OF MASTER OF SCIENCE
IN ELECTRICAL ELECTRONICS ENGINEERING

Advisor: Assist. Prof. Dr. Ali Osman ÖZKAN

2021, 92 Pages

Jury

Assist. Prof. Dr. Ali Osman ÖZKAN

Assoc. Prof. Dr. Bayram AKDEMİR

Assist. Prof. Dr. Sabri ALTUNKAYA

Recently, artificial intelligence applications have been used effectively in many areas such as military, economics, medicine… Especially, in the healthcare sector, it is one of the applications of artificial intelligence to predict a patient's diagnosis from data stored on computers. However, as is known, these stored data have very large dimensions and will allow us to estimate the outcome in the most accurate way if they are evaluated equally. For more efficient use of this data, normalization methods are used. In this study, the diabetes data set, breast cancer disease data set, liver disease data set and heart disease data set are normalized with minimum and maximum (min-max) normalization method, decimal scaling normalization method, z-score normalization method, norm normalization method and these data sets are also evaluated without normalizing. These normalized data sets and raw data sets were then classified using ORANGE program with various classification algorithms such as artificial neural networks (YSA), decision tree (KA), support vector method (DVM), k nearest neighbor (k-NN) and Naive Bayes in 4 different k-fold crossover criteria (2,5,10,20) and classificaition accuracies were evaluated. The results were analyzed statistically and it was observed that normalization methods can improve the performance of artificial intelligence classification methods.

Keywords: Artificial intelligence, min-max normalization method, decimal scaling normalization method, z-score normalization method, norm normalization method, ANN, SVM, DT, k-NN, Naïve Bayes, ORANGE program

ÖNSÖZ

Tez çalışması boyunca belirtikleri görüş ve önerilerle tezin yönlenmesine yardımcı olan danışmanım sayın Dr. Öğr. Üyesi Ali Osman ÖZKAN' a, tez süresince verdikleri destek ve anlayıştan dolayı bölüm başkanımız sayın Prof. Dr. Mehmet Akif ERİŞMİŞ ve tüm hayatım boyunca beni bu zamana kadar yetiştiren aileme teşekkürlerimi sunuyorum.

Hakan YÜCE
KONYA-2021

İÇİNDEKİLER

KISALTMALAR

ARFF	**:**	Attribute Relationship File Format (Öznitelik İlişkisi Dosya Biçimi)
MAD	**:**	Mean Absolute Deviation (Medyan Mutlak Deviasyon)
Min-Mak	**:**	Minimum – Maximum (Minimum-Maksimum)
YSA	**:**	Yapay Sinir Ağları
DVM	**:**	Destek Vektör Makinesi
CSV	**:**	Comma Seperated Values (Virgülle Ayrılmış Değerler)
ANFIS	**:**	Adaptive Network Fuzzy Inference Systems (Adaptif Ağ Tabanlı Bulanık Çıkarım Sistemi)
SPECT	**:**	Single Photon Emission Computed Tomograph (Tek Foton Emisyon Bilgisayar Tomografi)
k-NN	**:**	k -Nearest Neighbors (k En Yakın Komşu)
KA	**:**	Karar Ağacı
MSE	**:**	Mean Squared Error (Ortalama Karesel Hata)
MIAS	**:**	Mammographic Image Analysis Society (Mamografik Görüntü Analizi Derneği)
RA	**:**	Romatoid artrit
VEP	**:**	Visual Evoked Potentials (Göresel Uyarılmış Potansiyeller)
WEKA	**:**	Waikato Environment for Knowledge Analysis
UCI	**:**	University of California,Irvine
HOMA	**:**	Homeostasis Model Assessment (Homeostaz Modeli Değerlendirmesi)
MCP	**:**	Monocyte chemoattractant protein (Monosit Kemoatraktan Protein)
BMI	**:**	Body Mass Index (Vücut Kütle İndeksi)

EŞİTLİKLER

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

1. GİRİŞ

Günümüzde tıpta herhangi bir hastalığa ait veri miktarı giderek artmakta ve bu verilerden hastalık hakkında tahminler yapılmaktadır. Bu tahminlere destek olabilecek makine öğrenmesi tabanlı çalışmalar da giderek artmaktadır. Bu tahminleri yapacak olan sınıflandırma algoritmaları bu ham verilerle eğitilmiş ve eğitilen algoritmalar hastalar hakkında tahminler de bulunmuştur. Fakat son zamanlarda sınıflandırma algoritmaları ham veri ile eğitilmekten vazgeçilmiş olup eğitimden önce ham veriler z-skor normalizasyon yöntemi, min-mak normalizasyon yöntemi, ondalık ölçekleme yöntemi, norm yöntemi, medyan yöntemi vb. gibi normalizasyon yöntemlerinden biri kullanılarak verinin boyutu değiştirilmeden değerleri ölçeklenmiştir. Daha sonra normalize edilen veri seti YSA, KA, k-NN, DVM ve Naive Bayes gibi sınıflandırma yöntemleri ile sınıflandırma işlemine tabi tutulmuştur. Normalize işlemine tabi tutulan veri setinin sınıflandırma performansının daha iyi olduğu görülmüştür.

1.1 Literatür Taraması

Özellikle literatüre bakıldığında; farklı hastalıklar hakkında tahmin yapılmadan önce hastalığa ait veriler min -mak normalizasyon, ondalık ölçekleme normalizasyon, z-skor normalizasyon v.b. normalizasyon yöntemleri ile normalize edilerek veriler YSA, KA, k-NN, DVM ve Naive Bayes gibi sınıflandırma yöntemleri ile sınıflandırma işlemine tabi tutulmuştur. Yapılan yeni sınıflandırmalar ham veri ile yapılan sınıflandırmaya göre daha iyi sonuç verdiği görülmüştür. Bu alanda yapılan çalışmalardan bazıları aşağıda sıralanmıştır.

Özkan ve Durğun (2016) yaptığı çalışmada 40 sağlıklı kişi ve 40 romatoid artritli (RA) hastası olan kişilerin sağ ve sol el ulnar ve radyal arterlerinden Doppler sinyallerini almışlardır. Sonra bu sinyallere parametrik olmayan işaret işleme yöntemlerinden biri olan Welch periodogram yöntemini uygulayarak işaretlerin öz niteliğini çıkarmışlardır. Sinyallerin çıkarılmasından sonra en yaygın 3 normalizasyon yöntemi (z-skor, ondalık ölçekleme ve minimum-maksimum) ve YSA sınıflandırma yöntemini kullanarak WEKA programı vasıtasıyla sınıflandırma performansını incelemişlerdir. Sınıflama işlemi ham veriler kullanılarak ve mevcut normalizasyon yöntemi ile normalize edilmiş veriler olmak üzere her bir el için dört farklı veri seti ile gerçekleşmiştir. YSA ile sınıflama işlemi bittikten sonra performansı ölçmek amacıyla 10-kat çaprazlama yöntemi kullanılarak veri

kümeleri ayırma, sınıflama doğruluğu, seçicilik ve duyarlılık durumları incelenmiştir. Sonuçlar incelendiğinde hem sağ el hem de sol eldeki verilere dayalı sınıflandırma sonuçlarında normalizasyon yöntemlerinin bu hastalık üzerinde sınıflandırma performansını artırdığı gözlenmiştir ve en doğru sınıflandırma sonucuna z-skor normalizasyon yöntemi kullanılarak elde edilmiştir.

Singh ve arkadaşları (2015) yaptığı çalışmada bazı popüler normalleştirme tekniklerinin özelliklerini araştırmış ve değerlendirmişlerdir. Çalışmalarında meme kanserinin ultrasonik görüntülerini normalize ederek sınıflandırıcının performansı üzerindeki etkisini incelemişler. Veri setine, normalizasyon tekniklerini değerlendirmek için geri yayılım yapay sinir ağı ve destek vektör makinesini kullanarak sınıflandırma yapmışlar ve normalizasyon yöntemi olarak minimum-maksimum yöntemi, Z-skor yöntemi, Softmax yöntemi ve D-minimum-maksimum yöntemini kullanmışlardır. Sonuç olarak normalizasyon tekniklerinin sınıflandırma doğruluğuna önemli bir etkiye sahip olduğunu göstermişlerdir.

Jayalakshmi ve Santhakumaran (2011) yaptığı çalışmada diyabet hastalarını sınıflandırmada normalizasyon yöntemlerinin etkisini izlemişlerdir. Bu çalışma için Pima Hintlilerinin diyabet hastalığı veri kümesi kullanılmıştır. Sonuç olarak da deneysel olarak da diyabet hastalığını yapay sinir ağları sınıflandırma algoritması kullanarak sınıflandırma işleminde performansın normalizasyon yöntemlerine bağlı olduğunu görülmüştür. Bu çalışmada, geri yayılım yapay sinir ağı modelinde en iyi normalizasyon yöntemi olarak istatiksel sütun normalizasyonu önerilmiştir.

Atomi (2012) yaptığı çalışmada yapay sinir ağlarının son zamanlarda tıp, biyoloji, finans, ekonomi ve benzeri birçok uygulamada kullanıldığından bahsetmiş. Burada YSA'nın yakınsamasını artırmak için farklı ön işleme tekniklerini kullanmıştır. Özellikle min-mak normalizasyon yöntemi, z-skor normalizasyon yöntemi ve ondalık ölçekleme normalizasyon yöntemi kullanmış ve farklı ön işleme tekniklerinin YSA'nın hesaplama verimliliğini oldukça artırdığını göstermiştir.

Huang ve Qin (2018) yaptığı çalışmada moleküler sınıflandırmanın performansını artırmak için normalizasyon yöntemlerinin kullanılabileceğini vurgulamışlar ve bu çalışmada medyan normalizasyon yöntemi, nicelik normalizasyon yöntemi ve varyans stabilize normalizasyon yöntemlerini kullanmışlardır. Deneylerinde bir çift mikroRNA mikrodizi veri kümesinden yeniden örneklemeye dayalı simülasyonlar kullanmışlardır. Veriler deneysel olarak elde edildiğinde işlemeden kaynaklı kusurlara sahip olabileceğini vurgulamışlar ve bu kusurların çeşitli problemlere neden olabileceğinin vurgulamışlardır.

Sonuç olarak normalizasyon yöntemlerinin bu kusurların etkisini azaltarak sınıflandırıcı performansını artırabileceği görülmüştür.

Ahidha ve Premalatha (2017) yaptığı çalışmada mikrodizi verilerinin özellik seçimi ve sınıflandırılması, makine öğreniminde en önemli zorluklardan biri olduğundan bahsetmiş ve özellik seçimi tekniklerinin arkasındaki etkenin, kanser/tümör mikrodizi ekspresyon verilerinin sınıflandırılmasında hayati bir rol oynayan ayrımcılık özelliği alt kümelerinin seçilmesinden bahsetmişlerdir. Bu çalışmada, bulanık Gauss üyelik fonksiyonu ile normalize edilen miRNA verilerinde F-skoru ve ilgili bilgi kazancını birleştiren yeni bir özellik seçim yaklaşımını kullanarak DVM ve YSA sınıflandırma yöntemleri ile sınıflandırma işlemine tabi tutmuşlardır. Deneysel sonuçlar, önerilen yaklaşımın son teknoloji özellik seçme algoritmalarına kıyasla daha iyi bir sınıflandırma doğruluğu sağladığını göstermişlerdir.

İleri ve arkadaşları (2018) konuşmacının cinsiyetini tanımlamak için yaptıkları çalışmada normalizasyon yöntemlerinin etkisini incelemek istemişler ve çalışmada; kısa süreli ortalamanın etkisi ve varyans normalizasyonu, kısa süreli spektral ortalama ve ölçekleme normalizasyonu, min-mak normalizasyonu, z-skor normalizasyonu ve standart sapma normalizasyonu yöntemlerini kullanmışlardır. Bu yöntemlerden herhangi birini kullanmadan sınıflandırıcı olarak destek vektör yöntemini kullandıklarında 384 konuşmacıdan 375 konuşmacının cinsiyetini doğru tahmin etmişlerdir. Başarı % 97.6562 olarak elde edilmiştir. Fakat standart sapma normalizasyon yöntemi hariç diğer normalizasyon yöntemlerinde bu başarıya yaklaşamamışlardır. Standart sapma normalizasyon yönteminde ise; 384 konuşmacıdan 377 konuşmacının cinsiyeti doğru tahmin edilmiştir. Başarı % 98.1771 olarak elde edilmiştir. Sonuç olarak normalizasyon yönteminin sınıflandırıcı performansını artırabileceği görülmüştür.

Borkin ve arkadaşları (2019) sınıflandırma model performansında veri normalizasyonun etkisini araştırmak istemişler ve bunun için Parkinson hastası olan bireylerin veri setlerini kullanmışlardır. Borkin ve arkadaşları sınıflandırma işlemini XGBoost sınıflandırma modeli ile normalizasyon işlemini ise, min-mak normalizasyon yöntemi ile yapmışlardır. Sonuçları verileri normalize etmeden ve normalize ederek karşılaştırmışlar ve verileri normalize etmeden daha doğru bir sınıflandırma yaptığını gözlemlemişlerdir. Fakat her veri normalizasyonundan sonra yapılan sınıflandırma işleminin daha az doğruluk vermeyeceğini vurgulamışlardır. Çünkü buradaki verilerin lineer dönüşüme hassas olmayabileceğini vurgulamışlardır.

Akdemir (2009) yaptığı çalışmada her zaman yapılan sütun temelli normalizasyon yöntemi yerine yeni bir normalizasyon yöntemi olarak satır temelli normalizasyon yöntemini kullanmıştır. Satır tabanlı normalizasyon yaparken önce her satırda özelliklerin birimleri farklı olabileceği için öncelikle bu birimler ortadan kaldırılmış sonra normalizasyon yapılmıştır. Çalışmada sınıflama performansı ölçümü için SPECT verisi, kalp verisi, Doppler, hepatit ve VEP verilerini kullanmıştır. Bu yöntemi bu ham verilere uygulayarak onları normalize etmiş ve ardından bu normalize edilen verileri en yaygın kullanılan sınıflandırma yöntemlerinden olan ANFIS ve YSA 'da kullanmıştır. Aynı ham verileri geleneksel normalizasyon yöntemleri normalize edip aynı sınıflandırma yöntemleri ile sınıflandırmıştır. Bu yöntemleri karşılaştırdığında önerilen yeni metodun sınıflandırma performansına olumlu etki ettiğini gözlemlemiştir.

Mustaffa ve Yusof (2010) yaptıkları çalışmada gelecek dang salgını hakkında tahminde bulunmak istemişler. Tahmin işleminde DVM ve YSA sınıflandırma yöntemlerini kullanmadan önce min-mak normalizasyon yöntemi, ondalık ölçekleme normalizasyon yöntemi ve z-skor normalizasyon yöntemini uygulamışlar ve sonuçları tahmin doğruluğu ve MSE olarak değerlendirmişlerdir. En iyi sonuca DVM yöntemi ile ulaşmışlar ve ondalık ölçekleme normalizasyon yönteminin sınıflandırma performansı artırdığını gözlemlemişlerdir.

1.2 Çalışmanın Amacı ve Önemi

Günümüzde hastalara ait çok büyük veriler bilgisayar ortamında saklanmakta ve bu devasa veriler bilgisayarlar tarafından yorumlanarak hastalar hakkında tahminler yapmaktadır. Hastalara ait teşhis makine öğrenmesinin bir alt dalı sınıflandırma algoritmaları vasıtasıyla gerçekleştirilmektedir. Bu işlem sırasında bu sınıflandırma algoritmaları hastalığa ait ham verilerle eğitildikten sonra muhtemel hastalık hakkında tahmin yapılması istenmektedir. Fakat bazen bu veriler birbirine göre çok uç değer alabilmekte, bazen varyanslardan çabuk etkilenebilmekte, bazen faklı nedenlerden dolayı sınıflandırma algoritmaları iyi eğitilememesine neden olarak sınıflandırmada negatif bir etkiye sahip olabilmektedir. Bu olumsuz etkiyi ortadan kaldırmak için çeşitli çözümler ortaya atılmış ve bu çözümlerden bir tanesi de sınıflandırma algoritmaları eğitilmeden önce ham veriyi normalize etme fikridir.

Yapılan bu çalışmada Pima Hintlilerinin diyabet hastalığı verisi, göğüs kanseri hastalığı verisi, karaciğer hastalığı verisi ve kalp hastalığı verisi DVM, YSA, KA, k-NN ve

Naive Bayes sınıflandırma yöntemlerinde 4 farklı k-kat çaprazlama (2,5,10,20) kriteri uygulanmadan önce min-mak normalizasyon yöntemi, ondalık ölçekleme normalizasyon yöntemi, z-skor normalizasyon yöntemi ve norm normalizasyon yöntemi uygulanmıştır. Normalizasyon yöntemi sonrasında ORANGE programı kullanılarak sınıflandırma işlemi gerçekleştirilmiş ve sonuçlar ham veri setlerinin sınıflandırma performansı ile karşılaştırılmıştır. Sonuç olarak normalizasyon yöntemleri uygulanarak sınıflandırıcı performansının artabileceği görülmüştür.

2. NORMALİZASYON YÖNTEMLERİ

Normalizasyon axb veri boyutuna sahip bir veri setini bir uzaydan başka bir uzaya taşır. Bu taşımada yeni maksimum ve minimum noktaları oluşur ancak veri setinin axb olan boyutunda herhangi bir değişiklik olmaz. Burada ham verinin aksine normalize edilmiş veri sayesinde sınıflandırıcının kararlılığı artabilecektir. Fakat şunu bilmeliyiz ki her veri seti için normalizasyon gerekmez. Özellikler farklı aralıklara sahip olduğu zaman gerekir (Akdemir,2009).

Örneğin bu farklılık birimsel farklılıktan kaynaklanabilir ya da diğer bir farklılıktan kaynaklanabilir. İki özellik içeren veri setini göz önüne alalım. Bu özelliklerden bir tanesi yaş olsun ve 20-40 aralığında değişsin. Diğer özellik ise bu kişilerin aldıkları maaş olsun ve 2000-20000 TL arasında değişsin. Görüldüğü gibi yaş ile maaş verisi arasındaki oran 100 kattır. Bu iki özelliğin aralığı farklıdır. Biz bir analiz yaptığımız zaman; örneğin regresyon analizi, maaş özelliği sonucu daha fazla etkiler. Fakat bu özelliğin daha önemli olduğunu bize söylemez. Etkilerin eş miktarda olmasını sağlamak için bu iki özelliği normalize etmeliyiz. Ek olarak 10000 satırlık verimiz var. Bu verilerden örneğin bazı satırlarda maaş değerlerini 500000 TL ya da daha büyük veya çok az değerler girdik. Bu değerlerin normalize edilmesi ile etkilerini ortadan kaldırabiliriz. Ayrıca veri setinin özellik çıkarımından sonra oluşturulan yeni veri setinin boyutu fazla olabilir. Veri setinde ilgisiz/fazla özellikler olabilir. Bu özellikler sınıflama performansını azaltabilir ve sınıflandırıcının hesaplama maliyetini artırabilir (Polat, 2008; Akdemir B.,2009)). Ayrıca unutmamalıyız ki çok katmanlı ağ modelinin girdi ve çıktıların ölçeklenmesi ağın performansını yakından etkilemektedir. Böylece değerlerin dağılımı daha düzenli olacaktır. Görülmelidir ki normalizasyon sadece girdi değere uygulanmayıp aynı zamanda çıktı değere de uygulanabilir. Çünkü bu çıktılar başka bir YSA için veri seti olabilir (Yavuz S., Deveci M.,2012).

Bu çalışmada min-mak normalizasyon yöntemi, ondalık ölçekleme normalizasyon yöntemi, z-skor normalizasyon yöntemi, medyan normalizasyon yöntemi, D_minimum-maksimum normalizasyon yöntemi, norm normalizasyon yöntemi, medyan-mod normalizasyon yöntemi ve ortalama-mod normalizasyon yöntemi açıklanmış ve min-mak normalizasyon yöntemi, ondalık ölçekleme normalizasyon yöntemi, z-skor normalizasyon yöntemi ve norm normalizasyon yönteminin DVM, YSA, KA, k-NN ve Naive Bayes gibi sınıflandırma algoritmalarının sınıflandırma doğruluk performanslarına etkisi incelenmiştir. Şimdi ilk olarak bu normalizasyon yöntemlerini inceleyelim.

2.1 Minimum Maksimum Normalizasyon Yöntemi

Mühendislik uygulamalarında en fazla tercih edilen normalizasyon yöntemlerinden biridir. Verileri doğrusal olarak normalize eder. Bu yöntem mühendislik çalışmalarında tıbbi veriler, görüntü verileri gibi kaynağı mühendislik menşeli olmayan veri setlerinde de yaygın olarak kullanılmaktadır. Bu normalizasyon yönteminde veri negatif değerli olsa bile negatif işaret ortadan kalkar. Minimum, verinin alabileceği en düşük değerdir. Bu değer 0'dır. Maksimum ise, en büyük değerdir. Bu değer ise 1'dir. Diğer değerler 0 ile 1 arasında değerler alacaktır. Kısaca bu normalizasyon yönteminde değerler 0-1 arasına sıkıştırılır. Min-mak normalizasyonunun dezavantajı keskin değerleri çok iyi ele alamaz. Yani bir veri setinde 100 değerimiz olsun. Bu değerlerden 99 tanesi 60' tan küçük bir tanesi 99 ise, 99 olan değer eskisi kadar aktif değildir (www.codecademy.com/articles/normalization ; Akdemir B.,2009; Yavuz S., Deveci M.,2012).

Bir veriyi 0-1 arasına sıkıştırmak için Eşitlik 2.1 kullanılır.

$$x' = \frac{(x_i - x_{min})}{(x_{max} - x_{min})} \tag{2.1}$$

Eşitlik 2.1'de:

x' = Normalize edilmiş değeri

x_i = Normalize edilecek değeri

x_{min} = Veri setindeki en küçük değeri

x_{max} = Veri setindeki en büyük değeri ifade etmektedir.

Çizelge 2.1'de min-mak normalizasyon yöntemini daha iyi anlamak için kullanılacak bir veri seti görülmektedir. Bu veri seti 4 farklı kişiye ait 4 farklı özelliğe sahiptir.

Çizelge 2.1 Min-mak yöntemi için normalize edilmemiş veri seti

	Özellik-1	**Özellik-2**	**Özellik-3**	**Özellik-4**
Kişi-1	5	10	22	4
Kişi-2	2	51	15	3
Kişi-3	8	20	14	15
Kişi-4	11	2	1	24

Çizelge 2.1'de görülen 4 kişiye ait veriler min-mak normalizasyon yöntemi ile 0-1 aralığına sırasıyla sütun sütun ölçeklenecektir. Buna göre:

Özellik-1 sütunu için veri setinin minimum ve maksimum değeri bulunmalıdır. Buna göre,

Özellik-1 sütununda en büyük değer = 11

Özellik-1 sütununda en küçük değer = 2

Sonrasında Eşitlik 2.1 kullanılarak Özellik-1 sütununun ölçeklenmiş değerleri elde edilir.

Özellik_{11} = (5-2) / (11-2) = 0.33

Özellik_{12} = (2-2) / (11-2) = 0

Özellik_{13} = (8-2) / (11-2) = 0.66

Özellik_{14} = (11-2) / (11-2) = 1

Özellik-2 sütunu için veri setinin minimum ve maksimum değeri bulunmalıdır. Buna göre,

Özellik-2 sütununda en büyük değer = 51

Özellik-2 sütununda en küçük değer = 2

Sonrasında Eşitlik 2.1 kullanılarak Özellik-2 sütununun ölçeklenmiş değerleri elde edilir.

Özellik_{21} = (10-2) / (51-2) = 0.16

Özellik_{22} = (51-2) / (51-2) = 1

Özellik_{23} = (20-2) / (51-2) = 0.367

Özellik_{24} = (2-2) / (51-2) = 0

Özellik-3 sütunu için veri setinin minimum ve maksimum değeri bulunmalıdır. Buna göre,

Özellik-3 sütununda en büyük değer = 22

Özellik-3 sütununda en küçük değer = 1

Sonrasında Eşitlik 2.1 kullanılarak Özellik-3 sütununun ölçeklenmiş değerleri elde edilir.

Özellik_{31} = (22-1) / (22-1) = 1

Özellik_{32} = (15-1) / (22-1) = 0.76

Özellik_{33} = (14-1) / (22-1) = 0.71

Özellik_{34} = (1-1) / (22-1) = 0

Özellik-4 sütunu için veri setinin minimum ve maksimum değeri bulunmalıdır. Buna göre,

Özellik-4 sütununda en büyük değer = 24

Özellik-4 sütununda en küçük değer = 3

Sonrasında Eşitlik 2.1 kullanılarak Özellik-4 sütununun ölçeklenmiş değerleri elde edilir.

Özellik_{41} = (4-3) / (24-3) = 0.04

Özellik_{42} = (3-3) / (24-3) = 0

Özellik_{43} = (15-3) / (24-3) = 0.57

Özellik_{44} = (24-3) / (24-3) = 1

Yukarıdaki yapılan işlemlerin sonrasında Çizelge 2.1'de verilen 4 kişiye ait olan 4 farklı özelliğin min-mak normalizasyon yöntemi uygulanarak 0-1 aralığına ölçeklenmiş değerleri bulunmuştur. Bu ölçeklenmiş değerler Çizelge 2.2'de gösterilmiştir.

Çizelge 2.2 Min-mak yöntemi için normalize edilmiş veri seti

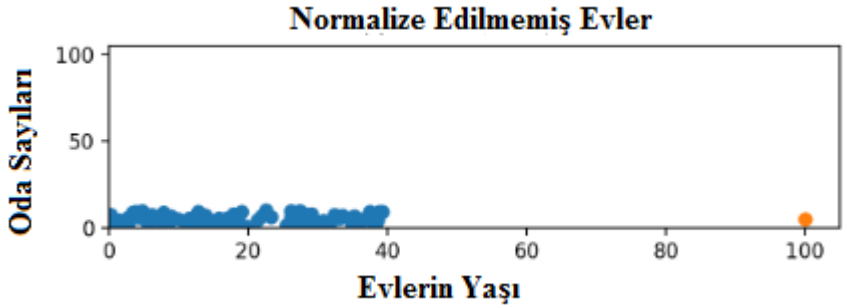
	Özellik-1	**Özellik-2**	**Özellik-3**	**Özellik-4**
Kişi-1	0.33	0.16	1	0.04
Kişi-2	0	1	0.76	0
Kişi-3	0.66	0.367	0.71	0.57
Kişi-4	1	0	0	1

Çizelge 2.2'de görüldüğü gibi her sütunda maksimum olarak 1 değeri, minimum olarak 0 değeri mevcuttur. Yani *x* değerimiz minimum durumunda y=0, x değerimiz

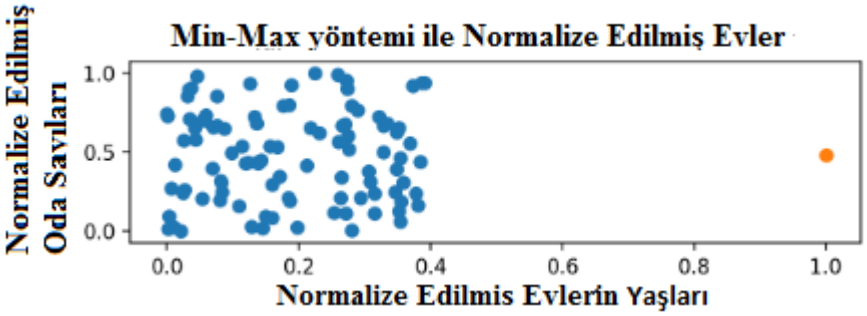
maksimum durumunda iken y=1 değerini alır. Bu y değerleri bizim belirlediğimiz ölçek aralığına bağlı olacaktır (Yavuz S., Deveci M.,2012).

Bu normalizasyon yönteminin dezavantajlardan biride, yeni bir veri eklendiği zaman yeni veri ilgili sütunda maksimum ya da minimum olabilir. Bu durum normalizasyonun tekrar yapılmasına neden olacaktır (https://ec.europa.eu/jrc/en/coin/10-step-guide/step-5).

Bu normalizasyon yöntemini Şekil 2.1 ve Şekil 2.2'de anlatalım. Şekil 2.1'de evlere ait oda sayıları ve evlerin yaşları bilgilerini gösteren bir grafiksel dağılım görülmektedir. Oda sayıları yaklaşık Şekil 2.1'de görüldüğü gibi 0-10 arası değişirken, evlerin yaşı 0-40 arası değişiyor. Fakat bir evin yaşı istisna olarak (uç olarak) 100'dür. Bu evlerin yaş ve oda sayıları normalize edildiği zaman oda sayılarını ve evlerin yaşlarını Şekil 2.2'de görüldüğü gibi 0-1 arasına normalize edebiliyoruz. Fakat evlerin yaşı ağırlıklı olarak sadece 0-0.4 arasına normalize edilebiliyor. Grafikte görülen uç nokta 100 değerinden sebeple, buradaki herhangi bir evin fiyat tahmini yapılacağı zaman y değerlerinin x değerlerine göre baskın olacağı görülmektedir (www.codecademy.com/articles/normalization).



Şekil 2.1 Evlerin yaşlarını ve oda dağılımını gösteren normalize edilmemiş veri dağılımı grafiği (www.codecademy.com/articles/normalization)



Şekil 2.2 Evlerin yaşlarını ve oda dağılımını gösteren normalize edilmiş veri dağılımı grafiği (www.codecademy.com/articles/normalization)

Sonuç olarak min-mak yöntemi uç noktalara iyi odaklanamamıştır (www.codecademy.com/articles/normalization).

2.2 Ondalık Ölçekleme Normalizasyon Yöntemi

Ondalık ölçekleme yöntemi minimum-maksimum yöntemi kadar yaygın kullanılmamasına rağmen literatürde yer almaktadır. Bu normalizasyon yönteminde amaç veri seti değerlerini 1'den küçük yapmak için mevcut değerleri 10 ve 10' un katı değerlere bölünmesidir. Bu 10' un kuvveti değeri mevcut değeri 1'den küçük yapan en küçük değer olmalıdır (Akdemir B.,2009).

Bu normalizasyon yöntemi için Eşitlik 2.2 kullanılır.

$$A' = \frac{A_i}{10^j} \tag{2.2}$$

Eşitlik 2.2'de:

A' = Normalize edilmiş veriyi

A_i = Normalize edilecek değeri

j = A' değerini 1 den küçük yapan değeri ifade etmektedir.

Çizelge 2.3'de ondalık ölçekleme normalizasyon yöntemini daha iyi anlamak için kullanılacak bir veri seti görülmektedir. Bu veri seti 4 farklı kişiye ait 4 farklı özelliğe sahiptir.

Çizelge 2.3 Ondalık ölçekleme yöntemi için normalize edilmemiş veri seti

	Özellik-1	**Özellik-2**	**Özellik-3**	**Özellik-4**
Kişi-1	5	10	22	4
Kişi-2	2	51	15	3
Kişi-3	8	200	14	15
Kişi-4	3	2	1	24

Çizelge 2.3'te görülen 4 kişiye ait 4 özellik ondalık ölçekleme normalizasyon yöntemi kullanılarak Özellik-1 sütunundan Özellik-4 sütununa sırasıyla normalize edilecektir. Buna göre:

Özellik-1 sütununu 1’den küçük yapacak en küçük j değeri 1’dir ve Özellik-1 sütununun yeni değerlerini bulmak için her bir özellik 10 değerine bölünmelidir.
Buna göre Özellik-1 sütununun ölçeklenmiş değerleri:

$\text{Özellik}_{11} = 5 / 10 = 0.5$

$\text{Özellik}_{12} = 2 / 10 = 0.2$

$\text{Özellik}_{13} = 8 / 10 = 0.8$

$\text{Özellik}_{14} = 3 / 10 = 0.3$

Özellik-2 sütununu 1’den küçük yapacak en küçük j değeri 3’tür ve Özellik-2 sütununun yeni değerlerini bulmak için her bir özellik 1000 değerine bölünmelidir.
Buna göre Özellik-2 sütununun ölçeklenmiş değerleri:

$\text{Özellik}_{21} = 10 / 1000 = 0.01$

$\text{Özellik}_{22} = 51 / 1000 = 0.051$

$\text{Özellik}_{23} = 200 / 1000 = 0.2$

$\text{Özellik}_{24} = 2 / 1000 = 0.002$

Özellik-3 sütununu 1’den küçük yapacak en küçük j değeri 2’dir ve Özellik-3 sütununun yeni değerlerini bulmak için her bir özellik 100 değerine bölünmelidir. Buna göre Özellik-3 sütununun ölçeklenmiş değerleri:

$\text{Özellik}_{31} = 22 / 100 = 0.22$

$\text{Özellik}_{32} = 15 / 100 = 0.15$

$\text{Özellik}_{33} = 14 / 100 = 0.14$

$\text{Özellik}_{34} = 1 / 100 = 0.01$

Özellik-4 sütununu 1’den küçük yapacak en küçük j değeri 2’dir ve Özellik-2 sütununun yeni değerlerini bulmak için her bir özellik 100 değerine bölünmelidir. Buna göre Özellik-4 sütununun ölçeklenmiş değerleri:

$$\text{Özellik}_{41} = 4 / 100 = 0.04$$

$$\text{Özellik}_{42} = 3 / 100 = 0.03$$

$$\text{Özellik}_{43} = 15 / 100 = 0.15$$

$$\text{Özellik}_{44} = 24 / 100 = 0.24$$

Yukarıdaki yapılan işlemlerin sonrasında Çizelge 2.3'de verilen 4 kişiye ait olan 4 farklı özelliğin ondalık ölçekleme normalizasyon yöntemi uygulanarak ölçeklenmiş değerleri bulunmuştur. Bu ölçeklenmiş değerler Çizelge 2.4'de gösterilmiştir.

Çizelge 2.4 Ondalık ölçekleme yöntemi için normalize edilmiş veri seti

	Özellik-1	**Özellik-2**	**Özellik-3**	**Özellik-4**
Kişi-1	0.5	0.01	0.22	0.04
Kişi-2	0.2	0.051	0.15	0.03
Kişi-3	0.8	0.2	0.14	0.15
Kişi-4	0.3	0.002	0.01	0.24

Çizelge 2.4'te 4 kişiye ait olan 4 farklı özelliğin ondalık ölçekleme normalizasyon yöntemi uygulanarak ölçeklenmiş değerleri gösterilmektedir. Çizelgede görüldüğü gibi ondalık ölçekleme normalizasyon yöntemi ile veriler 0-1 arasına sıkıştırılmıştır.

Ondalık ölçekleme yönteminin dezavantajlarından bir tanesi minimum-maksimum yönteminde olduğu gibidir. Örneğin Özellik-1 sütununa yeni bir veri ekleyelim. Bu veri 10' dan büyük olduğunda Özellik-1 de bulunan veriyi 1'den küçük yapmak için veriler 100 değerine bölünme durumunda kalacaktır. Böylece Özellik-1'de bulunan diğer verilerde 100'e bölünme durumunda kalacak ve önceki değerleri değişecektir. Kısaca herhangi bir yeni veri ekleme durumunda tüm sütun verileri tekrar normalize edilmesi gerekecektir. Özellikle bu durum finansal verilerde önemli olabilir. Çünkü finansal verilerde sürekli değişkenlik gösterebilmektedir.
(www.codecademy.com/articles/normalization, (https://ec.europa.eu/jrc/en/coin/10-step-guide/step-5).

2.3 Z-Skor Normalizasyon Yöntemi

Bu yöntem istatiksel normalizasyon yöntemi olarak da bilinmektedir. Bilindiği gibi veri seti içinde bazı uç değerler vardır. Bu değerler sonuçlara daha fazla etki yapacaktır. Bu

veri seti içindeki uç verilerin diğer veriler gibi modele tahmin için eş katkı sağlaması gerekir. Z-skor yöntemiyle mevcut verilerin standart sapması ve ortalaması hesaplanarak normalizasyon işlemi gerçekleştirilir. Böylece veri seti içindeki uç değerlerin etkisi azaltılabilir.
(www.codecademy.com/articles/normalization; Yavuz,2013).

Bu normalizasyon yöntemi için Eşitlik 2.3 kullanılır.

$$x' = \frac{x_i - \mu_i}{\sigma_i} \tag{2.3}$$

Eşitlik 2.3'te:

x_i = Normalize edilecek değeri

μ_i = Veri setinin ortalama değeri

σ = Verideki standart sapmayı ifade etmektedir.

Eşitlik 2.3'te yer alan standart sapmanın hesaplaması için Eşitlik 2.4 kullanılır.

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i-1}^{N}(x_i - x')^2} \tag{2.4}$$

Eşitlik 2.4'te:

N = Dizinin eleman sayısını

x_i = Dizinin i. elemanını

x' = Dizinin elemanlarının aritmetik ortalamasını

İfade etmektedir.

Z-skor normalizasyon yöntemini daha iyi anlamak için Çizelge 2.5'de 4 farklı kişiye ait 4 farklı özelliği değerlendirelim.

Çizelge 2.5 Z-skor yöntemi için normalize edilmemiş veri seti

	Özellik-1	**Özellik-2**	**Özellik-3**	**Özellik-4**
Kişi-1	2	7	22	0
Kişi-2	25	0	17	4
Kişi-3	0	2	1	8
Kişi-4	12	3	8	12

Çizelge 2.5'de görülen 4 kişiye ait 4 özellik z-skor normalizasyon yöntemi ile Özellik-1 sütunundan Özellik-4 sütununa sırasıyla normalize edilecektir. Buna göre:

Özellik-1 sütunundaki verilerin ortalaması Eşitlik 2.3'te görüldüğü gibi

$\mu_i = (2+25+0+12)/4$

$= 9.75$ olarak hesaplanmıştır.

Özellik-1 sütunundaki verilerin standart sapması Eşitlik 2.4'te görüldüğü gibi

$$\sigma = \sqrt{\frac{(A_{11} - \mu_i)^2 + (A_{12} - \mu_i)^2 + (A_{13} - \mu_i)^2 + (A_{14} - \mu_i)^2}{3}}$$

$= 11.44188$ olarak hesaplanmıştır.

Eşitlik 2.3'e göre hesaplanan Özellik-1 ortalama değeri ve Eşitlik 2.4'e göre hesaplanan Özellik-1 standart sapma değerine bağlı 4 kişiye ait Özellik-1 sütununun yeni değerleri:

$\text{Özellik}_{11} = -0.67734$

$\text{Özellik}_{12} = 1.322823$

$\text{Özellik}_{13} = -0.85213$

$\text{Özellik}_{14} = 0.196646$

Özellik-2 sütunundaki verilerin ortalaması Eşitlik 2.3'te görüldüğü gibi

$\mu_i = (7+0+2+3)/4 = 3$ olarak hesaplanmıştır.

Özellik-2 sütunundaki verilerin standart sapması Eşitlik 2.4'te görüldüğü gibi

$$\sigma = \sqrt{\frac{(B_{11} - \mu_i)^2 + (B_{12} - \mu_i)^2 + (B_{13} - \mu_i)^2 + (B_{14} - \mu_i)^2}{3}}$$

$= 2.94392$ olarak hesaplanmıştır.

Eşitlik 2.3'e göre hesaplanan Özellik-2 ortalama değeri ve Eşitlik 2.4'e göre hesaplanan Özellik-2 standart sapma değerine bağlı 4 kişiye ait Özellik-2 sütununun yeni değerleri:

Özellik_{21} = 1.358732

Özellik_{22} = -1.01905

Özellik_{23} = -0.33968

Özellik_{24} = 0

Özellik-3 sütunundaki verilerin ortalaması Eşitlik 2.3'te görüldüğü gibi

$\mu_i = (22+17+1+8) / 4$

= 12 olarak hesaplanmıştır.

Özellik-3 sütunundaki verilerin standart sapması Eşitlik 2.4'te görüldüğü gibi

$$\sigma = \sqrt{\frac{(C_{11} - \mu_i)^2 + (C_{12} - \mu_i)^2 + (C_{13} - \mu_i)^2 + (C_{14} - \mu_i)^2}{3}}$$

= 9.345231 olarak hesaplanmıştır.

Eşitlik 2.3'e göre hesaplanan Özellik-3 ortalama değeri ve Eşitlik 2.4'e göre hesaplanan Özellik-3 standart sapma değerine bağlı 4 kişiye ait Özellik-3 sütununun yeni değerleri:

Özellik_{31} = 1.070065

Özellik_{32} = 0.535032

Özellik_{33} = -1.17707

Özellik_{34} = -0.42803

Özellik-4 sütunundaki verilerin ortalaması Eşitlik 2.3'te görüldüğü gibi

$\mu_i = (0+4+8+12) / 4$

=6 olarak hesaplanmıştır.

Özellik-4 sütunundaki verilerin standart sapması Eşitlik 2.4'te görüldüğü gibi

$$\sigma = \sqrt{\frac{(D_{11} - \mu_i)^2 + (D_{12} - \mu_i)^2 + (D_{13} - \mu_i)^2 + (D_{14} - \mu_i)^2}{3}}$$

= 5.163978 olarak hesaplanmıştır.

Eşitlik 2.3'e göre hesaplanan Özellik-4 ortalama değeri ve Eşitlik 2.4'e göre hesaplanan Özellik-4 standart sapma değerine bağlı 4 kişiye ait Özellik-4 sütununun yeni değerleri:

Özellik_{41} = -1.1619

Özellik_{42} = -0.3873

Özellik_{43} = 0.387298

Özellik_{44} = 1.161895

Yukarıdaki yapılan işlemlerin sonrasında Çizelge 2.5'de verilen 4 kişiye ait olan 4 farklı özelliğin z-skor normalizasyon yöntemi uygulanarak ölçeklenmiş değerleri bulunmuştur. Bu ölçeklenmiş değerler Çizelge 2.6'da gösterilmiştir.

Çizelge 2.6 Z-skor yöntemi için normalize edilmiş veri seti

	Özellik-1	**Özellik-2**	**Özellik-3**	**Özellik-4**
Kişi-1	-0.67734	1.358732	1.070065	-1.1619
Kişi-2	1.322823	-1.01905	0.535032	-0.3873
Kişi-3	-0.85213	-0.33968	-1.17707	0.387298
Kişi-4	0.196646	0	-0.42803	1.161895

Z-skor normalizasyon yöntemi veri normal bir dağılım (Gauss dağılımı) takip ederse faydalı olacaktır (Peshawa J. M. A,2015).

2.4 Medyan Normalizasyon Yöntemi

Bu normalizasyon yönteminde her veri setinin medyanı bulunur. Şayet orta noktada iki değer varsa bu iki sayının ortalaması alınır. Tek bir değer varsa o değer medyan değeridir. Medyan aşırı sapmalardan etkilenmez. Yani mevcut veri setinde diğer verilere göre aşırı yüksek ya da düşük verinin olması bu normalizasyon yönteminde daha az etkilidir. Bu normalizasyon yöntemi verileri ölçeklendirir ve her veri aynı medyana sahip olur. (tr.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/mean-

median-basics/a/mean-median-and-mode-review;Yavuz S. Deveci M.,2012; Välikangas T, Suomi T, and Elo L.L.,2016).

Bu normalizasyon yöntemi için Eşitlik 2.5 kullanılır.

$$x' = \frac{x_i}{\text{Medyan}(a_i)} \tag{2.5}$$

Eşitlik 2.5'te:

x' = Normalize edilmiş değeri

x_i = Normalize edilecek değeri

$\text{Medyan}(a_i)$ = Girdi setinin medyanını ifade etmektedir.

Çizelge 2.7'de medyan normalizasyon yöntemini daha iyi anlamak için kullanılacak bir veri seti görülmektedir. Bu veri seti 4 farklı kişiye ait 4 farklı özelliğe sahiptir.

Çizelge 2.7 Medyan yöntemi için normalize edilmemiş veri seti

	Özellik-1	**Özellik-2**	**Özellik-3**	**Özellik-4**
Kişi-1	2	7	22	1
Kişi-2	8	17	14	4
Kişi-3	10	2	1	8
Kişi-4	12	3	8	12

Çizelge 2.7'de görülen 4 kişiye ait 4 özellik medyan normalizasyon yöntemi ile Özellik-1 sütunundan Özellik-4 sütununa sırasıyla normalize edilecektir. Buna göre:

Çizelge 2.7'ye göre Özellik-1 sütunun küçükten büyüğe sıralandığı zaman orta olan veriler 8 ve 10 değeri olacaktır. Burada tek değer olmadığı için medyan değerimiz bu iki değerin ortalaması olan 9 değeridir.

Özellik-1 sütununun medyanı = (8 + 10) / 2 = 9

Özellik_{11} = 2 / 9 = 0.22

Özellik_{12} = 8 / 9 = 0.88

Özellik_{13} = 10 / 9 = 1.11

Özellik_{14} = 12 / 9 = 1.33

Çizelge 2.7'ye göre Özellik-2 sütunun küçükten büyüğe sıralandığı zaman orta olan verileri 3 ve 7 değeri olacaktır. Bu iki değerin ortalaması olan 5 değeridir.

Özellik-2 sütununun medyanı = (3 + 7) / 2 = 5

Özellik_{21} = 7 / 5 = 1.4

Özellik_{22} = 17 / 5 = 3.4

Özellik_{23} = 2 / 5 = 0.4

Özellik_{24} = 3 / 5 = 0.6

Çizelge 2.7'ye göre Özellik-3 sütunun küçükten büyüğe sıralandığı zaman orta olan verileri 8 ve 14 değeri olacaktır. Bu iki değerin ortalaması olan 11 değeridir.

Özellik-3 sütununun medyanı = (8 + 14) / 2 = 11

Özellik_{31} = 22 / 11 = 2

Özellik_{32} = 14 / 11 = 1.27

Özellik_{33} = 1 / 11 = 0.09

Özellik_{34} = 8 / 11 = 0.72

Çizelge 2.7'ye göre Özellik-4 sütunun küçükten büyüğe sıralandığı zaman orta olan verileri 4 ve 8 değeri olacaktır. Bu iki değerin ortalaması olan 6 değeridir.

Özellik-4 sütununun medyanı = (4 + 8) / 2 = 6

Özellik_{41} = 1 / 6 = 0.16

Özellik_{42} = 4 / 6 = 0.66

Özellik_{43} = 8 / 6 = 1.33

Özellik_{44} = 12 / 6 = 2

Yukarıdaki yapılan işlemlerin sonrasında Çizelge 2.7'de verilen 4 kişiye ait olan 4 farklı özelliğin medyan normalizasyon yöntemi uygulanarak ölçeklenmiş değerleri bulunmuştur. Bu ölçeklenmiş değerler Çizelge 2.8'de gösterilmiştir.

Çizelge 2.8 Medyan yöntemi için normalize edilmiş veri seti

	Özellik-1	**Özellik-2**	**Özellik-3**	**Özellik-4**
Kişi-1	0.22	1.4	2	0.16
Kişi-2	0.88	3.4	1.27	0.66
Kişi-3	1.11	0.4	0.09	1.33
Kişi-4	1.33	0.6	0.72	2

2.5 D_Minimum-Maksimum Normalizasyon Yöntemi

Bu normalizasyon yönteminde veriler 0.1 ile 0.9 arasına ölçeklenir. Normalizasyon yapılarak veriler boyutsuz hale getirilir. Lineer bir dönüşüm oluşur. Yeni ölçeklenmiş veriler ile standart sapma azaltılarak uç verilerin etkisi azaltılır. Ama bu normalizasyon yöntemi keskin değerleri çok iyi alamaz (Yavuz S. Deveci M.,2012;www.oreilly.com/library/view/regressionanalysiswith/9781788627306/6bb0d820-6200-4bfe-aa91-e7b7ffa2a9c1.xhtml).

Bu normalizasyon yöntemi için Eşitlik 2.6 kullanılır.

$$x' = 0.8 * \frac{(x_i - x_{min})}{(x_{max} - x_{min})} + 0.1 \qquad (2.6)$$

Eşitlik 2.6'da

x' = Normalize edilmiş değeri

x_i = Normalize edilecek değeri

x_{min} = Veri setindeki en küçük değeri

x_{max} = Veri setindeki en büyük değeri ifade etmektedir.

Çizelge 2.9'da D_min-mak normalizasyon yöntemini daha iyi anlamak için kullanılacak bir veri seti görülmektedir. Bu veri seti 4 farklı kişiye ait 4 farklı özelliğe sahiptir.

Çizelge 2.9 D_min-mak yöntemi için normalize edilmemiş veri seti

	Özellik-1	**Özellik-2**	**Özellik-3**	**Özellik-4**
Kişi-1	5	1	22	4
Kişi-2	9	51	15	3
Kişi-3	8	30	18	40
Kişi-4	17	10	1	24

Çizelge 2.9'de görülen 4 kişiye ait veriler D_min-mak normalizasyon yöntemi ile 0.1-0.9 aralığına Özellik-1 sütunundan Özellik-4 sütununa sırasıyla ölçeklenecektir. Buna göre:

Özellik-1 sütunu için veri setinin minimum ve maksimum değeri bulunmalıdır. Buna göre,

Özellik-1 sütununda en büyük değer = 17

Özellik-1 sütununda en küçük değer = 5

Sonrasında Eşitlik 2.6 kullanılarak Özellik-1 sütununun ölçeklenmiş değerleri elde edilir.

$$\text{Özellik}_{11} = 0.8 * ((5 - 5) / (17 - 5) + 0.1) = 0.1$$

$$\text{Özellik}_{12} = 0.8 * ((9 - 5) / (17 - 5) + 0.1) = 0.36$$

$$\text{Özellik}_{13} = 0.8 * ((8 - 5) / (17 - 5) + 0.1) = 0.3$$

$$\text{Özellik}_{14} = 0.8 * ((17 - 5) / (17 - 5) + 0.1) = 0.9$$

Özellik-2 sütunu için veri setinin minimum ve maksimum değeri bulunmalıdır. Buna göre,

Özellik-2 sütununda en büyük değer = 51

Özellik-2 sütununda en küçük değer = 1

Sonrasında Eşitlik 2.6 kullanılarak Özellik-2 sütununun ölçeklenmiş değerleri elde edilir.

$$\text{Özellik}_{21} = 0.8 * ((1 - 1) / (51 - 1) + 0.1) = 0.1$$

$$\text{Özellik}_{22} = 0.8 * ((51 - 1) / (51 - 1) + 0.1) = 0.9$$

$$\text{Özellik}_{23} = 0.8 * ((30 - 1) / (51 - 1) + 0.1) = 0.56$$

$$\text{Özellik}_{24} = 0.8 * ((10 - 1) / (51 - 1) + 0.1) = 0.24$$

Özellik-3 sütunu için veri setinin minimum ve maksimum değeri bulunmalıdır. Buna göre,

Özellik-3 sütununda en büyük değer = 22

Özellik-3 sütununda en küçük değer = 1

Sonrasında Eşitlik 2.6 kullanılarak Özellik-3 sütununun ölçeklenmiş değerleri elde edilir.

$Özellik_{31}$ = 0.8 * ((22 – 1) / (22 – 1) + 0.1) = 0.9

$Özellik_{32}$ = 0.8 * ((15 – 1) / (22 – 1) + 0.1) = 0.63

$Özellik_{33}$ = 0.8 * ((18 – 1) / (22 – 1) + 0.1) = 0.74

$Özellik_{34}$ = 0.8 * ((1 – 1) / (22 – 1) + 0.1) = 0.1

Özellik-4 sütunu için veri setinin minimum ve maksimum değeri bulunmalıdır. Buna göre,

Özellik-4 sütununda en büyük değer = 40

Özellik-4 sütununda en küçük değer = 3

Sonrasında Eşitlik 2.6 kullanılarak Özellik-4 sütununun ölçeklenmiş değerleri elde edilir.

$Özellik_{41}$ = 0.8 * ((4 – 3) / (40 – 3) + 0.1) = 0.12

$Özellik_{42}$ = 0.8 * ((43– 3) / (40 – 3) + 0.1) = 0.1

$Özellik_{43}$ = 0.8 * ((40 – 3) / (40 – 3) + 0.1) = 0.9

$Özellik_{44}$ = 0.8 * ((24 – 3) / (40 – 3) + 0.1) = 0.55

Yukarıdaki yapılan işlemlerin sonrasında Çizelge 2.9'de verilen 4 kişiye ait olan 4 farklı özelliğin D_min_mak normalizasyon yöntemi uygulanarak ölçeklenmiş değerleri bulunmuştur. Bu ölçeklenmiş değerler Çizelge 2.10'da gösterilmiştir.

Çizelge 2.10 D_Min-mak yöntemi için normalize edilmiş veri seti

	Özellik-1	**Özellik-2**	**Özellik-3**	**Özellik-4**
Kişi-1	0.1	0.1	0.9	0.12
Kişi-2	0.36	0.9	0.63	0.1
Kişi-3	0.3	0.56	0.74	0.9
Kişi-4	0.9	0.24	0.1	0.55

Çizelge 2.10'da 4 kişiye ait olan 4 farklı özelliğin D_min_mak normalizasyon yöntemi uygulanarak ölçeklenmiş değerleri gösterilmektedir. Bu normalizasyon yönteminde veriler min-mak yöntemine göre 0.9-0.1 arasına ölçeklenmiştir. Kısaca min-mak yöntemini en genel formül ile düzenlersek 7 numaralı denklemi kullanmalıyız.

Bu denkleme göre yeni minimum ve maksimum noktalarını belirleyip ölçeklendirme yapabiliriz. Yeni hesaplama için Eşitlik 2.7 kullanılır.

$$x' = \frac{x_i - x_{min}}{x_{max} - x_{min}} (New_{max} - New_{min}) - New_{min} \quad (2.7)$$

Yukarıdaki eşitliğe göre:

0.8 = 0.9-0.1 eşitliği 0.8= $New_{max} - New_{min}$ eşitliğinden geldiği görülecektir (Yavuz S. Deveci M.,2012).

2.6 Norm Normalizasyon Yöntemi

Herhangi bir vektörün normu ya da uzunluğu Öklid mesafesine eşittir. Norm normalizasyonu aynı zamanda vektör normalizasyonu olarak isimlendirilir (Gautam,2015). Norm hesabı için Eşitlik 2.8 kullanılır (Eesa A.S., Arabo W. K.,2017).

$$|x| = \sqrt{x_1^2 + x_2^2 + x_3^2 + ... + x_i^2} \quad (2.8)$$

Eşitlik 2.8'de:

$|x|$=Normalize edilecek verilerin normunu ifade etmektedir.

Norm normalizasyon için Eşitlik 2.9 kullanılır.

$$x' = \frac{x_i}{|x|} \quad (2.9)$$

Eşitlik 2.9'da:

x' = Normalize edilmiş veriyi

x_i = Normalize edilecek değeri ifade etmektedir.

Çizelge 2.11'de norm normalizasyon yöntemini daha iyi anlamak için kullanılacak bir veri seti görülmektedir. Bu veri seti 4 farklı kişiye ait 4 farklı özelliğe sahiptir.

Çizelge 2.11 Norm normalizasyon yöntemi için normalize edilmemiş veri seti

	Özellik-1	**Özellik-2**	**Özellik-3**	**Özellik-4**
Kişi-1	5	1	22	4
Kişi-2	3	10	2	3
Kişi-3	8	4	7	8
Kişi-4	1	1	1	8

Çizelge 2.11'de görülen 4 kişiye ait veriler norm normalizasyon yöntemi ile Özellik-1 sütunundan Özellik-4 sütununa sırasıyla normalize edilecektir.

Buna göre:

$|\text{Özellik-1}|$ = Özellik-1 sütununun norm değeri

$= \sqrt{5^2 + 3^2 + 8^2 + 1^2}$

= 9.94 olarak hesaplanır.

Özellik-1 sütununa ait yeni değerler.

$\text{Özellik}_{11} = 5 / 9.94 = 0.50$

$\text{Özellik}_{12} = 3 / 9.94 = 0.30$

$\text{Özellik}_{13} = 8 / 9.94 = 0.80$

$\text{Özellik}_{14} = 1 / 9.94 = 0.10$

$|\text{Özellik-2}|$ = Özellik-2 sütununun norm değeri

$= \sqrt{1^2 + 10^2 + 4^2 + 1^2}$

= 10.86 olarak hesaplanmıştır.

Özellik-2 sütununa ait yeni değerler.

$\text{Özellik}_{21} = 1 / 10.86 = 0.09$

$\text{Özellik}_{22} = 10 / 10.86 = 0.92$

$\text{Özellik}_{23} = 4 / 10.86 = 0.36$

$\text{Özellik}_{24} = 1 / 10.86 = 0.09$

$|\text{Özellik-3}|$ = Özellik-3 sütununun norm değeri

$= \sqrt{22^2 + 2^2 + 7^2 + 1^2}$

= 23.19 olarak hesaplanmıştır.

Özellik-3 sütununa ait yeni değerler.

Özellik_{31} = 22 / 23.19 = 0.95

Özellik_{32} = 2 / 23.19 = 0.08

Özellik_{33} = 7 / 23.19 = 0.30

Özellik_{34} = 1 / 23.19 = 0.04

$|\text{Özellik-4}|$ = Özellik-4 sütununun norm değeri

$= \sqrt{4^2 + 3^2 + 8^2 + 8^2}$

= 12.36 olarak hesaplanmıştır.

Özellik-4 sütununa ait yeni değerler.

Özellik_{41} = 4 / 12.36 = 0.32

Özellik_{42} = 3 / 12.36 = 0.24

Özellik_{43} = 8 / 12.36 = 0.65

Özellik_{44} = 8 / 12.36 = 0.65

Yukarıdaki yapılan işlemlerin sonrasında Çizelge 2.11'de verilen 4 kişiye ait olan 4 farklı özelliğin norm normalizasyon yöntemi uygulanarak ölçeklenmiş değerleri bulunmuştur. Bu ölçeklenmiş değerler Çizelge 2.12'de gösterilmiştir.

Çizelge 2.12 Norm normalizasyon yöntemi için normalize edilmiş veri seti

	Özellik-1	**Özellik-2**	**Özellik-3**	**Özellik-4**
Kişi-1	0.5	0.09	0.95	0.32
Kişi-2	0.3	0.92	0.08	0.24
Kişi-3	0.8	0.36	0.30	0.65
Kişi-4	0.1	0.09	0.04	0.65

Ayrıca örnek bir sütunun normalize edilmiş verinin norm değerinin 1`e eşittir (Abdi H., 2010). Buna göre,

$|x|$ = Normalize edilmiş Özellik-1 sütununun Normu

$= \sqrt{0.5^2 + 0.3^2 + 0.8^2 + 0.1^2}$

=1 olarak hesaplanmıştır.

2.7 Medyan-Mod Normalizasyon Yöntemi

Medyan-Mod normalizasyonu yöntemi için Eşitlik 2.10 kullanılır.

$$x' = \frac{x_i - \text{Medyan}(x_i)}{\text{MAD}(x_i)} \tag{2.10}$$

Bu normalizasyon yöntemi oldukça anormal skorlara duyarsızdır. O giriş dağılımını korumaz ve skorları ortak bir aralığa dönüştürmez. Bundan dolayı MAD değeri hesaplanır (Basheer I.A, Hajmeer M,2000).

MAD (mean absolute deviation) değeri hesabı için Eşitlik 2.11 kullanılır.

$$p = \frac{1}{N}\sum_i^N |x_i - \text{Median}(x_i)| \tag{2.11}$$

2.8 Ortalama-Mod Normalizasyon Yöntemi

Bu normalizasyon yönteminde verilen veri setinin medyanı yerine mod değeri kullanılır. Ortalama-Mod normalizasyonu yöntemi için Eşitlik 2.12 kullanılır.

$$x' = \frac{x_i - \text{Ortalama}(xi)}{\text{MAD}(x_i)} \tag{2.12}$$

Bu normalizasyon yöntemi de medyan-mod normalizasyonuna benzerdir ve oldukça anormal skorlara duyarsızdır. O da giriş dağılımını korumaz ve skorları ortak bir aralığa dönüştürmez. Bundan dolayı MAD değeri hesaplanır (Basheer I.A, Hajmeer M, 2000).

MAD değeri hesabı için Eşitlik 2.13 kullanılır.

$$p = \frac{1}{N}\sum_{i}^{N} | x_i - \text{Ortalama}(x_i) | \quad (2.13)$$

2.9 Normalizasyon Yöntemi Seçimi

Normalizasyon yöntemi seçimi sahip olunan hem veri setine hem de çalışmanın amaçlarına bağlı olmakla birlikte genelde z-skor en yaygın olarak kullanılan metottur. Fakat yöntem seçimi farklılık göstermektedir (Nandakumar K., 2005)

Örneğin bazen daha iyi öğrenmeyi sağlamak için min-mak yöntemi yerine D_min_max yöntemi kullanılır. Böylece sigmoid fonksiyonunun doyması önlenir. Burada yeni verimiz 0-1 arasına ölçeklendirmek yerine 0.1-0.9 arasına sıkıştırılır. Ayrıca veri setindeki dağılımı değiştirmek istemezsek de min-mak yöntemi iyi olabilir. Böylece veri setimiz bozulmaz. Şayet veri setimizde çok uç değerler mevcut ise kuvvetli z-skor yöntemi kullanılabilir. Bu yöntemle uç noktaların etkisi azaltılabilir. Şayet normal bir dağılım istiyorsak z-skor yöntemini de kullanabiliriz. Ek olarak veri setimizin varyanslardan etkilenmemesini istiyorsak medyan yöntemi uygun bir yöntem olabilir. Bazen normalizasyon yapılmadan önce farklı yaklaşımlarda olabilmektedir. Örneğin çok değerli veriler olduğu zaman verinin logaritmasını almak gerekebilir (Basheer I.A. ve Hajmee M.,2000).

Fakat unutulmamalıdır ki veriler sütun sütun normalize edileceğinden sütun bazında bağımsız olacaktır. Burada bir sütunda birim ağırlık türünde, bir sütunda uzunluk biriminde, bir sütunda doğru-yanlış seçim biçiminde olsa bile minimum ve maksimum değerleri bağımsız olacaktır (Akdemir B., 2009).

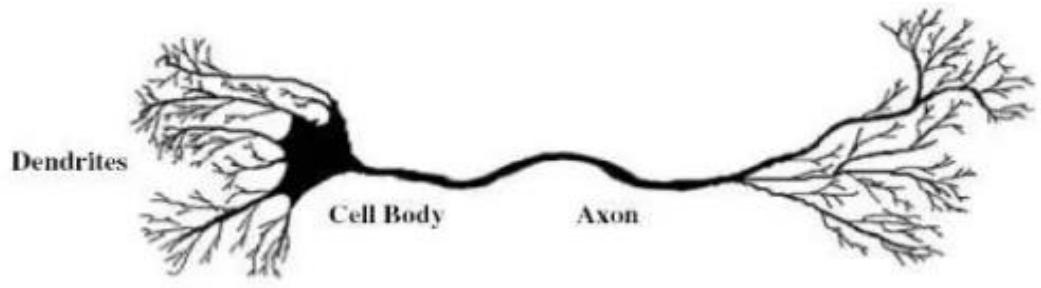
3. MATERYAL VE METOD

Yapılan bu çalışmada Pima Hintlilerinin diyabet hastalığı verisi, göğüs kanseri hastalığı verisi, karaciğer hastalığı verisi ve kalp hastalığı verisi YSA, DVM, Naive Bayes, k-NN ve KA gibi sınıflandırma yöntemleri kullanılarak 4 farklı k-kat çaprazlama (2,5,10,20) kriteri altında sınıflandırma işlemine tutulmuştur. Sınıflandırma işlemi ORANGE programı kullanılarak yapılmıştır. Şimdi yukarıda belirtilen sınıflandırma yöntemleri ve ORANGE programını inceleyelim.

3.1 Yapay Sinir Ağları

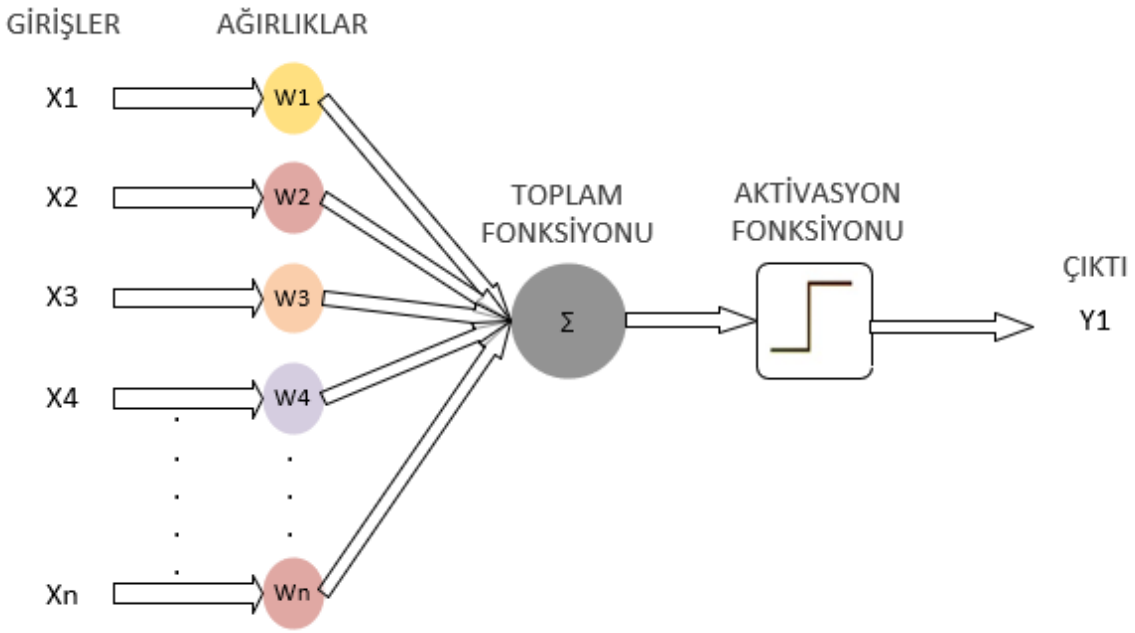
Yapay sinir ağları (YSA) sorunlara çözüm bulmak amacıyla insan beyninin düşünme, öğrenme, taklit etme, tahmin etme gibi birçok özelliklerini kullanarak geliştirilmiş bilgisayar yazılımlarıdır. Bildiğimiz gibi insan beynine bilgi ulaşır, değerlendirilir ve bu değerlendirme sonuçlanır. Amaç beynimizi matematiksel olarak modellemesidir. Bu modelleme düşüncesi makineler insan gibi düşünebilir mi fikrini ortaya atan İngiliz matematikçi ve bilgisayar bilimci olan Alan Mathison Turing tarafından ortaya atılmıştır (Yazıcı,2007; https://kod5.org/yapay-sinir-aglari-ysa-nedir/).

Şekil 3.1'de insan sinir ağı genel görünümü görülmektedir. YSA sınıflandırma algoritması da insan sinir ağı modeline benzetilmektedir. Burada dentrites (dentrit) toplama fonksiyonunu, Cell body aktivasyon fonksiyonunu ve axon (akson) çıkış elemanını ifade etmektedir (https://kod5.org/yapay-sinir-aglari-ysa-nedir/)



Şekil 3.1 İnsan sinir ağı genel görünümü
(https://kod5.org/yapay-sinir-aglari-ysa-nedir/)

Şekil 3.2'de Şekil 3.1'de görülen insan sinir ağı modelinin YSA algoritmasında modellenmesi gösterilmiştir. Bu model üç temel kurala sahiptir; bunlar çarpma, toplama ve aktivasyon fonksiyonu. Burada dışarıdan gelen input bir ağırlıkla çarpılır ve sonra bir bias değeri ile toplanır. Sonuç çeşitli aktivasyon fonksiyonlarından geçirilerek çıkış olarak iletilir (Mohamed,2017).

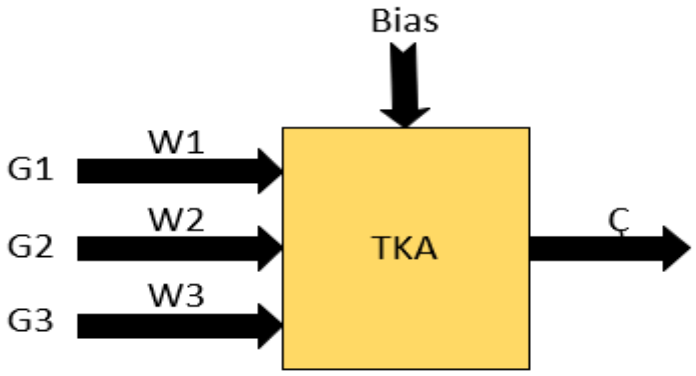


Şekil 3.2 YSA modeli

3.1.1 Tek katmanlı algılayıcılar

Bir yapay sinir ağı tek katmanlı algılayıcılardan oluşuyorsa sadece girdi ve çıktı katmanından oluşuyor demektir. Girdi üniteleri doğrudan çıktı ünitelerine bağlanır. Tek katmanlı algılayıcıların basit problemleri çözmede iyiyken problem karmaşıklaştıkça çözümden uzaklaştığı gözlemlenmiştir. Tek katmalı algılayıcılar doğrusal bir çıktıya sahip olup 1 veya -1 değerini alır. (Öztemel,2006;https://medium.com/@k.ulgen90/makine-%C3%B6%C4%9Frenimi b%C3%B6l%C3%BCm-3-4b160df1f4c8).

Şekil 3.3'de tek katmanlı algılayıcı modeli görülmektedir. Şekilde girişler, toplama-aktivasyon görevlerini yerine getiren bir gövde ve çıkış yer almaktadır. Bu bakımdan çok-girişli tek çıkışlı bir yapıya sahiptirler (Öğücü M. O.,2006)

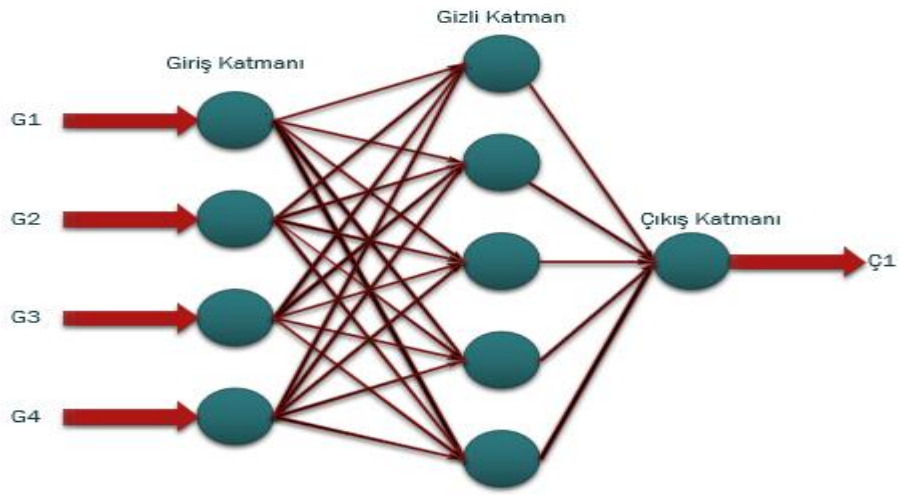


Şekil 3.3 Tek katmanlı algılayıcılar

3.1.2 Çok katmanlı algılayıcılar

Tek katmanlı algılayıcılardaki eksikliği gidermek için ortaya çıkmıştır. Burada yapısal olarak doğrusal bir aktivasyon fonksiyonuna sahip olmayan birçok nöronun belli bir üstünlük içerisinde bağlanmasıdır (Öztürk,2018)

Şekil 3.4'te çok katmanlı algılayıcı modeli gösterilmiştir. Görüldüğü gibi çok katmanlı algılayıcı girdi katmanı, gizli katman ve çıktı katmanından oluşmaktadır. Bazı uygulamalarda birden çok gizli katman yer alabilir. Şekilde görüldüğü gibi ara katmana gelen bilgi giriş katmanı ile iletilir. Giriş katmanından gelen bilgiler gizli katmanda işlenir ve çıkış katmanına iletilir. Çıkış katmanı ara katman yani gizli katmandan gelen bilgiye göre bir çıkış üretir (Gör İ.; Öğücü M. O.,2006).

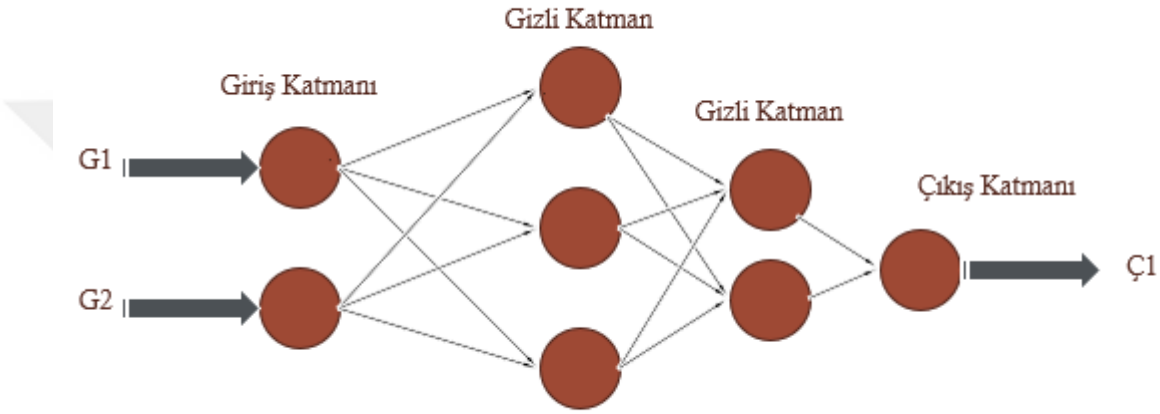


Şekil 3.4 Çok katmanlı algılayıcılar

3.1.3 İleri beslemeli yapay sinir ağları

Bu ağ modelinde nöronlar girişten çıkışa doğru düzenli katmanlar biçimindedirler. Bir katmandan bir sonraki katmana bağlantı vardır. Girişten bilgiler gizli katmana iletilir sonra sırasıyla çıkışa iletilir (Öztürk,2018).

Şekil 3.5'te ileri beslemeli yapay sinir ağı modeli görülmektedir. Burada bir katmandan sadece kendisinden sonra gelen katmana bir bağlantı vardır (Öztürk K., Şahin M. E., 2018)

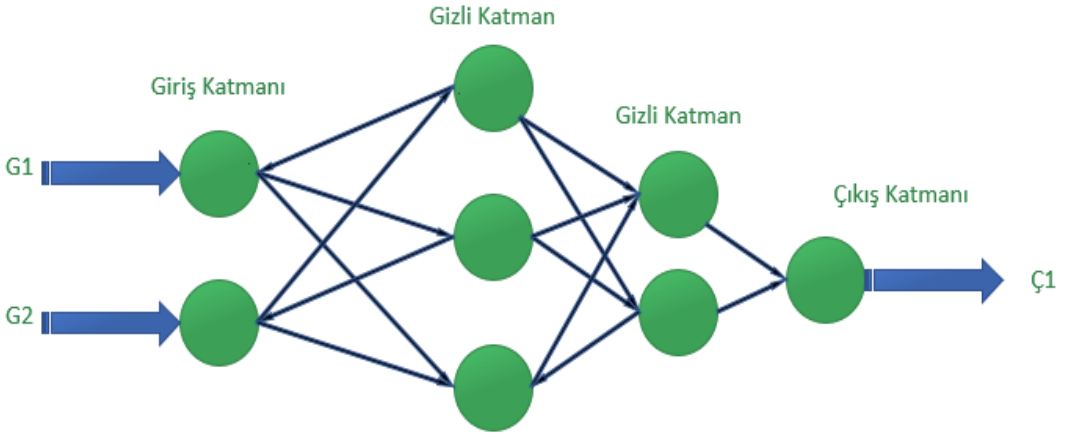


Şekil 3.5 İleri beslemeli yapay sinir ağları

3.1.4 Geri beslemeli yapay sinir ağları

Bu ağ yapısı ileri beslemeli yapay sinir ağlarından farklı olarak bir nöronun çıktısı hem kendinden sonraki nörona hem de kendi katmanından önceki bir nörona veya kendi katmanında bulunan başka bir nörona girdi olarak verilebilir. Doğrusal bir ilerleme yoktur (Öztürk,2018).

Şekil 3.6'da geri beslemeli yapay sinir ağı modeli görülmektedir. Burada bir sinyal ileri ya da geri yönde ilerleyebilir. Geri beslemeli yapay sinir ağları çok güçlüdürler ve karmaşık olabilirler (https://msatechnosoft.in/blog/artificial-neural-network-types-feed-forward-feedback-structure-perceptron-machine-learning-applications/)

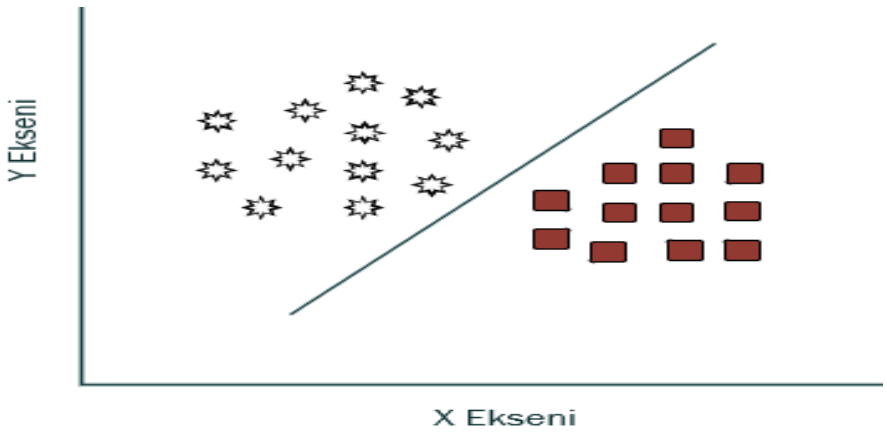


Şekil 3.6 Geri beslemeli yapay sinir ağı

3.2 Destek Vektör Makinesi

Destek vektör makinesi (DVM) istatiksel öğrenme teorisini esas alan parametrik olmayan denetimli öğrenim yöntemlerinden biridir. İlk defa Vapnik tarafından 1992 yılında tanıtılmıştır. Bu yöntemin çalışma prensibi, iki sınıfı birbirinden ayıran en uygun karar verme fonksiyonunun tanımlanmasıdır. Kısaca en uygun hiper düzlemin tanımlanması şeklindedir (Kavzoğlu,2010; Mohamed,2017)

Şekil 3.7'de DVM sınıflandırma algoritmasını tanımlamak için bir şekil görülmektedir. DVM sınıflandırma algoritmasında amaç iki sınıfı en iyi ayıran sınırın tanımlanmasıdır. Burada yıldız ve kare sınıflarını en iyi ayıran düzlem çizilmiştir (https://veribilimcisi.com/2017/07/19/destek-vektor-makineleri-support-vector-machine/).



Şekil 3.7 DVM sınıflandırma

Bu yöntem ikili sınıflandırma amaçlı geliştirilmiştir ve şayet az sayıda örnekleme verisine sahip isek bu yöntemle daha doğru sınıflandırma sonuçlarına sahip oluruz. İlk olarak iki sınıflı doğrusal olmayan verilerin sınıflandırılması için geliştirilen bu yöntem daha sonra çok sınıflı doğrusal olmayan verilerin sınıflandırılması için genelleştirilmiştir. (Üstüner,2013).

DVM'nin ana avantajı göreceli olarak eğitimi kolay bir sınıflandırma yöntemidir ve yüksek boyutlu verileri göreceli olarak iyi ölçekler. Modelin karmaşıklığı ve hatası arasındaki değiş tokuş kolay bir şekilde kontrol edilir. Doğruluk performansı yüksektir ve iyi genelleme yapabilir. DVM' nin ana dezavantajı özellikler yorumlanamazsa sınıflama yapması zordur. Ayrıca DVM pahalıdır ve lineerden uzak veri setlerini ayırmak için kullanılan iyi bir kernel fonksiyonuna sahip olmalıdır (Mohamed,2017).

3.3 Naive Bayes

Bu sınıflandırma yöntemi ismini Matematikçi Thomas Bayes'den almıştır. Bu algoritma olasılık işlemleri temelli bir dizi hesaplama ile sisteme girilen verilerin sınıfını belirlemeye yardımcı olur. Her verinin sınıflandırmaya katkı sağladığı varsayılır. Sınıflandırma işleminin temeli Bayes teoremine dayanmaktadır ve genellikle veri boyutu büyük olduğu zaman tercih edilir (https://kodedu.com/2014/05/naive-bayes-siniflandirma-algoritmasi/; Jadhav,2013)

Bayes teoremi yöntemi için Eşitlik 3.1 kullanılır.

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)} \tag{3.1}$$

P(A): A olayının gerçekleşme durumu
P(B): B olayının gerçekleşme durumu
P(A\B): B olayının olması durumunda A olayının gerçekleşme durumu
P(B\A): A olayının olması durumunda B olayının gerçekleşme durumu

Çok özellikli Bayes teoremi yöntemi için Eşitlik 3.2 kullanılır.

$$P(C/F1,...Fn) = \frac{P(C)p(F1,...Fn/C)}{p(F1,...Fn)} \tag{3.2}$$

Burada C hedefimizi göstermekte olup F ise özelliklerimizi temsil etmektedir. Özetle belirtilmelidir ki Naive Bayes yöntemi bütün koşullu olasılıkların çarpımı olarak değerlendirilir. (https://erdincuzun.com/makine_ogrenmesi/naive-bayes-classifier/)

Sınıflandırma yönteminde sistem belirli miktarda sınıfı olan öğretilmiş veri ile beslenir. Öğretilmiş veriler üzerinde yapılan olasılık hesabı ile sisteme sunulan belirli bir sınıfa ait olan veriler üzerinde sınıflandırma yapılmaya çalışılır. Bilinmelidir ki öğretilmiş veri sayısı ne kadar fazla ise test verilerinin sınıflandırma doğruluğu daha da yüksek olacaktır. (https://kodedu.com/2014/05/naive-bayes-siniflandirma-algoritmasi/)

Naive Bayes eğitildikten sonra performansı yüksek olarak çalışır fakat eğitilmiş bir sistemin güncellenmesi kaynak ve zaman bakımından maliyetli olur. Nedeni ise her veri kümesi tekrar sınıflandırma işlemi için tekrar taranması gerekir. Ayrıca iyi bir sonuç almak için Naive Bayes yöntemi çok fazla kayıt verisi gerektirir (Jadhav,2013).

Naive Bayes yöntemi metin madenciliği, duygu analizi, spam filtrelemede, v.b. uygulamalarda kullanılmaktadır. Bu yöntemin hızlı ve verimli olması, ilgisiz özelliklere duyarsız olması gibi iyi özellikleri vardır. En büyük dezavantajları ise sınıflandırıcılarının bağımsız olması gerekliliğidir (https://devhunteryz.wordpress.com/2019/12/02/naive-bayes-siniflandirici/), (https://medium.com/@Emreyz/y%C3%B6ntemler-3-naive-bayes-899314be2018), (https://medium.com/yapay-zeka-makine-%C3%B6%C4%9Frenmesi-derin-%C3%B6%C4%9Frenme/denetimli-%C3%B6%C4%9Frenme-d7237c50b10b).

3.4 k- En Yakın Komşu Algoritması

k-En Yakın Komşu yöntemi (k-NN), sınıflandırma yöntemlerinde kullanılan denetimli öğrenme yöntemlerinden birisidir. Bu yönteme göre, sınıflandırılması yapılacak verilerin, normal verilere göre davranışları incelenerek en yakın olduğu düşünülen k adet veri bulunur. Sonrasında bu k adet verinin ortalaması alınır ve bu eşiğe göre sınıflandırma işlemi yapılır. k-NN verilerin dağılımı hakkında çok az ya da önceden hiç bilgi olmadığı zaman temel ve en basit sınıflandırma tekniğidir. Bu yöntemde önemli olan verilerin özelliklerinin net olmasıdır. (Çalışkan S. B., Soğukpınar İ., 2008; Bolandraftar,2013)

k-NN yöntemi uzaklığı hesaplarken genelde 3 yöntemden faydalanmaktadır. Bunlar:

- Öklid Uzaklığı
- Minkowski Uzaklığı

- Manhattan Uzaklığı

Öklid uzaklığı yöntemi için Eşitlik 3.3 kullanılır.

$$d(a,b) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (3.3)$$

Manhattan uzaklığı yöntemi için Eşitlik 3.4 kullanılır.

$$d(a,b) = \sum_{i=1}^{k}|x_i - y_i| \quad (3.4)$$

Minkowski uzaklığı yöntemi için Eşitlik 3.5 kullanılır.

$$d(a,b) = (\sum_{i=1}^{k}(|x_i - y_i|)^q)^{1/q} \quad (3.5)$$

k-NN sınıflandırma yönteminde k bilinmeyen noktanın en yakın komşularını temsil etmekte olup k=1 olduğu zaman en basit halidir ve her bir örnek onu çevreleyen örneklere benzer olarak sınıflandırma işlemine tabi tutulacaktır. Şayet bir örneğin sınıfı bilinmiyorsa sınıflandırma işlemini ona en yakın örneğin sınıfına göre olacaktır (Bolandraftar,2013). k-NN, tembel öğrenme türüdür; buradaki işlev sadece yerel olarak yaklaştırılır ve tüm hesaplama, sınıflandırma işlemi yapılana kadar ertelenir.
(https://veribilimcisi.com/2017/07/20/k-en-yakin-komsu-k-nearest-neighborsknn/)

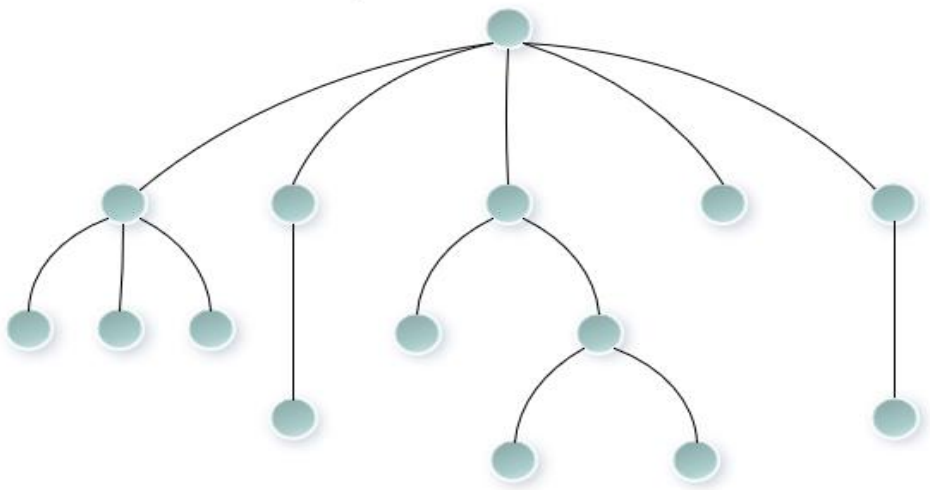
k-NN sıklıkla veri sıkıştırma, istatistik, görüntü işleme, veri madenciliği ve makine öğrenmesi gibi çeşitli uygulamalarda kullanılır. (https://medium.com/yapay-zeka-makine-%C3%B6%C4%9Frenmesi-derin-%C3%B6%C4%9Frenme/denetimli %C3%B6%C4%9Frenme-d7237c50b10b)

3.5 Karar Ağaçları

Karar ağaçları (KA) veri madenciliğinde en sık kullanılan sınıflandırma ve tahmin yöntemlerinden bir tanesidir. KA kolay yorumlanma ve anlaşılma özelliklerine sahip olmakla beraber düşük maliyetli ve güvenilir bir yöntemdir. KA veri iyi şekilde açıklamayan çok karmaşık ağaçlar ortaya çıkarabilir. Bazen ezbere öğrenme yapabilir.
(Çalış A, Kayapınar S, Çetinyokuş T., 2014, https://medium.com/@k.ulgen90/makine%C3%B6%C4%9Frenimib%C3%B6l%C3%BCm-5-karar-a%C4%9Fa%C3%A7lar%C4%B1-c90bd7593010)

Bir KA, çok sayıda özelliğe sahip bir veri kümesini, bir dizi karar kuralları uygulayarak daha düşük birimlere ayırmak için kullanılır. Bir KA kök, iç ve yaprak düğümlerinden meydana gelir. Her iç düğümün öznitelik üzerinde bir test şartı vardır. Her bir dalın test koşulunun sonucunu temsil eden bir yaprağı vardır ve her yaprak düğümünün bir sınıf etiketi ile atandığı ağaç yapısı gibi bir akış şemasıdır. İlk düğüm kök düğümdür (Jadhav, 2013).

Şekil 3.8'de KA sınıflandırma algoritmasını tanımlamak için bir şekil görülmektedir. KA sınıflandırma algoritması tek bir düğümle başlar ve bir dizi sorular sorarak belirli noktalara dallanarak ulaşır. Şekil 3.8'de görüldüğü gibi her daire bir karar noktasını ifade etmektedir ve karar sonrası yeni dallanmalar oluşmaktadır. Yeni dallanmalardan yeni karar noktaları oluşabilir veya sonlanabilir (https://medium.com/@ekrem.hatipoglu/machine-learning-prediction-algorithms-decision-tree-random-forest-part-5-2970905c021e).



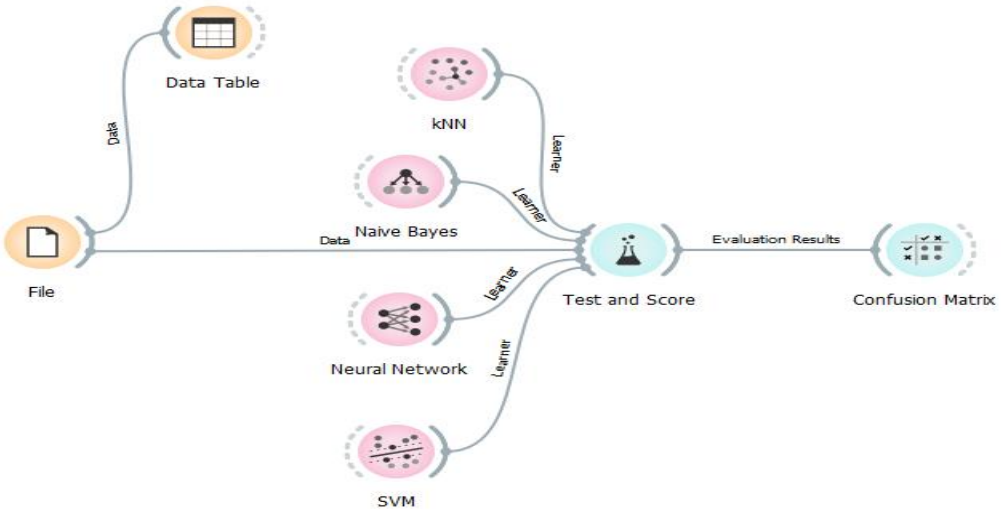
Şekil 3.8 KA sınıflandırma algoritması

Karar ağaçları alt düğümlere ayrılırken birden fazla algoritma kullanır. Alt düğümler oluştukça alt düğümlerin saflığı artacaktır. Algoritma seçimi hedef değişkenin tipine göre değişecektir. (https://medium.com/@k.ulgen90/makine-%C3%B6%C4%9Frenimi-b%C3%B6l%C3%BCm-5-karar-a%C4%9Fa%C3%A7lar%C4%B1-c90bd7593010)

3.6 ORANGE Programı

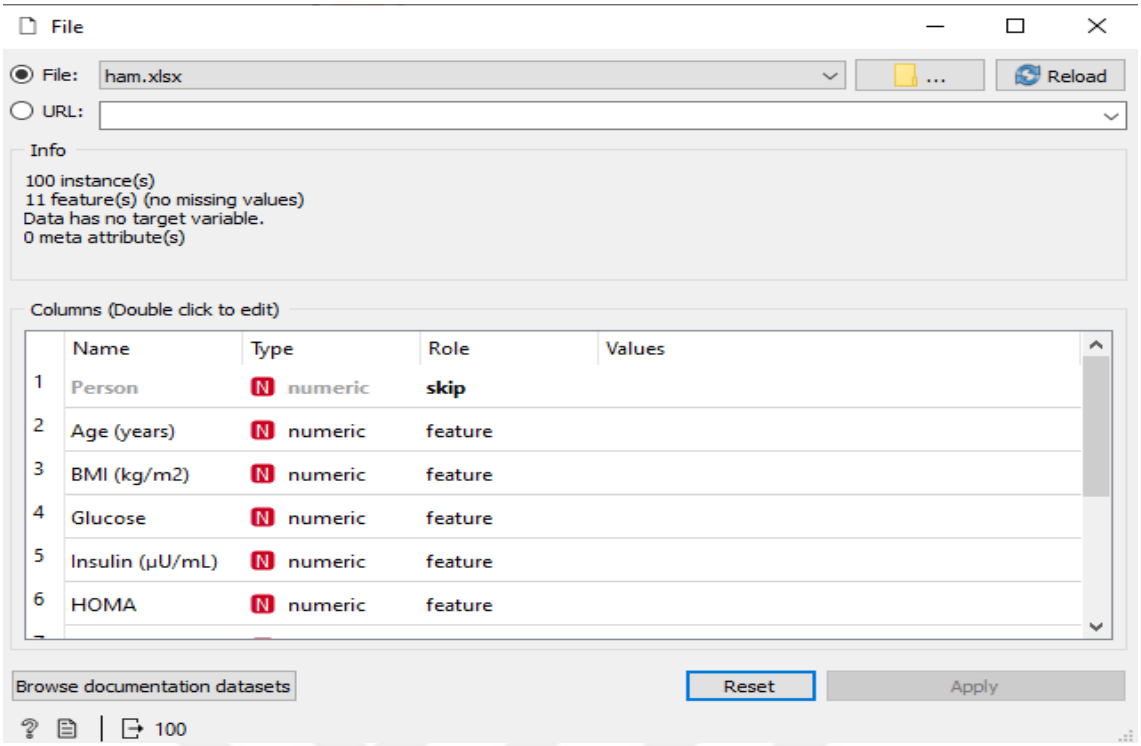
ORANGE programı açık kaynak kodlu bir program olup veri görselleştirme, makine öğrenimi ve veri madenciliği için kullanılmaktadır. Bu program veri analizi için iş akışlarını çeşitli görsel araç kutucukları ile oluşturarak işleyişi kolaylaştırmaktadır. ORANGE programı excel, virgül ve sekme ile ayrılmış dosyaları ve Google e-tablolar gibi çevrim içi dosyaları da okuyabilme yeteneğine sahiptir.
(https://en.wikipedia.org/wiki/Orange_(software)) (https://orangedatamining.com/)
(https://orangedatamining.com/faq/#)

Şekil 3.9'da örnek bir ORANGE programı analizi görülmektedir. Şekilde analiz edilecek dosya 'File' kısmında eklenir. Eklenen veriler 'Data Table' kısmında incelenir ve istenen kategoriye atanabilir. 'Test and Score' kısmında Şekil 3.9'da görülen k-NN, Naive Bayes, Neural Network ve SVM gibi sınıflandırma algoritmalarının performanslarını izlenebilir. Bu tezde ağaç sınıflandırma yöntemi de tezimiz incelemesi içerisinde olacaktır. Her bir sınıflandırma yöntemi kendi görseli üzerinde seçilerek gerekli ayarlamaları yapılabilir. 'Confusion Matrix' kısmında ise karışıklık matrisi görülerek sınıflandırma yönteminin seçicilik ve duyarlılık performansları da değerlendirilebilir.



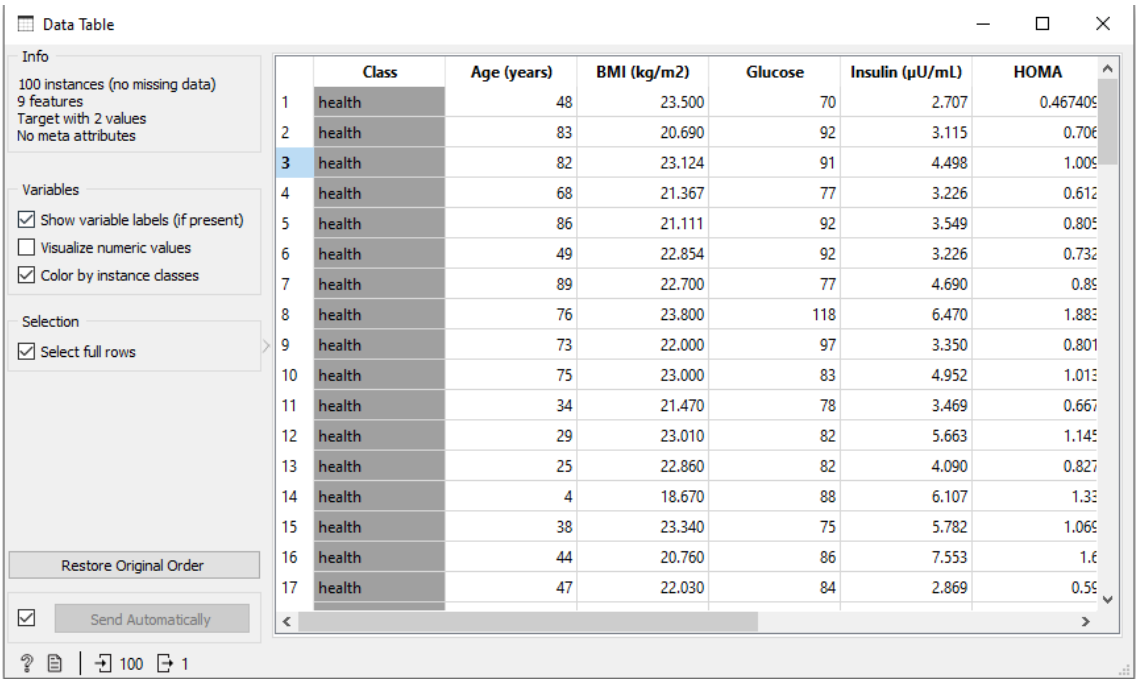
Şekil 3.9 ORANGE program örnek analiz

Şekil 3.10'da 'File' aracı içeriği görülmektedir. Bu kısımda internet üzerinden veri setimizi seçebiliriz. Ya da kendi hazırladığımız excel ya da başka uzantılı veri dosyalarını ekleyebiliriz. Ayrıca veri setimiz içinde yer alan özelliklerin tipini ve rolünü değiştirebiliriz.



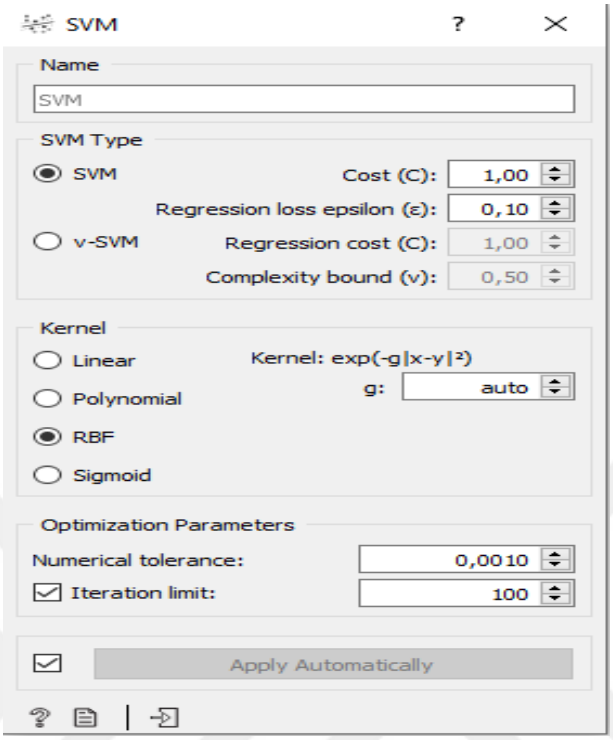
Şekil 3.10 File aracı inceleme

Şekil 3.11'de 'Data Table' aracı içeriği görülmektedir. Bu kısımda 'File' aracı vasıtasıyla yüklediğimiz dosya içerindeki verilerimi kontrol edebiliriz.



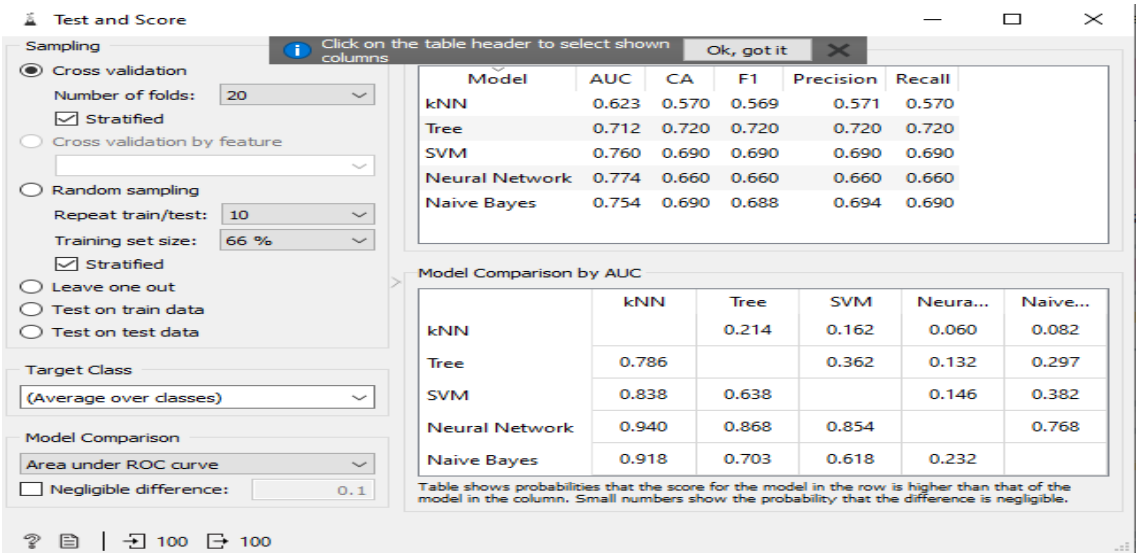
Şekil 3.11 Data Table aracı inceleme

Şekil 3.12 DVM sınıflandırma yönteminin default ayarlarını göstermektedir. Diğer sınıflandırma yöntemleri de benzer şekilde seçilerek ayarları yapılabilir.



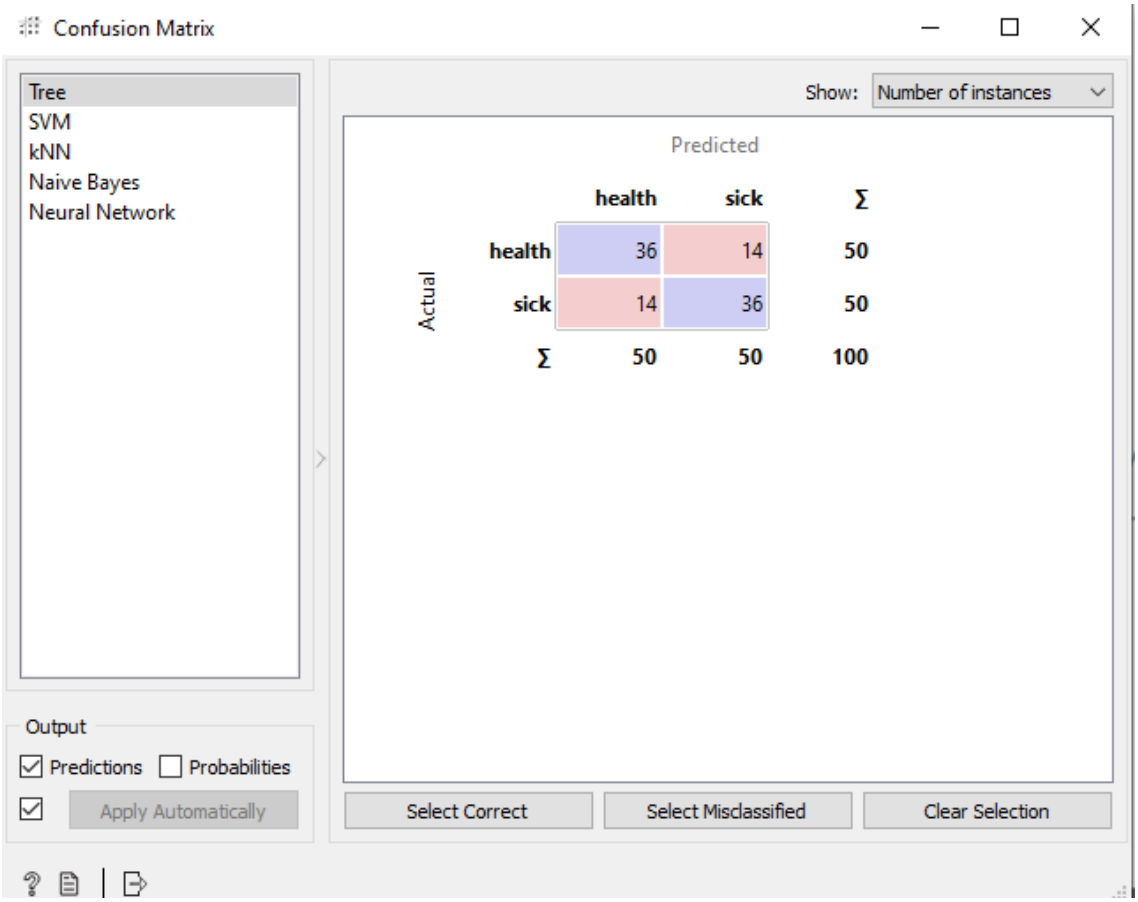
Şekil 3.12 DVM sınıflandırma yöntemi ayarları değiştirme

Şekil 3.13'de görüldüğü gibi mevcut sınıflandırma yöntemlerinin performansları farklı k-kat çaprazlama kriterlerinde **Test and Score** aracı sayesinde izlenebilir. Ya da farklı ayarlamalar yapılabilir.



Şekil 3.13 Sınıflandırma yöntemlerinin performansı

Şekil 3.14'de görüldüğü hata matriside **Confusion Matrix** aracı sayesinde izlenebilir. Bu değerler sayesinde sınıflandırma algoritmasının seçicilik ve duyarlılık performansı değerlendirilebilir.



Şekil 3.14 Hata matrisi

3.7 Değerlendirme Adımları

Veri setimiz ister herhangi bir normalizasyon yöntemi ile normalize edilerek veya normalize edilmeden ham veri olarak sınıflandırma işlemine tabi tutulduktan sonra sonuç doğruluk performansı olarak değerlendirilmiştir. Ayrıca karışıklık matrisi kullanılarak sınıflandırma yönteminin seçicilik ve duyarlılık performansı da değerlendirilebilir.

3.7.1 Sınıflama doğruluğu

Bir sınıflama sistemince yapılan gerçek ve tahmin edilmiş olan sınıflamalar hakkındaki bilgiye karışıklık matrisi ile ulaşabiliriz (Akdemir B.,2009). Bu matris axa

boyutunda olup satırlar; doğru karar sınıflarına, sütunlar ise; sınıflandırıcı tarafından alınan kararlara karşılık gelir. (https://e-abm.com/how-to-establish-quality-and-correctness-of-classification-models-part-3-confusion-matrix/)

Çizelge 3.1'de karışıklık matrisi tablosu görülmektedir. Bu matris tablosu kullanılarak sınıflandırma performansı incelenen sınıflandırma algoritmasının seçicilik ve duyarlılık performansı incelenebilir.

Burada:

TN: Bir örneğin negatif olduğu doğru tahmin sayısını

FP: Bir örneğin pozitif olduğu yanlış tahmin sayısını

FN: Bir örneğin negatif olduğu yanlış tahmin sayısını

TP: Bir örneğin pozitif olduğu doğru tahmin sayısını ifade etmektedir.

Çizelge 3.1 Karışıklık matrisi (Akdemir B.,2009)

Gerçek	Tahmin Edilen	
	Negatif	Pozitif
Negatif	TN	FN
Pozitif	FP	TP

Sınıflama doğruluğu için Eşitlik 3.6 kullanılır.

$$\text{Sınıflama Doğruluğu (\%)} = \frac{TP + TN}{TP + TN + FN + FP} * 100 \qquad (3.6)$$

Bu çalışmada, Pima Hintlilerinin diyabet hastalığı verisi, göğüs kanseri hastalığı verisi, karaciğer hastalığı verisi ve kalp hastalığı verisi DVM, YSA, KA, k-NN ve Naive Bayes sınıflandırma yöntemlerinde 4 farklı k-kat çaprazlama (2,5,10,20) kriteri altında doğruluk performansı karşılaştırılmıştır.

4. ÇALIŞMADA KULLANILAN VERİ SETLERİ

Veri setlerimiz Pima Hintlilerinin Diyabet hastalığı verisi, göğüs kanseri hastalığı verisi, karaciğer hastalığı verisi ve kalp hastalığı verisidir. Bu veri setleri datahub internet adresinden ve UCI internet adresinden alınmıştır.

Kullanılan veri setleri her zaman sınıflama için kullanılacak diye bir genel düşünce yoktur. Bu veriler ileriye dönük bir tahmin yapmak içinde kullanılabilir. Ayrıca bazen elde edilen veri setleri istenilen duruma göre indirgenebilir. Bazen bu durum eksik veriden de kaynaklanabilir. Sınıflandırma yaparken bu veri setlerinden örneğin bir tümörün iyi ya da kötü huylu olduğunu, Diyabet hastalığının var olup olmadığı gibi durumları anlayabiliriz. Sınıflandırma işleminde veri setleri ham veriler ve ham verilerin normalize edilmesi olarak kullanılmıştır. Sınıflandırma işlemi için ise YSA, DVM, KA, Naive Bayes ve KNN metotları kullanılmıştır. Sınıflama işlemi sonunda sonuçlar sınıflama doğruluğu performansları olarak değerlendirilmiştir (https://datahub.io/search?q=diabet, Akdemir B.,2009).

4.1 Diyabet Hastalığı Verisi

Diyabet; pankreasın kan diyabetini düzenleyen bir hormon olan insülini yeterli miktarda üretememesi veya üretmiş olduğu insülinin kullanımında bozukluk sonucu kandaki diyabet düzeyinin yükselmesiyle gelişen bir hastalıktır (Koç ve Güler ,2015).

Diyabet hastalığı verimiz datahub internet adresinden elde edilmiştir. Veri setinin asıl kaynağı olarak UCI internet sitesi olarak gösterilmektedir. Veri setimizde 50 pozitif ve 50 negatif olmak üzere toplam 100 kişiye ait veriler bulunmaktadır. Veri setimiz Çizelge 4.1’de gösterilen genel özelliklere sahiptir.

Çizelge 4.1 Diyabet hastalığı veri seti genel özellikleri

Pima Hintlilerin Diyabet Hastalığı Veri seti	
Ana Sahibi	Ulusal Diyabet, Sindirim ve Böbrek Hastalıkları Enstitüsü
Veri Türü	Sağlık
Veri Tabanı Vericisi	Vincent Sigillito Araştırma Merkezi,RMI Grup Lideri Uygulamalı Fizik Laboratuvarı Johns Hopkins Üniversitesi Johns Hopkins Road Laurel, MD 20707 (301) 953-6231
Veri Tabanı Vericisi Mail	vgs@aplcen.apl.jhu.edu
Kullanım Amacı	Sınıflandırma
Veri Alınma Tarihi	9.04.1990
Örnek Sayısı	768
Özellik Sayısı	8
Özellik Türü	Sayısal
Eksik Değer	Yok
Son Özellik	Sınıflandırma
Sınıflandırma Dağılımı	1: Pozitif/ 0: Negatif

Veriye ait 8 özellik aşağıdaki gibidir:

1.Özellik: Hamile kalma sayısı

2.Özellik: Oral glukoz tolerans testi 2 saat plazma glikoz konsantrasyonu

3.Özellik: Diyastolik kan basıncı (mm Hg)

4.Özellik: Triceps deri kalınlığı (mm)

5.Özellik: 2 saatlik serum insülini (mu U/ml)

6.Özellik: Vücut kütle indeksi (kg/m^2)

7.Özellik: Diyabet hastalığı soyağacı durumu ()

8.Özellik: Yaş (yıl)

4.2 Göğüs Kanseri Hastalığı Verisi

Meme kanseri, meme dokularının hücrelerinde normal olmayan dönüşümler sonucu tümör adı verilen kitleler oluşturmasıdır. İyi huylu ve kötü huylu olmak üzere iki çeşit tümör mevcuttur ve meme kanseri, dünyadaki en tehlikeli hastalık olmak ile beraber en

yaygın kanser türlerinden biridir (Abdulkareem ve Kasapbaşlı,2020). Kadın hastalıklarında önde gelen ikinci kanser türüdür (Sun ve arkadaşları,2017)

Göğüs kanseri verimiz, UCI makine öğrenmesi veri bankasından elde edilmiştir. Veri setimizde 50 pozitif ve 50 negatif olmak üzere toplam 100 kişiye ait veriler bulunmaktadır ve Çizelge 4.2'de gösterilen genel özelliklere sahiptir.

Çizelge 4.2 Göğüs kanseri hastalığı veri seti genel özellikleri

Göğüs Kanseri Hastalığı Veri seti	
Veri Türü	Sağlık
Veri Tabanı Kaynak	Miguel Patrício(miguelpatricio '@' gmail.com) José Pereira (jafcpereira '@' gmail.com) Joana Crisóstomo (joanacrisostomo '@' hotmail.com) Paulo Matafome (paulomatafome '@' gmail.com)
Kullanım Amacı	Sınıflandırma
Veri Alınma Tarihi	Kasım,1998
Örnek Sayısı	116
Özellik Sayısı	10
Özellik Türü	Sayısal
Eksik Değer	Var
Son Özellik	Sınıflandırma
Sınıflandırma Dağılımı	Var/Yok

Veriye ait 10 özellik aşağıdaki gibidir:

1.Özellik: Yaş (yıl)

2.Özellik: BMI (kg/m^2)

3.Özellik: Glukoz (mg/dL)

4.Özellik: İnsülin (µU/mL)

5.Özellik: HOMA

6.Özellik: Leptin (ng/mL)

7.Özellik: Adiponektin (µg/mL)

8.Özellik: Resistin (ng/mL)

9.Özellik: MCP-1(pg/dL)

10.Özellik: Sınıflandırma (sağlıklı/hasta)

4.3 Karaciğer Hastalığı Verisi

Karaciğer vücudumuzdaki üçgen şeklindeki en büyük organ olup vücuttaki glikoz, yağ, vitamin, hormon, …v.b. birçok kimyasalın dengelenmesi görevini yerine getirmektedir. Karaciğer hastalığının erken teşhisinde hayatta kalma olasılığı artacaktır (Muthuselvan S.ve arkadaşları, 2018).

Karaciğer verilerimiz UCI internet adresinden elde edilmiştir. Veri setimizde 50 pozitif ve 50 negatif olmak üzere toplam 100 kişiye ait veriler bulunmaktadır ve verimiz Çizelge 4.3'te ki genel özelliklere sahiptir.

Çizelge 4.3 Karaciğer hastalığı veri seti genel özellikleri

Karaciğer hastalığı Veri seti	
Veri Türü	Sağlık
Veri Tabanı Kaynak-1	Bendi Venkata Ramana-ramana.bendi '@' gmail.com Bilgi Teknolojileri Bölümü-Aditya Teknoloji ve Yönetim Enstitüsü. Tekkali.532201-Andhra Pradesh-Hindistan
Veri Tabanı Kaynak-2	Prof. M.Surendra Prasad Babu - drmsprasadbabu '@' yahoo.co.in Bilgisayar Bilimi ve Sistem Mühendisliği Bölümü, Andhra Üniversitesi Mühendislik Fakültesi
Veri Tabanı Kaynak-3	Prof. N. B. Venkateswarlu venkat_ritch '@' yahoo.com Bilgi Teknolojileri Bölümü-Aditya Teknoloji ve Yönetim Enstitüsü. Tekkali.532201-Andhra Pradesh-Hindistan
Kullanım Amacı	Sınıflandırma
Veri Alınma Tarihi	21.05.2012
Örnek Sayısı	583
Özellik Sayısı	10
Özellik Türü	Sayısal
Eksik Değer	Var
Son Özellik	Sınıflandırma
Sınıflandırma Dağılımı	Var/Yok

Bu 10 özellikten 7 tanesi kullanılmaktadır. Veriye ait 7 özellik aşağıdaki gibidir:

1.Özellik: Yaş (Yıl)

2.Özellik: Toplam bilirubin miktarı

3.Özellik: Suda çözünebilen bilirubin

4.Özellik: Alkalen fosfataz

5.Özellik: Alanin Aminotransferaz

6.Özellik: Aspartat Aminotransferaz

7.Özellik: Albümin ve globülin oranı

4.4 Kalp Hastalığı

Kalp iki fonksiyona sahip kaslı pompa şeklindeki bir organımızdır. Birinci görevi, vücudun dokularından kanı toplayıp ve onu akciğerlere iletmek. İkincisi ise, onu akciğerlerden alıp vücudun bütün dokusuna iletmek şeklindedir (Weinhaus A. J. ve Roberts K. P).

Kalp hastalığı veri setimiz UCI makine öğrenmesi bankası kalp veri seti tabanından alınmıştır. Veri setimizde 50 pozitif ve 50 negatif olmak üzere toplam 100 kişiye ait veriler bulunmaktadır ve verimiz Çizelge 4.4'te verilen genel özelliklere sahiptir.

Çizelge 4.4 Kalp hastalığı veri seti genel özellikleri

Kalp Hastalığı Veri seti	
Ver Türü	Sağlık
Kullanım Amacı	Sınıflandırma
Veri Alınma Tarihi	9.04.1990
Örnek Sayısı	270
Özellik Sayısı	13
Özellik Türü	Sayısal
Eksik Değer	Yok
Son Özellik	Sınıflandırma
Sınıflandırma Dağılımı	1: var / 0: yok

Veriye ait 13 özellik aşağıdaki gibidir:

1.Özellik: Yaş (yıl)

2.Özellik: Cinsiyet (kadın/erkek)

3.Özellik: Göğüs ağrısı tipi (1 ile 4 arası)

4.Özellik: Dinlenme durumunda kan basıncı (tansiyon))

5.Özellik: Serum kolesterol (mg/dl)

6.Özellik: Tokluk Diyabet düzeyi >120 mg/dl (1=doğru/0=yanlış)

7.Özellik: Dinlenme halinde Elektrokardiyografı düzeyi (0,1,2)

8.Özellik: Maksimum kalp atış değeri(sürekli)

9.Özellik: Egzersiz durumunda göğüs ağrısı (0=hayır/1=evet)

10.Özellik: Dinlenme halinde ST değeri (sürekli)

11.Özellik: Pik egzersiz halinde ST segmentinin eğimi (1-2)

12.Özellik: Büyük damarların sayısı (0-3)

13.Özellik: Hasar oranı (3=normal,6=kalıcı,7=geri düzeltile bilinen hasar)

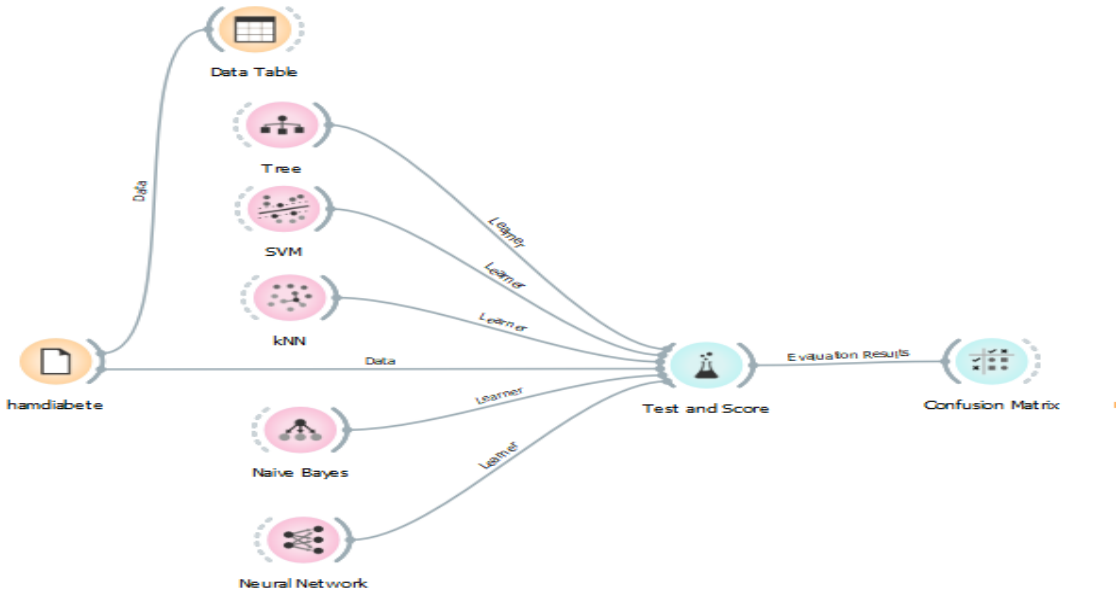
5. SONUÇLAR VE ÖNERİLER

Bu bölümde Pima Hintlilerinin diyabet hastalığı verisi, göğüs kanseri hastalığı verisi, karaciğer hastalığı verisi ve kalp hastalığı verisi sınıflandırma işlemlerine tabi tutulmuş ve sonuçları 4 farklı k-kat çaprazlamada (2,5,10,20) sınıflama doğruluğu değerlendirilmiş ve karşılaştırılmıştır. Değerlendirme sonucunda önerilerde bulunulmuştur.

5.1 Diyabet Hastalığı Sınıflandırma Performans Sonuçları

Diyabet hastalığının sınıflandırma performansını değerlendirmek için 50 sağlıklı ve 50 hasta verisi olmak üzere toplam 100 kişiden değerler alınmıştır. Sınıflandırma işlemi ham verilere, minimum maksimum normalizasyon yöntemi uygulanmış verilere, ondalık ölçekleme normalizasyon yöntemi uygulanmış verilere, z-skor normalizasyon yöntemi uygulanmış verilere ve norm normalizasyon yöntemi uygulanmış verilere olmak üzere 5 farklı durum için değerlendirilmiştir. Diyabet hastalığı verisinin ham veri ve normalize edilmiş durumu DVM, YSA, k-NN, KA ve Naive Bayes gibi sınıflandırma yöntemleri ile 4 farklı k-kat çaprazlamada (2,5,10,20) sınıflama doğruluğu değerlendirilmiştir.

Şekil 5.1'de diyabet hastalığı ham veri setinin ORANGE programında KA, DVM, k-NN, Naive Bayes ve YSA gibi sınıflandırma algoritmaları kullanılarak sınıflandırma işlemine tabi tutulması gösterilmiştir.



Şekil 5.1 Diyabet hastalığı ham verisinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Diyabet hastalığı ham verisine 5 faklı sınıflandırma yöntemi kullanılarak yapılan sınıflandırma işlemlerinin sonucu 2,5,10 ve 20-kat çaprazlamada elde edilen sınıflama doğruluğu Çizelge 5.1'de gösterilmiştir.

Çizelge 5.1 Ham diyabet hastalığı veri setinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Sınıflandırma Yöntemi	Sınıflama doğruluğu (%)				
	k-kat çaprazlama				
	2	**5**	**10**	**20**	**Ortalama**
k-NN	**62**	57	59	57	58.75
KA	66	68	65	**72**	67.75
DVM	67	**69**	**69**	**69**	**68.5**
YSA	62	**71**	65	66	66
Naive Bayes	65	68	**69**	**69**	67.75
Ortalama	64.4	**66.6**	65.4	**66.6**	

Çizelge 5.1'e göre;

k-NN sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 2-kat çaprazlamada % 62 olarak bulunmuştur.

KA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 20-kat çaprazlamada % 72 olarak elde edilmiştir.

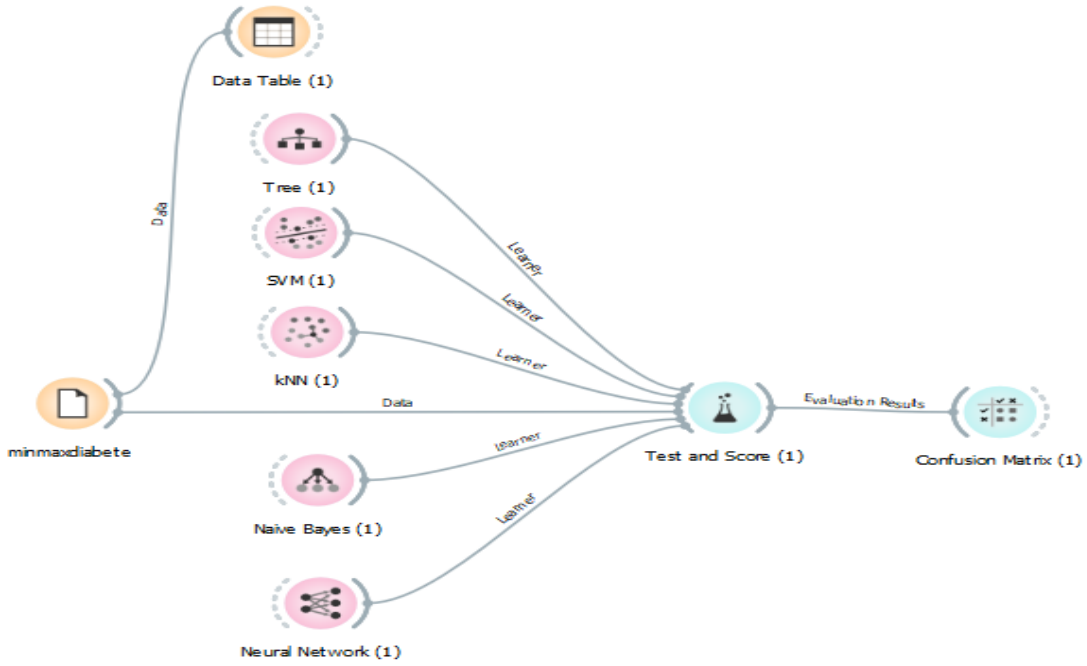
DVM sınıflandırma yönteminde, 2-kat çaprazlama kriteri hariç diğer çaprazlama değerlerinde % 69 sınıflama doğruluğu elde edilmiştir. 5 sınıflandırma yöntemi içinde ortalamalara bakıldığında ise % 68.5 sınıflama doğruluğuyla en yüksek sonuca DVM sınıflandırma yönteminde ulaşılmıştır.

YSA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 5-kat çaprazlamada % 71 olarak bulunmuştur.

Naive Bayes sınıflandırma yöntemindeyse, 10 ve 20-kat çaprazlamada % 69 olarak en yüksek sınıflama doğruluğu elde edilmiştir.

k-kat çaprazlamanın diyabet hastalığı ham verilerine etkisine bakıldığında; en yüksek sınıflama doğruluğu 5 ve 10-kat çaprazlamada ortalama % 66.6 olmuştur.

Şekil 5.2'de diyabet hastalığı veri setinin minimum maksimum normalizasyon yöntemi uygulanması sonrası ORANGE programında KA, DVM, k-NN, Naive Bayes ve YSA sınıflandırma algoritmaları kullanılarak sınıflandırma işlemine tabi tutulması gösterilmiştir.



Şekil 5.2 Diyabet hastalığı minimum maksimum normalizasyon yöntemi uygulanmış verisinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Diyabet hastalığı verisine minimum maksimum normalizasyon yöntemiyle 5 farklı sınıflandırma yöntemi kullanılarak yapılan sınıflandırma işlemlerinin sonucu 2,5,10 ve 20-kat çaprazlamada elde edilen sınıflama doğruluğu Çizelge 5.2'de gösterilmiştir.

Çizelge 5.2 Minimum maksimum normalizasyon yöntemi uygulanmış diyabet hastalığı veri setinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Sınıflandırma Yöntemi	Sınıflama doğruluğu (%)				
	Çaprazlama Sayısı				
	2	**5**	**10**	**20**	**Ortalama**
k-NN	60	**69**	67	68	66
KA	66	67	65	**71**	67.25
DVM	67	**69**	**69**	**69**	**68.5**
YSA	62	**71**	65	66	66
Naive Bayes	65	68	**69**	**69**	67.75
Ortalama	64	68.8	**67**	**68.6**	

Çizelge 5.2'ye göre;

k-NN sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 5-kat çaprazlamada % 69 olarak bulunmuştur.

KA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 20-kat çaprazlamada % 71 olarak elde edilmiştir.

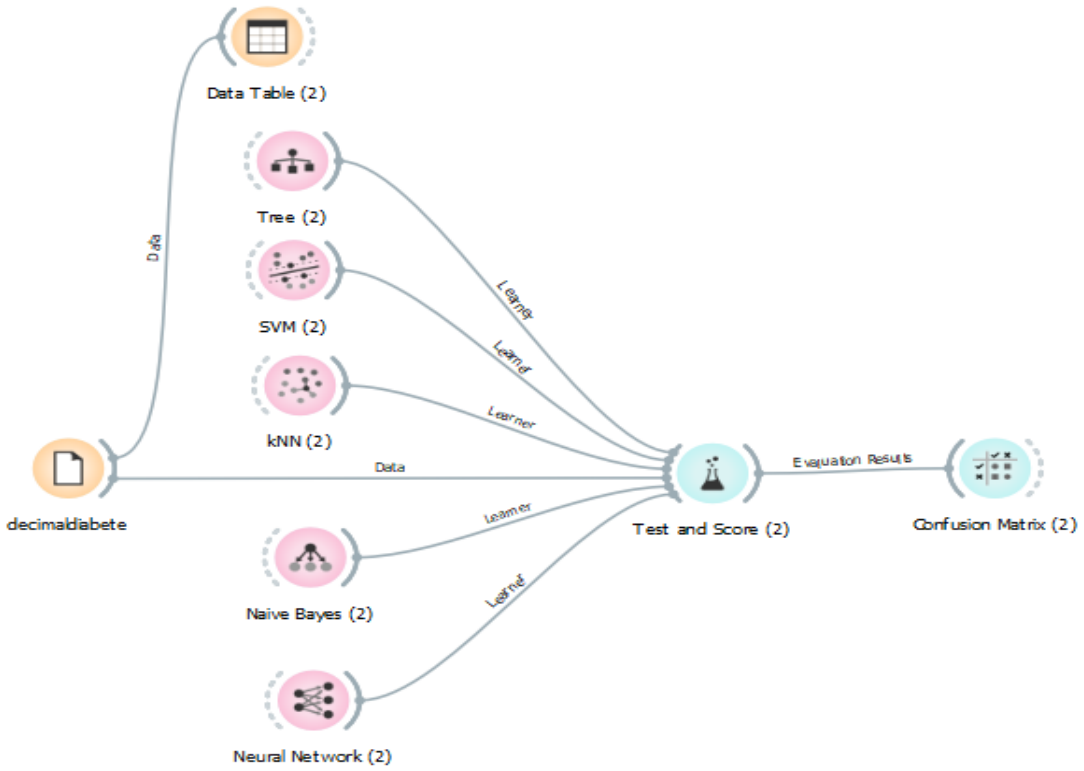
DVM sınıflandırma yönteminde, 2-kat çaprazlama kriteri hariç diğer çaprazlama değerlerinde % 69 sınıflama doğruluğu elde edilmiştir. 5 sınıflandırma yöntemi içinde ortalamalara bakıldığında ise % 68.5 sınıflama doğruluğuyla en yüksek sonuca DVM sınıflandırma yönteminde ulaşılmıştır.

YSA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 5-kat çaprazlamada % 71 olarak bulunmuştur.

Naive Bayes sınıflandırma yöntemindeyse, 10 ve 20-kat çaprazlamada % 69 olarak en yüksek sınıflama doğruluğu elde edilmiştir.

k-kat çaprazlamanın diyabet hastalığı minimum maksimum normalizasyon verilerine etkisine bakıldığında; en yüksek sınıflama doğruluğu 5-kat çaprazlamada ortalama % 68.8 olmuştur.

Şekil 5.3'te diyabet hastalığı veri setinin ondalık ölçekleme normalizasyon yöntemi uygulanması sonrası ORANGE programında KA, DVM, k-NN, Naive Bayes ve YSA sınıflandırma algoritmaları kullanılarak sınıflandırma işlemine tabi tutulması gösterilmiştir.



Şekil 5.3 Diyabet hastalığı ondalık ölçekleme normalizasyon yöntemi uygulanmış verisinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Diyabet hastalığı verisine ondalık ölçekleme normalizasyon yöntemiyle 5 farklı sınıflandırma yöntemi kullanılarak yapılan sınıflandırma işlemlerinin sonucu 2,5,10 ve 20-kat çaprazlamada elde edilen sınıflama doğruluğu Çizelge 5.3'de gösterilmiştir.

Çizelge 5.3 Ondalık ölçekleme normalizasyon yöntemi uygulanmış diyabet hastalığı veri setinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Sınıflandırma Yöntemi	Sınıflama Doğruluğu (%)				
	Çaprazlama Sayısı				
	2	**5**	**10**	**20**	**Ortalama**
k-NN	59	61	**63**	60	60.75
KA	66	67	67	**70**	67.5
DVM	67	**69**	**69**	**69**	**68.5**
YSA	63	**71**	65	65	66
Naive Bayes	65	**70**	69	**70**	**68.5**
Ortalama	64	**67.6**	66.6	66.8	

Çizelge 5.3'e göre;

k-NN sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 63 olarak bulunmuştur.

KA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 20-kat çaprazlamada % 70 olarak elde edilmiştir.

DVM sınıflandırma yönteminde, 2-kat çaprazlama kriteri hariç diğer çaprazlama değerlerinde % 69 sınıflama doğruluğu elde edilmiştir.

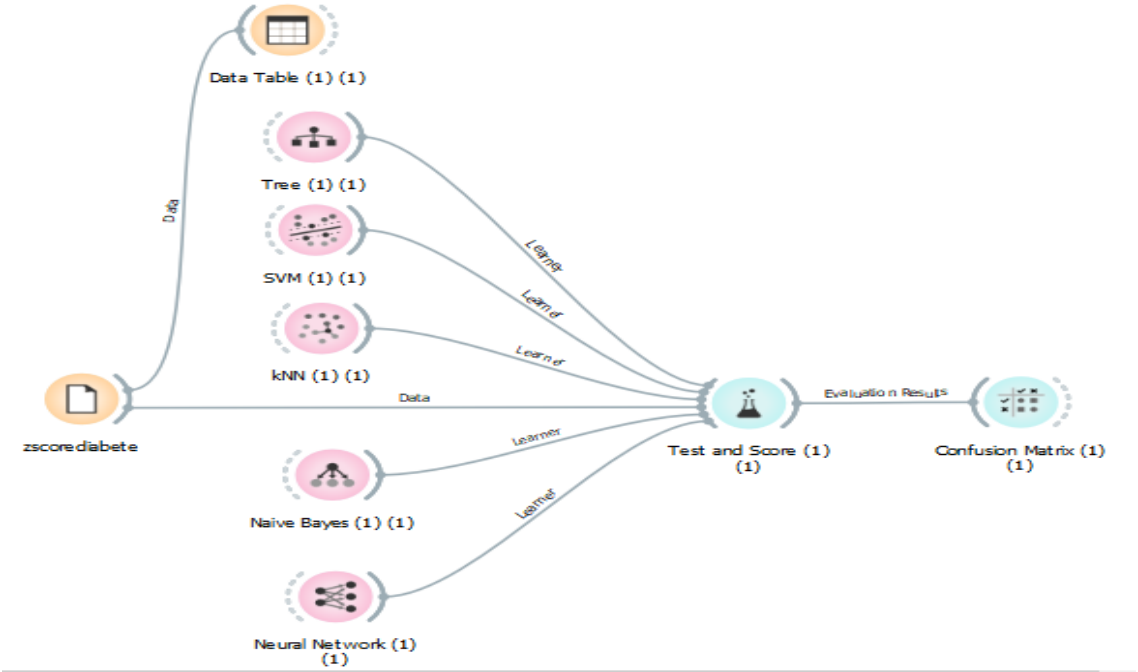
YSA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 5-kat çaprazlamada % 71 olarak bulunmuştur.

Naive Bayes sınıflandırma yöntemindeyse, 5 ve 20-kat çaprazlamada % 70 olarak en yüksek sınıflama doğruluğu elde edilmiştir.

5 sınıflandırma yöntemi içinde ortalamalara bakıldığında ise % 68.5 sınıflama doğruluğuyla en yüksek sonuca DVM ve Naive Bayes sınıflandırma yöntemlerinde ulaşılmıştır.

k-kat çaprazlamanın diyabet hastalığı ondalık ölçekleme normalizasyon verilerine etkisine bakıldığında; en yüksek sınıflama doğruluğu 5-kat çaprazlamada ortalama % 67.6 olmuştur.

Şekil 5.4'te diyabet hastalığı veri setinin z-skor normalizasyon yöntemi uygulanması sonrası ORANGE programında KA, DVM, k-NN, Naive Bayes ve YSA sınıflandırma algoritmaları kullanılarak sınıflandırma işlemine tabi tutulması gösterilmiştir.



Şekil 5.4 Diyabet hastalığı z-skor normalizasyon yöntemi uygulanmış verisinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Diyabet hastalığı verisine z-skor normalizasyon yöntemiyle 5 farklı sınıflandırma yöntemi kullanılarak yapılan sınıflandırma işlemlerinin sonucu 2,5,10 ve 20-kat çaprazlamada elde edilen sınıflama doğruluğu Çizelge 5.4'de gösterilmiştir.

Çizelge 5.4 Z-skor ölçekleme normalizasyon yöntemi uygulanmış diyabet hastalığı veri setinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Sınıflandırma Yöntemi	Sınıflama Doğruluğu (%)				
	Çaprazlama Sayısı				
	2	**5**	**10**	**20**	**Ortalama**
k-NN	64	**72**	71	69	**69**
KA	66	68	65	**72**	67.75
DVM	67	**69**	**69**	**69**	68.5
YSA	62	**71**	65	66	66
Naive Bayes	65	68	**69**	**69**	67.75
Ortalama	64.8	**69,6**	67.8	69	

Çizelge 5.4'e göre;

k-NN sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 5-kat çaprazlamada % 72 olarak bulunmuştur. 5 sınıflandırma yöntemi içinde ortalamalara bakıldığında ise % 69 sınıflama doğruluğuyla en yüksek sonuca ulaşılmıştır.

KA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 20-kat çaprazlamada % 72 olarak elde edilmiştir.

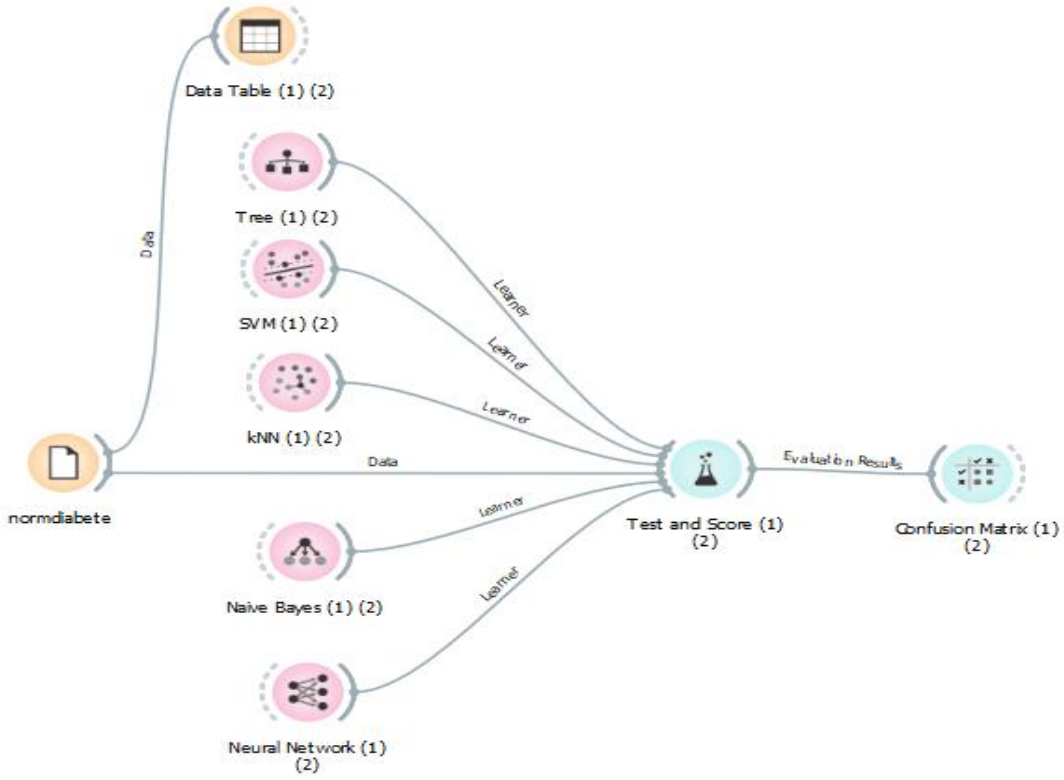
DVM sınıflandırma yönteminde, 2-kat çaprazlama kriteri hariç diğer çaprazlama değerlerinde % 69 sınıflama doğruluğu elde edilmiştir.

YSA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 5-kat çaprazlamada % 71 olarak bulunmuştur.

Naive Bayes sınıflandırma yöntemindeyse, 10 ve 20-kat çaprazlamada % 69 olarak en yüksek sınıflama doğruluğu elde edilmiştir.

k-kat çaprazlamanın diyabet hastalığı z-skor normalizasyon verilerine etkisine bakıldığında; en yüksek sınıflama doğruluğu 5-kat çaprazlamada ortalama % 69.6 olmuştur.

Şekil 5.5'te diyabet hastalığı veri setinin norm normalizasyon yöntemi uygulanması sonrası ORANGE programında KA, DVM, k-NN, Naive Bayes ve YSA sınıflandırma algoritmaları kullanılarak sınıflandırma işlemine tabi tutulması gösterilmiştir.



Şekil 5.5 Diyabet hastalığı norm normalizasyon yöntemi uygulanmış verisinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Diyabet hastalığı verisine norm normalizasyon yöntemiyle 5 farklı sınıflandırma yöntemi kullanılarak yapılan sınıflandırma işlemlerinin sonucu 2,5,10 ve 20-kat çaprazlamada elde edilen sınıflama doğruluğu Çizelge 5.5'de gösterilmiştir.

Çizelge 5.5 Norm ölçekleme normalizasyon yöntemi uygulanmış diyabet hastalığı veri setinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Sınıflandırma Yöntemi	Sınıflama Doğruluğu (%)				
	Çaprazlama Sayısı				
	2	**5**	**10**	**20**	**Ortalama**
k-NN	56	**63**	**63**	**63**	61.25
KA	65	67	66	**71**	67.25
DVM	66	**69**	**69**	**69**	**68.25**
YSA	63	**70**	66	67	66.5
Naive Bayes	65	70	**71**	65	67.75
Ortalama	63	**67.8**	67	67	

Çizelge 5.5'e göre;

k-NN sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 5,10 ve 20-kat çaprazlamada % 63 olarak bulunmuştur.

KA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 20-kat çaprazlamada % 71 olarak elde edilmiştir.

DVM sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 5,10 ve 20-kat çaprazlamada % 69 sınıflama doğruluğu elde edilmiştir. . 5 sınıflandırma yöntemi içinde ortalamalara bakıldığında ise % 68.25 sınıflama doğruluğuyla en yüksek sonuca ulaşılmıştır.

YSA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 5-kat çaprazlamada % 70 olarak bulunmuştur.

Naive Bayes sınıflandırma yöntemindeyse, 10-kat çaprazlamada % 71 olarak en yüksek sınıflama doğruluğu elde edilmiştir

k-kat çaprazlamanın diyabet hastalığı norm normalizasyon verilerine etkisine bakıldığında; en yüksek sınıflama doğruluğu 5-kat çaprazlamada ortalama % 67.8 olmuştur.

Normalizasyon yöntemlerinin sınıflandırma performansına etkisini değerlendirmek için ayrı ayrı değerlendirilen k-kat çaprazlamaların sınıflama doğruluklarının ortalamaları Çizelge 5.6'da gösterilmiştir.

Çizelge 5.6 Normalizasyon yöntemlerinin diyabet hastalığı veri setinin sınıflandırma performansına etkisinin karşılaştırması

Sınıflandırma Yöntemleri	**Ham Veri Sınıflama Doğruluğu (%)**	**Sınıflama Doğruluğu (%)**			
		Minimum Maksimum	**Ondalık Ölçekleme**	**Z-skor**	**Norm Yöntemi**
k-NN	58.75	66	60.75	**69**	61.25
KA	67.75	67.25	67.5	67.75	67.25
DVM	**68.5**	**68.5**	**68.5**	68.5	68.25
YSA	66	66	66	66	**66.5**
Naive Bayes	67.75	67.75	**68.5**	67.75	67.75
Ortalama	65.75	67.1	66,25	**67.8**	66.2

Çizelge 5.6'ya göre,

k-NN sınıflandırma yönteminde normalizasyon yöntemlerinin sınıflama performansına doğrudan etkisi olmuştur. En iyi performans artışı z-skor normalizasyon yönteminde % 69 sınıflama doğruluğu elde edilmiştir.

KA sınıflandırma yönteminde normalizasyon yöntemlerinin sınıflama performansına pek bir etkisinin olmadığı görülmüştür. Sadece z-skor normalizasyon yönteminde sınıflama doğruluğu % 67.75 olarak ham verideki sınıflama doğruluğuna ulaşabilmiştir.

DVM sınıflandırma yönteminde normalizasyon yöntemlerinin performansını artırmadığı hatta norm normalizasyon yöntemi sonrası % 68.25 ile azalttığı görülmüştür.

YSA sınıflandırma yönteminde sadece norm normalizasyon yönteminde sınıflama doğruluğu az da olsa arttığı (% 66.5) görüldü. Diğer normalizasyon yöntemlerinde performansı değişmemiştir.

Naive Bayes sınıflandırma yönteminde sadece ondalık ölçekleme normalizasyon yönteminde sınıflama doğruluğunu arttırdığı (% 68,5) görülmüştür. Diğer normalizasyon yöntemlerinin bir etkisinin olmadığı görülmüştür.

Sonuç olarak normalizasyon yöntemlerinin diyabet hastalığı verilerinin sınıflandırma performansına pek az bir etkisinin olduğu görülmüştür.

Diyabet hastalığı ham verisine ve 4 faklı normalizasyon yöntemlerine k-kat çaprazlamanın etkisini görmek için Çizelge 5.1'den Çizelge 5.5'e kadar ortalama sınıflama doğrulukları alınarak Çizelge 5.7'de toplu olarak gösterilmiştir.

Çizelge 5.7 Diyabet hastalığı verilerine k-kat çaprazlamanın etkisinin değerlendirilmesi

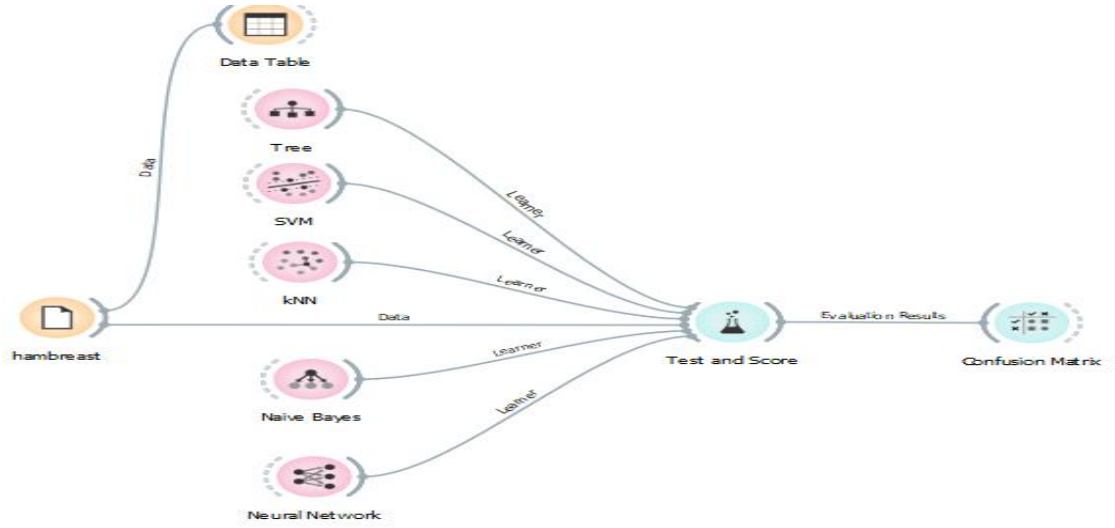
Diyabet hastalığı veri tipi	**Ortalama sınıflama doğruluğu (%)**			
	k-kat çaprazlama			
	2	**5**	**10**	**20**
Ham veri	64.4	**66.6**	65.4	**66.6**
Minimum maksimum normalizasyon	64	**68.8**	67	68.6
Ondalık ölçekleme normalizasyon	64	**67.6**	66.6	66.8
Z-skor normalizasyon	64.8	**69.6**	67.8	69
Norm normalizasyon	63	**67.8**	67	67

Çizelge 5.7'ye göre, diyabet hastalığı veri seti doğru k-kat çaprazlama seçiminin sınıflama doğruluklarına bakıldığında en yüksek sınıflama doğruluğu 5-kat çaprazlamada olmuştur.

5.2 Göğüs Kanseri Hastalığı Sınıflandırma Performans Sonuçları

Göğüs kanseri hastalığının sınıflandırma performansını değerlendirmek için 50 sağlıklı ve 50 hasta verisi olmak üzere toplam 100 kişiden değerler alınmıştır. Sınıflandırma işlemi ham verilere, minimum maksimum normalizasyon yöntemi uygulanmış verilere, ondalık ölçekleme normalizasyon yöntemi uygulanmış verilere, z-skor normalizasyon yöntemi uygulanmış verilere ve norm normalizasyon yöntemi uygulanmış verilere olmak üzere 5 farklı durum için değerlendirilmiştir. Göğüs kanseri hastalığı verisinin ham veri ve normalize edilmiş durumu DVM, YSA, k-NN, KA ve Naive Bayes gibi sınıflandırma yöntemleri ile 4 farklı k-kat çaprazlamada (2,5,10,20) sınıflama doğruluğu değerlendirilmiştir.

Şekil 5.7'de göğüs kanseri hastalığı ham veri setinin ORANGE programında KA, DVM, k-NN, Naive Bayes ve YSA gibi sınıflandırma algoritmaları kullanılarak sınıflandırma işlemine tabi tutulması gösterilmiştir.



Şekil 5.6 Göğüs kanseri hastalığı ham verisinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Göğüs kanseri hastalığı ham verisine 5 faklı sınıflandırma yöntemi kullanılarak yapılan sınıflandırma işlemlerinin sonucu 2,5,10 ve 20-kat çaprazlamada elde edilen sınıflama doğruluğu Çizelge 5.8'de gösterilmiştir.

Çizelge 5.8 Ham göğüs kanseri hastalığı veri setinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Sınıflandırma Yöntemi	Sınıflama doğruluğu (%)				
	k-kat çaprazlama				
	2	**5**	**10**	**20**	**Ortalama**
k-NN	**57**	48	54	**57**	54
KA	72	72	**82**	81	76.75
DVM	81	81	83	**85**	**82.5**
YSA	80	**82**	81	80	80.75
Naive Bayes	73	72	**76**	73	73.5
Ortalama	72.6	71	**75.2**	**75.2**	

Çizelge 5.8'e göre;

k-NN sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 2 ve 20-kat çaprazlamada % 57 olarak bulunmuştur.

KA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 82 olarak elde edilmiştir.

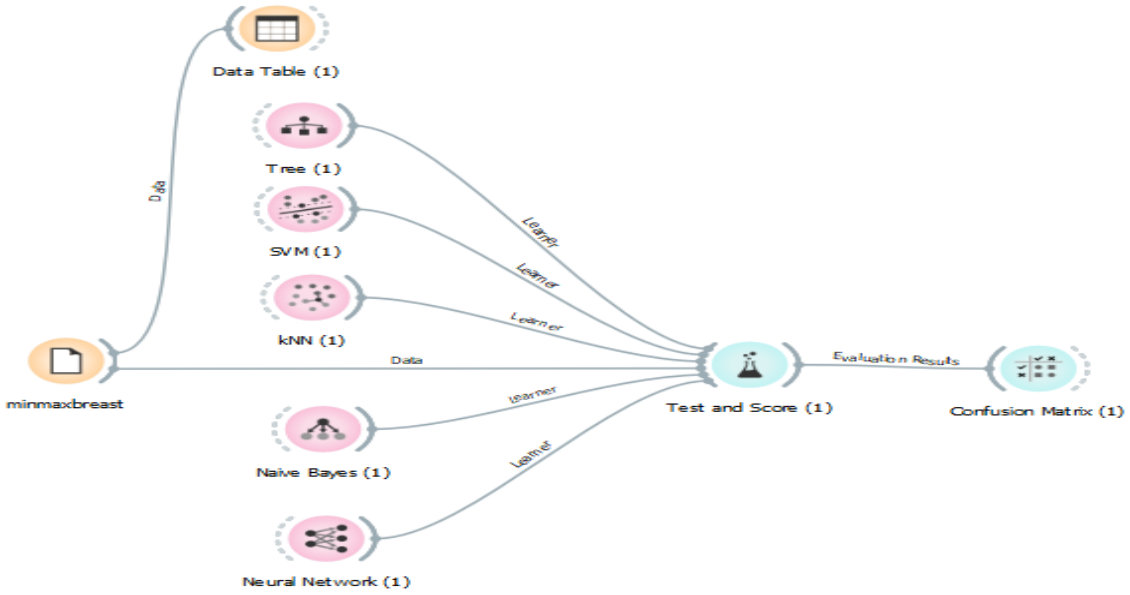
DVM sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 20-kat çaprazlamada % 85 olarak elde edilmiştir. 5 sınıflandırma yöntemi içinde ortalamalara bakıldığında ise % 82.5 sınıflama doğruluğuyla en yüksek sonuca DVM sınıflandırma yönteminde ulaşılmıştır.

YSA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 5-kat çaprazlamada % 82 olarak bulunmuştur.

Naive Bayes sınıflandırma yöntemindeyse, 10-kat çaprazlamada % 76 olarak en yüksek sınıflama doğruluğu elde edilmiştir.

k-kat çaprazlamanın göğüs kanseri hastalığı ham verilerine etkisine bakıldığında; en yüksek sınıflama doğruluğu 10 ve 20-kat çaprazlamada ortalama % 75.2 olmuştur.

Şekil 5.7'de Göğüs kanseri hastalığı veri setinin minimum maksimum normalizasyon yöntemi uygulanması sonrası ORANGE programında KA, DVM, k-NN, Naive Bayes ve YSA sınıflandırma algoritmaları kullanılarak sınıflandırma işlemine tabi tutulması gösterilmiştir.



Şekil 5.7 Göğüs kanseri hastalığı minimum maksimum normalizasyon yöntemi uygulanmış verisinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Göğüs kanseri hastalığı verisine minimum maksimum normalizasyon yöntemiyle 5 farklı sınıflandırma yöntemi kullanılarak yapılan sınıflandırma işlemlerinin sonucu 2,5,10 ve 20-kat çaprazlamada elde edilen sınıflama doğruluğu Çizelge 5.9'da gösterilmiştir.

Çizelge 5.9 Minimum maksimum normalizasyon yöntemi uygulanmış göğüs kanseri hastalığı veri setinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Sınıflandırma Yöntemi	Sınıflama doğruluğu (%)				
	k-kat çaprazlama				
	2	**5**	**10**	**20**	**Ortalama**
k-NN	**78**	77	**78**	**78**	77.75
KA	72	72	**84**	81	77.25
DVM	81	81	82	**84**	**82**
YSA	80	**82**	81	80	80.75
Naive Bayes	**75**	74	74	73	74
Ortalama	77.2	77.2	**79.8**	79.2	

Çizelge 5.9'e göre;

k-NN sınıflandırma yönteminde, 5-kat çaprazlama kriteri hariç diğer çaprazlama değerlerinde % 78 sınıflama doğruluğu elde edilmiştir.

KA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 84 olarak elde edilmiştir.

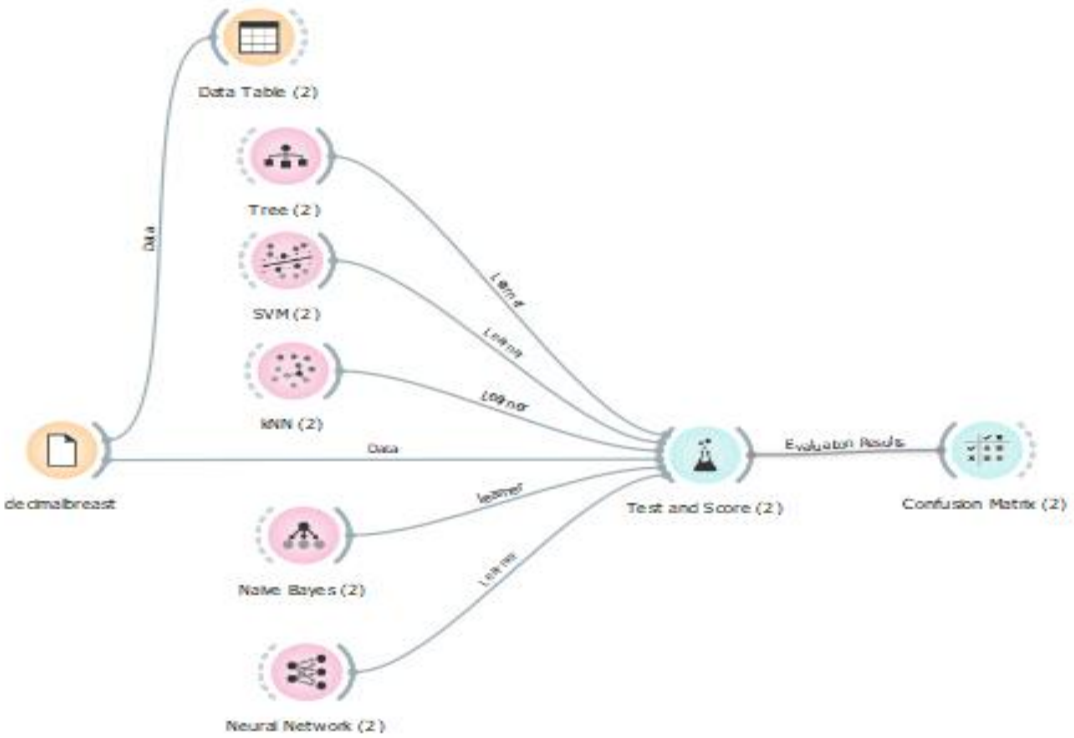
DVM sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 20-kat çaprazlamada % 84 olarak elde edilmiştir. 5 sınıflandırma yöntemi içinde ortalamalara bakıldığında ise % 82 sınıflama doğruluğuyla en yüksek sonuca DVM sınıflandırma yönteminde ulaşılmıştır.

YSA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 5-kat çaprazlamada % 82 olarak bulunmuştur.

Naive Bayes sınıflandırma yöntemindeyse, 2-kat çaprazlamada % 75 olarak en yüksek sınıflama doğruluğu elde edilmiştir.

k-kat çaprazlamanın göğüs kanseri hastalığı minimum maksimum normalizasyon verilerine etkisine bakıldığında; en yüksek sınıflama doğruluğu 10-kat çaprazlamada ortalama % 79.8 olmuştur.

Şekil 5.8'de göğüs kanseri hastalığı veri setinin ondalık ölçekleme normalizasyon yöntemi uygulanması sonrası ORANGE programında KA, DVM, k-NN, Naive Bayes ve YSA sınıflandırma algoritmaları kullanılarak sınıflandırma işlemine tabi tutulması gösterilmiştir.



Şekil 5.8 Göğüs kanseri hastalığı ondalık ölçekleme normalizasyon yöntemi uygulanmış verisinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Göğüs kanseri hastalığı verisine ondalık ölçekleme normalizasyon yöntemiyle 5 farklı sınıflandırma yöntemi kullanılarak yapılan sınıflandırma işlemlerinin sonucu 2,5,10 ve 20-kat çaprazlamada elde edilen sınıflama doğruluğu Çizelge 5.10'da gösterilmiştir.

Çizelge 5.10 Ondalık ölçekleme normalizasyon yöntemi uygulanmış göğüs kanseri hastalığı veri setinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Sınıflandırma Yöntemi	Sınıflama doğruluğu (%)				
	k-kat çaprazlama				
	2	**5**	**10**	**20**	**Ortalama**
k-NN	74	74	**75**	**75**	74.5
KA	70	73	**81**	80	76
DVM	81	81	81	**84**	**81.75**
YSA	80	**82**	81	80	80.75
Naive Bayes	**74**	**74**	72	72	73
Ortalama	75.8	76.8	78	**78.2**	

Çizelge 5.10'a göre;

k-NN sınıflandırma yönteminde, 10 ve 20-kat çaprazlamada % 75 olarak en yüksek sınıflama doğruluğu elde edilmiştir.

KA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 81 olarak elde edilmiştir.

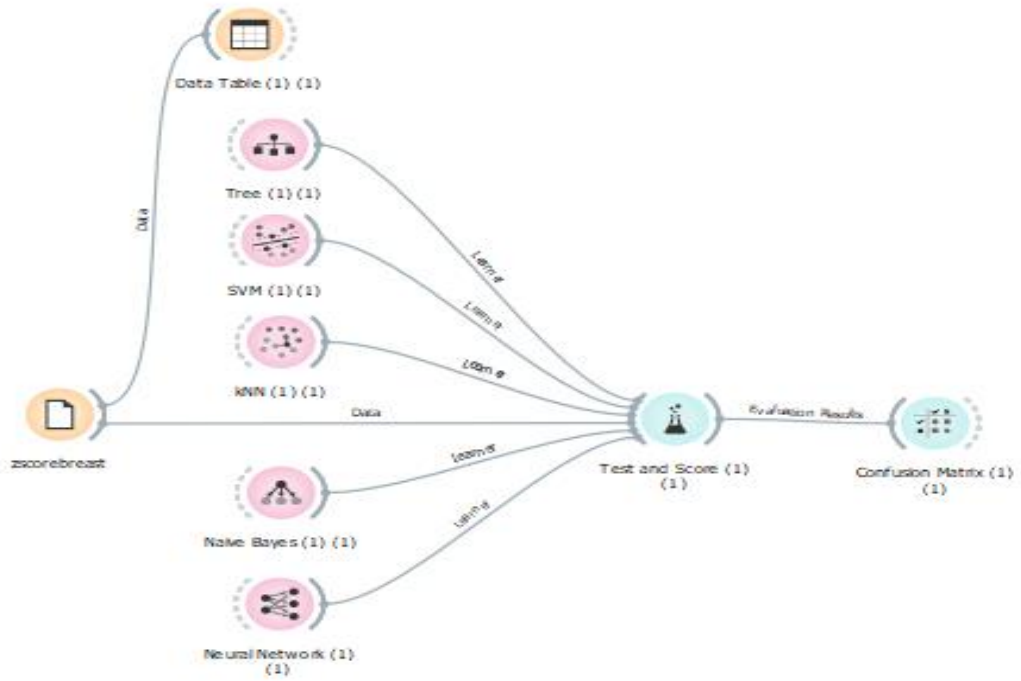
DVM sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 20-kat çaprazlamada % 84 olarak elde edilmiştir. 5 sınıflandırma yöntemi içinde ortalamalara bakıldığında ise % 81.75 sınıflama doğruluğuyla en yüksek sonuca DVM sınıflandırma yöntemi ile ulaşılmıştır.

YSA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 5-kat çaprazlamada % 82 olarak bulunmuştur.

Naive Bayes sınıflandırma yöntemindeyse, 2 ve 5-kat çaprazlamada % 74 olarak en yüksek sınıflama doğruluğu elde edilmiştir.

k-kat çaprazlamanın göğüs kanseri hastalığı ondalık ölçekleme verilerine etkisine bakıldığında; en yüksek sınıflama doğruluğu 20-kat çaprazlamada ortalama % 78.2 olmuştur.

Şekil 5.9'da göğüs kanseri hastalığı veri setinin z-skor normalizasyon yöntemi uygulanması sonrası ORANGE programında KA, DVM, k-NN, Naive Bayes ve YSA sınıflandırma algoritmaları kullanılarak sınıflandırma işlemine tabi tutulması gösterilmiştir.



Şekil 5.9 Göğüs kanseri hastalığı z-skor normalizasyon yöntemi uygulanmış verisinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Göğüs kanseri hastalığı verisine z-skor normalizasyon yöntemiyle 5 farklı sınıflandırma yöntemi kullanılarak yapılan sınıflandırma işlemlerinin sonucu 2,5,10 ve 20-kat çaprazlamada elde edilen sınıflama doğruluğu Çizelge 5.11'de gösterilmiştir.

Çizelge 5.11 Z-skor ölçekleme normalizasyon yöntemi uygulanmış göğüs kanseri hastalığı veri setinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Sınıflandırma Yöntemi	Sınıflama doğruluğu (%)				
	k-kat çaprazlama				
	2	**5**	**10**	**20**	**Ortalama**
k-NN	72	79	**81**	**81**	78.25
KA	72	72	**82**	81	76.75
DVM	81	81	83	**85**	**82.5**
YSA	80	**82**	81	80	80.75
Naive Bayes	74	72	**76**	73	73.75
Ortalama	75.8	77.2	**80.6**	80	

Çizelge 5.11'e göre;

k-NN sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10 ve 20-kat çaprazlamada % 81 olarak bulunmuştur.

KA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 82 olarak elde edilmiştir.

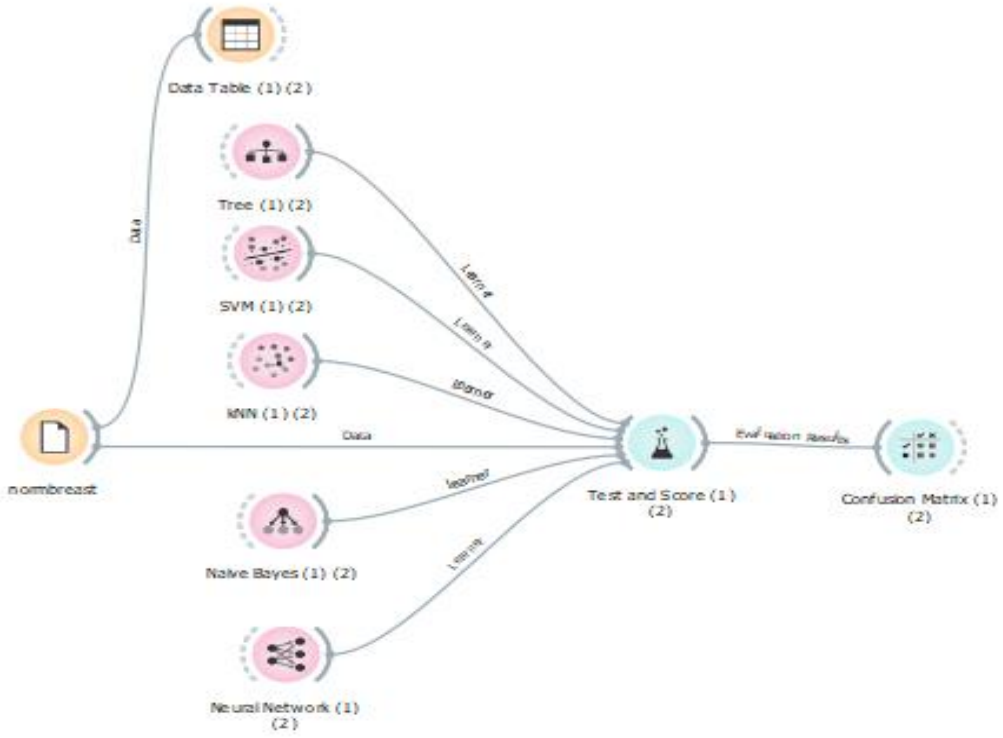
DVM sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 20-kat çaprazlamada % 85 olarak elde edilmiştir. 5 sınıflandırma yöntemi içinde ortalamalara bakıldığında ise % 82.5 sınıflama doğruluğuyla en yüksek sonuca ulaşılmıştır.

YSA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 5-kat çaprazlamada % 82 olarak bulunmuştur.

Naive Bayes sınıflandırma yöntemindeyse, 10-kat çaprazlamada % 76 olarak en yüksek sınıflama doğruluğu elde edilmiştir.

k-kat çaprazlamanın göğüs kanseri hastalığı z-skor normalizasyon verilerine etkisine bakıldığında; en yüksek sınıflama doğruluğu 10-kat çaprazlamada ortalama % 80.6 olmuştur.

Şekil 5.10'da göğüs kanseri hastalığı veri setinin norm normalizasyon yöntemi uygulanması sonrası ORANGE programında KA, DVM, k-NN, Naive Bayes ve YSA sınıflandırma algoritmaları kullanılarak sınıflandırma işlemine tabi tutulması gösterilmiştir.



Şekil 5.10 Göğüs kanseri hastalığı norm normalizasyon yöntemi uygulanmış verisinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Göğüs kanseri hastalığı verisine norm normalizasyon yöntemiyle 5 farklı sınıflandırma yöntemi kullanılarak yapılan sınıflandırma işlemlerinin sonucu 2,5,10 ve 20-kat çaprazlamada elde edilen sınıflama doğruluğu Çizelge 5.12'de gösterilmiştir.

Çizelge 5.12 Norm ölçekleme normalizasyon yöntemi uygulanmış göğüs kanseri hastalığı veri setinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Sınıflandırma Yöntemi	Sınıflama doğruluğu (%)				
	k-kat çaprazlama				
	2	**5**	**10**	**20**	**Ortalama**
k-NN	**71**	58	58	66	63.25
KA	72	72	**82**	81	76.75
DVM	81	81	83	**85**	**82.5**
YSA	80	**82**	81	80	80.75
Naive Bayes	73	72	**76**	73	74.25
Ortalama	75.4	73	76	**77**	

Çizelge 5.12'ye göre;

k-NN sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 2-kat çaprazlamada % 71 olarak bulunmuştur.

KA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 82 olarak elde edilmiştir.

DVM sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 20-kat çaprazlamada % 85 sınıflama doğruluğu elde edilmiştir. 5 sınıflandırma yöntemi içinde ortalamalara bakıldığında ise % 82.5 sınıflama doğruluğuyla en yüksek sonuca ulaşılmıştır.

YSA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 5-kat çaprazlamada % 82 olarak bulunmuştur.

Naive Bayes sınıflandırma yöntemindeyse, 10-kat çaprazlamada % 76 olarak en yüksek sınıflama doğruluğu elde edilmiştir.

k-kat çaprazlamanın göğüs kanseri hastalığı norm normalizasyon verilerine etkisine bakıldığında; en yüksek sınıflama doğruluğu 20-kat çaprazlamada ortalama % 77 olmuştur.

Normalizasyon yöntemlerinin sınıflandırma performansına etkisini değerlendirmek için ayrı ayrı değerlendirilen k-kat çaprazlamaların sınıflama doğruluklarının ortalamaları Çizelge 5.13'te gösterilmiştir.

Çizelge 5.13 Normalizasyon yöntemlerinin göğüs kanseri hastalığı veri setinin sınıflandırma performansına etkisinin karşılaştırması

Sınıflandırma Yöntemleri	**Ham Veri Sınıflama Doğruluğu (%)**	**Sınıflama Doğruluğu (%)**			
		Minimum Maksimum	**Ondalık Ölçekleme**	**Z-skor**	**Norm Yöntemi**
k-NN	54	77.75	74.5	**78.25**	63.25
KA	76.75	**77.25**	76	76.75	76.75
DVM	**82.5**	82	81.75	**82.5**	**82.5**
YSA	80.75	80.75	80.75	80.75	80.75
Naive Bayes	73.5	74	73	73.75	**74.25**
Ortalama	73.5	78.1	77.2	**78.4**	75.5

Çizelge 5.13'e göre,

k-NN sınıflandırma yönteminde normalizasyon yöntemlerinin sınıflama performansına doğrudan etkisi olmuştur. En iyi performans artışı z-skor normalizasyon yönteminde % 78.25 sınıflama doğruluğu elde edilmiştir.

KA sınıflandırma yönteminde normalizasyon yöntemlerinin sınıflama performansına pek bir etkisinin olmadığı görülmüştür. Sadece min-mak normalizasyon yönteminde sınıflama doğruluğu % 77.25 olarak daha iyi bir sonuca ulaşılmıştır.

DVM sınıflandırma yönteminde normalizasyon yöntemlerinin sınıflandırma performansını artırmadığı hatta minimum maksimum normalizasyon yöntemi ve ondalık ölçekleme normalizasyon yönteminde olumsuz etkilediği görülmüştür.

YSA sınıflandırma yönteminde de normalizasyon yöntemlerinin sınıflandırma performansını değiştirmediği görülmüştür.

Naive Bayes sınıflandırma yönteminde odalık ölçekleme normalizasyon yöntemi hariç diğer normalizasyon yöntemlerinin sınıflandırma algoritmalarının sınıflandırma performansına olumlu etkisi olmuştur. En yüksek sınıflama doğruluğu norm normalizasyon yöntemi ile % 74.25 olarak elde edilmiştir.

Sonuç olarak normalizasyon yöntemlerinin göğüs kanseri hastalığı verilerinin sınıflandırma performansına azda olsa olumlu etkisinin olabileceği görülmüştür.

Göğüs kanseri hastalığı ham verisine ve 4 faklı normalizasyon yöntemlerine k-kat çaprazlamanın etkisini görmek için Çizelge 5.8'den Çizelge 5.12'ye kadar ortalama sınıflama doğrulukları alınarak Çizelge 5.14'te toplu olarak gösterilmiştir.

Çizelge 5.14 Göğüs kanseri hastalığı verilerine k-kat çaprazlamanın etkisinin değerlendirilmesi

Göğüs kanseri hastalığı veri tipi	**Ortalama sınıflama doğruluğu (%)**			
	k-kat çaprazlama			
	2	**5**	**10**	**20**
Ham veri	72.6	71	**75.2**	**75.2**
Minimum maksimum normalizasyon	77.2	77.2	**79.8**	79.2
Ondalık ölçekleme normalizasyon	75.8	76.8	78	**78.2**
Z-skor normalizasyon	75.8	77.2	**80.6**	80
Norm normalizasyon	75.4	73	76	**77**

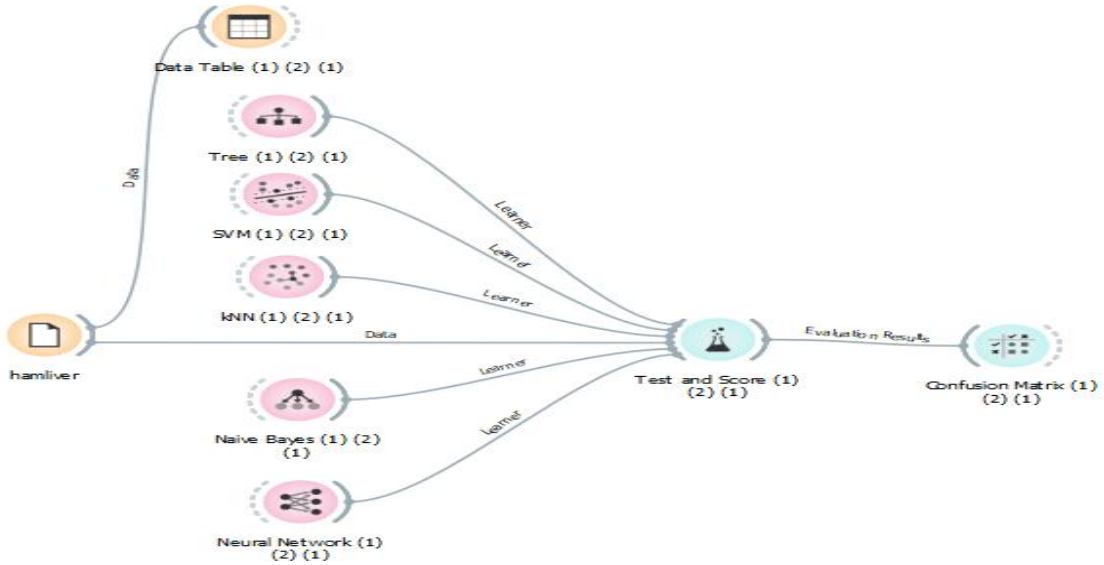
Çizelge 5.14'e göre, göğüs kanseri hastalığı veri seti doğru k-kat çaprazlama seçiminin sınıflama doğruluklarına bakıldığında en yüksek sınıflama doğruluğu genel olarak 10-kat çaprazlamada olmuştur.

5.3 Karaciğer Hastalığı Sınıflandırma Performans Sonuçları

Karaciğer hastalığının sınıflandırma performansını değerlendirmek için 50 sağlıklı ve 50 hasta verisi olmak üzere toplam 100 kişiden değerler alınmıştır. Sınıflandırma işlemi ham verilere, minimum maksimum normalizasyon yöntemi uygulanmış verilere, ondalık

ölçekleme normalizasyon yöntemi uygulanmış verilere, z-skor normalizasyon yöntemi uygulanmış verilere ve norm normalizasyon yöntemi uygulanmış verilere olmak üzere 5 farklı durum için değerlendirilmiştir. Karaciğer hastalığı verisinin ham veri ve normalize edilmiş durumu DVM, YSA, k-NN, KA ve Naive Bayes gibi sınıflandırma yöntemleri ile 4 farklı k-kat çaprazlamada (2,5,10,20) sınıflama doğruluğu değerlendirilmiştir.

Şekil 5.11'de karaciğer hastalığı ham veri setinin ORANGE programında KA, DVM, k-NN, Naive Bayes ve YSA gibi sınıflandırma algoritmaları kullanılarak sınıflandırma işlemine tabi tutulması gösterilmiştir.



Şekil 5.11 Karaciğer hastalığı ham verisinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Karaciğer hastalığı ham verisine 5 faklı sınıflandırma yöntemi kullanılarak yapılan sınıflandırma işlemlerinin sonucu 2,5,10 ve 20-kat çaprazlamada elde edilen sınıflama doğruluğu Çizelge 5.15'de gösterilmiştir.

Çizelge 5.15 Ham karaciğer hastalığı veri setinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Sınıflandırma Yöntemi	Sınıflama doğruluğu (%)				
	k-kat çaprazlama				
	2	**5**	**10**	**20**	**Ortalama**
k-NN	54	64	**68**	66	63
KA	59	**63**	62	58	60.5
DVM	**67**	65	**67**	66	66.25
YSA	**75**	71	74	70	72.5
Naive Bayes	74	73	74	**75**	**74**
Ortalama	65.8	67.2	**69**	67	

Çizelge 5.15'e göre;

k-NN sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 68 olarak bulunmuştur.

KA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 5-kat çaprazlamada % 63 olarak elde edilmiştir.

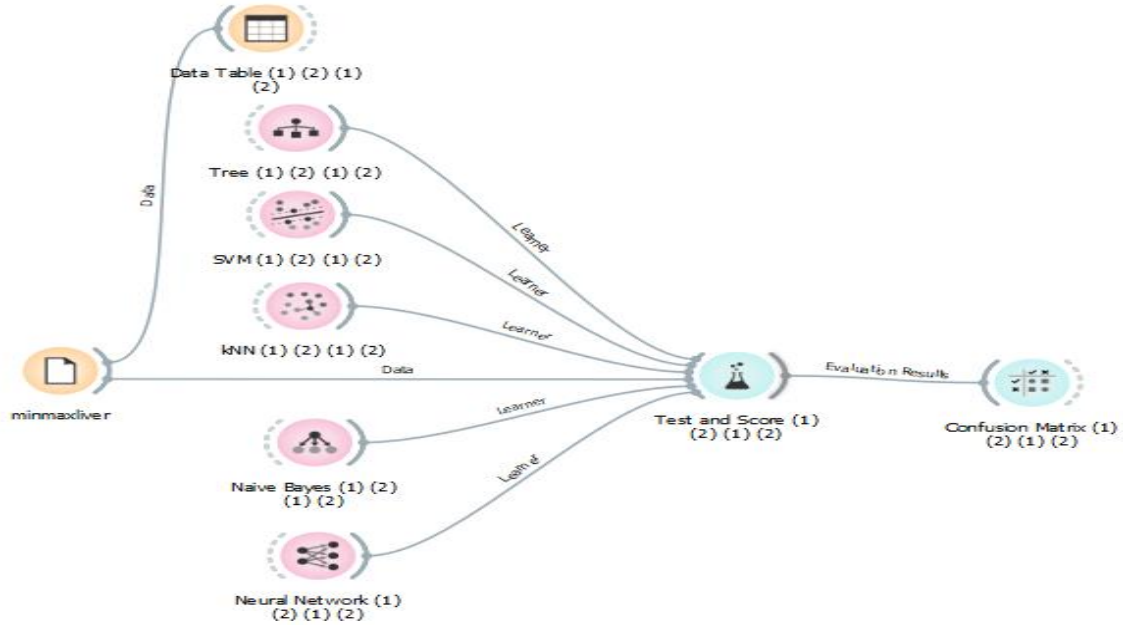
DVM sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 2-kat ve 10-kat çaprazlamada % 67 olarak bulunmuştur.

YSA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 2-kat çaprazlamada % 75 olarak bulunmuştur.

Naive Bayes sınıflandırma yöntemindeyse, 20-kat çaprazlamada % 75 olarak en yüksek sınıflama doğruluğu elde edilmiştir. 5 sınıflandırma yöntemi içinde ortalamalara bakıldığında ise % 74 sınıflama doğruluğuyla en yüksek sonuca Naive Bayes sınıflandırma yönteminde ulaşılmıştır.

k-kat çaprazlamanın karaciğer hastalığı ham verilerine etkisine bakıldığında; en yüksek sınıflama doğruluğu 10-kat çaprazlamada ortalama % 69 olmuştur.

Şekil 5.12'de karaciğer hastalığı veri setinin minimum maksimum normalizasyon yöntemi uygulanması sonrası ORANGE programında KA, DVM, k-NN, Naive Bayes ve YSA sınıflandırma algoritmaları kullanılarak sınıflandırma işlemine tabi tutulması gösterilmiştir.



Şekil 5.12 Karaciğer hastalığı minimum maksimum normalizasyon yöntemi uygulanmış verisinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Karaciğer hastalığı verisine minimum maksimum normalizasyon yöntemiyle 5 farklı sınıflandırma yöntemi kullanılarak yapılan sınıflandırma işlemlerinin sonucu 2,5,10 ve 20-kat çaprazlamada elde edilen sınıflama doğruluğu Çizelge 5.16'da gösterilmiştir.

Çizelge 5.16 Minimum maksimum normalizasyon yöntemi uygulanmış karaciğer hastalığı veri setinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Sınıflandırma Yöntemi	Sınıflama doğruluğu (%)				
	Çaprazlama Sayısı				
	2	**5**	**10**	**20**	**Ortalama**
k-NN	67	**68**	65	**68**	67
KA	59	62	**63**	61	61.25
DVM	**67**	65	**67**	66	66.25
YSA	**75**	71	74	71	72.75
Naive Bayes	74	74	74	**76**	**74.5**
Ortalama	68.4	68	**68.6**	68.4	

Çizelge 5.16'ya göre;

k-NN sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 5-kat ve 20-kat çaprazlamada % 68 olarak bulunmuştur.

KA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 63 olarak elde edilmiştir.

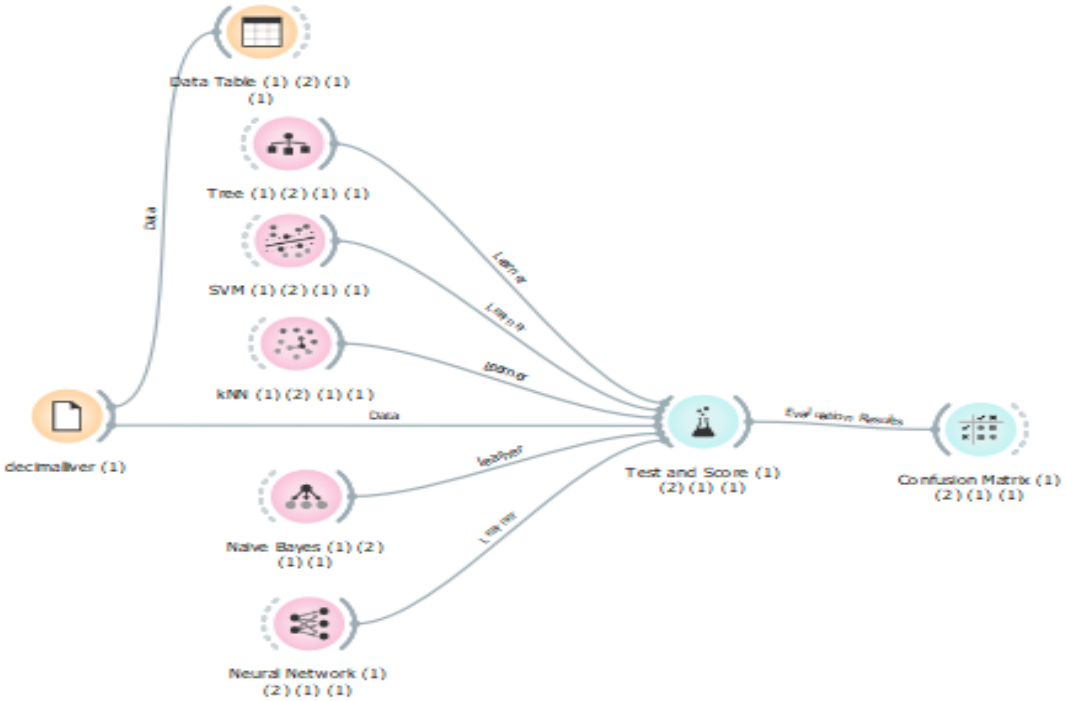
DVM sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 2-kat ve 10-kat çaprazlamada % 67 olarak bulunmuştur.

YSA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 2-kat çaprazlamada % 75 olarak bulunmuştur.

Naive Bayes sınıflandırma yöntemindeyse, 20-kat çaprazlamada % 76 olarak en yüksek sınıflama doğruluğu elde edilmiştir. 5 sınıflandırma yöntemi içinde ortalamalara bakıldığında ise % 74.5 sınıflama doğruluğuyla en yüksek sonuca Naive Bayes sınıflandırma yönteminde ulaşılmıştır.

k-kat çaprazlamanın karaciğer hastalığı minimum maksimum normalizasyon verilerine etkisine bakıldığında; en yüksek sınıflama doğruluğu 10-kat çaprazlamada ortalama % 68.6 olmuştur.

Şekil 5.13'te karaciğer hastalığı veri setinin ondalık ölçekleme normalizasyon yöntemi uygulanması sonrası ORANGE programında KA, DVM, k-NN, Naive Bayes ve YSA sınıflandırma algoritmaları kullanılarak sınıflandırma işlemine tabi tutulması gösterilmiştir.



Şekil 5.13 Karaciğer hastalığı ondalık ölçekleme normalizasyon yöntemi uygulanmış verisinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Karaciğer hastalığı verisine ondalık ölçekleme normalizasyon yöntemiyle 5 farklı sınıflandırma yöntemi kullanılarak yapılan sınıflandırma işlemlerinin sonucu 2,5,10 ve 20-kat çaprazlamada elde edilen sınıflama doğruluğu Çizelge 5.17'de gösterilmiştir.

Çizelge 5.17 Ondalık ölçekleme normalizasyon yöntemi uygulanmış karaciğer hastalığı veri setinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Sınıflandırma Yöntemi	Sınıflama Doğruluğu (%)				
	Çaprazlama Sayısı				
	2	**5**	**10**	**20**	**Ortalama**
k-NN	62	63	63	**68**	64
KA	58	**60**	58	51	56.75
DVM	**67**	65	**67**	65	66
YSA	73	71	**75**	69	72
Naive Bayes	75	**76**	74	**76**	**75.25**
Ortalama	67	67	**67.4**	65.8	

Çizelge 5.17'ye göre;

k-NN sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 20-kat çaprazlamada % 68 olarak bulunmuştur.

KA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 5-kat çaprazlamada % 60 olarak elde edilmiştir.

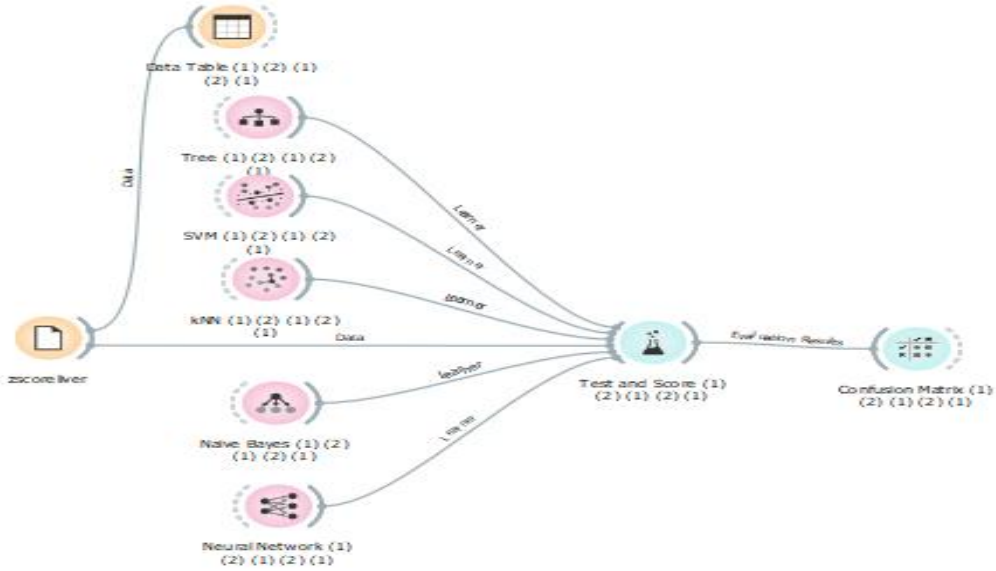
DVM sınıflandırma yönteminde, 2 ve 10-kat çaprazlamada % 67 olarak en yüksek sınıflama doğruluğu elde edilmiştir.

YSA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 75 olarak bulunmuştur.

Naive Bayes sınıflandırma yöntemindeyse, 5 ve 20-kat çaprazlamada % 76 olarak en yüksek sınıflama doğruluğu elde edilmiştir. 5 sınıflandırma yöntemi içinde ortalamalara bakıldığında ise % 75.25 sınıflama doğruluğuyla en yüksek sonuca Naive Bayes sınıflandırma yöntemi ile ulaşılmıştır.

k-kat çaprazlamanın karaciğer hastalığı ondalık ölçekleme normalizasyon verilerine etkisine bakıldığında; en yüksek sınıflama doğruluğu 10-kat çaprazlamada ortalama % 67.4 olmuştur.

Şekil 5.14'te karaciğer hastalığı veri setinin z-skor normalizasyon yöntemi uygulanması sonrası ORANGE programında KA, DVM, k-NN, Naive Bayes ve YSA sınıflandırma algoritmaları kullanılarak sınıflandırma işlemine tabi tutulması gösterilmiştir.



Şekil 5.14 Karaciğer hastalığı z-skor normalizasyon yöntemi uygulanmış verisinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Karaciğer hastalığı verisine z-skor normalizasyon yöntemiyle 5 farklı sınıflandırma yöntemi kullanılarak yapılan sınıflandırma işlemlerinin sonucu 2,5,10 ve 20-kat çaprazlamada elde edilen sınıflama doğruluğu Çizelge 5.18'de gösterilmiştir.

Çizelge 5.18 Z-skor ölçekleme normalizasyon yöntemi uygulanmış karaciğer hastalığı veri setinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Sınıflandırma Yöntemi	Sınıflama Doğruluğu (%)				
	Çaprazlama Sayısı				
	2	**5**	**10**	**20**	**Ortalama**
k-NN	65	63	65	**66**	64.75
KA	59	**63**	62	58	60.5
DVM	**67**	65	**67**	66	66.25
YSA	**75**	71	73	70	72.25
Naive Bayes	74	73	74	**75**	**74**
Ortalama	68	67	**68.2**	**67**	

Çizelge 5.18'e göre;

k-NN sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 20-kat çaprazlamada % 66 olarak bulunmuştur.

KA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 5-kat çaprazlamada % 63 olarak elde edilmiştir.

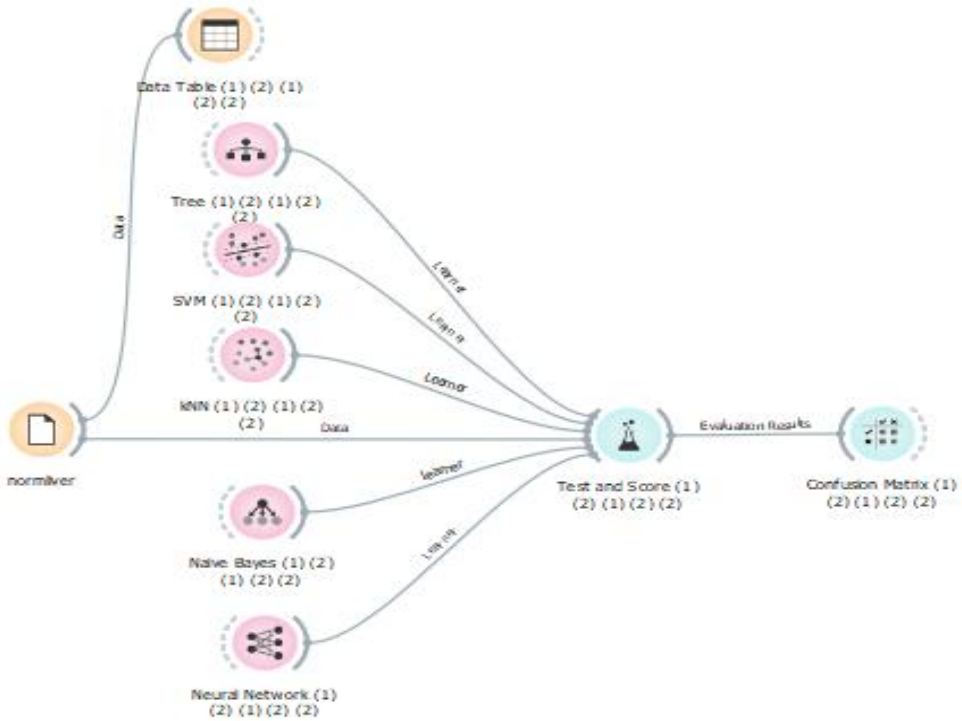
DVM sınıflandırma yönteminde, 2 ve 10-kat çaprazlamada % 67 olarak en yüksek sınıflama doğruluğu elde edilmiştir.

YSA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 2-kat çaprazlamada % 75 olarak bulunmuştur.

Naive Bayes sınıflandırma yöntemindeyse, 20-kat çaprazlamada % 75 olarak en yüksek sınıflama doğruluğu elde edilmiştir. 5 sınıflandırma yöntemi içinde ortalamalara bakıldığında ise % 74 sınıflama doğruluğuyla en yüksek sonuca ulaşılmıştır.

k-kat çaprazlamanın karaciğer hastalığı z-skor normalizasyon verilerine etkisine bakıldığında; en yüksek sınıflama doğruluğu 10-kat çaprazlamada ortalama % 68.2 olmuştur.

Şekil 5.15'te karaciğer hastalığı veri setinin norm normalizasyon yöntemi uygulanması sonrası ORANGE programında KA, DVM, k-NN, Naive Bayes ve YSA sınıflandırma algoritmaları kullanılarak sınıflandırma işlemine tabi tutulması gösterilmiştir.



Şekil 5.15 Karaciğer hastalığı norm normalizasyon yöntemi uygulanmış verisinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Karaciğer hastalığı verisine norm normalizasyon yöntemiyle 5 farklı sınıflandırma yöntemi kullanılarak yapılan sınıflandırma işlemlerinin sonucu 2,5,10 ve 20-kat çaprazlamada elde edilen sınıflama doğruluğu Çizelge 5.19'da gösterilmiştir.

Çizelge 5.19 Norm ölçekleme normalizasyon yöntemi uygulanmış karaciğer hastalığı veri setinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Sınıflandırma Yöntemi	Sınıflama Doğruluğu (%)				
	Çaprazlama Sayısı				
	2	**5**	**10**	**20**	**Ortalama**
k-NN	67	66	**68**	66	66.75
KA	59	62	**65**	64	62.5
DVM	**67**	65	**67**	65	66
YSA	73	70	**75**	69	71.75
Naive Bayes	**74**	**74**	72	72	**73**
Ortalama	68	67.4	**69.4**	67.2	

Çizelge 5.19'a göre;

k-NN sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 68 olarak bulunmuştur.

KA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 65 olarak elde edilmiştir.

DVM sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 2 ve 10-kat çaprazlamada % 67 sınıflama doğruluğu elde edilmiştir.

YSA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 75 olarak bulunmuştur.

Naive Bayes sınıflandırma yöntemindeyse, 2 ve 5-kat çaprazlamada % 74 olarak en yüksek sınıflama doğruluğu elde edilmiştir. 5 sınıflandırma yöntemi içinde ortalamalara bakıldığında ise % 73 sınıflama doğruluğuyla en yüksek sonuca ulaşılmıştır.

k-kat çaprazlamanın karaciğer hastalığı norm normalizasyon verilerine etkisine bakıldığında; en yüksek sınıflama doğruluğu 10-kat çaprazlamada ortalama % 69.4 olmuştur.

Normalizasyon yöntemlerinin sınıflandırma performansına etkisini değerlendirmek için ayrı ayrı değerlendirilen k-kat çaprazlamaların sınıflama doğruluklarının ortalamaları Çizelge 5.20'de gösterilmiştir.

Çizelge 5.20 Normalizasyon yöntemlerinin karaciğer hastalığı veri setinin sınıflandırma performansına etkisinin karşılaştırması

Sınıflandırma Yöntemleri	Ham Veri Sınıflama Doğruluğu (%)	Sınıflama Doğruluğu (%)			
		Minimum Maksimum	Ondalık Ölçekleme	Z-skor	Norm Yöntemi
k-NN	63	**67**	64	64.75	66.75
KA	60.5	61.25	56.75	60.5	**62.5**
DVM	66.25	66.25	66	66.25	66
YSA	72.5	**72.75**	72	72.25	71.75
Naive Bayes	74	74.5	**75.25**	74	73
Ortalama	67.25	**68.35**	66.8	67.55	68

Çizelge 5.20'ye göre,

k-NN sınıflandırma yönteminde normalizasyon yöntemlerinin sınıflama performansına doğrudan etkisi olmuştur. En iyi performans artışı min-mak normalizasyon yönteminde % 67 sınıflama doğruluğu elde edilmiştir.

KA sınıflandırma yönteminde normalizasyon yöntemlerinin sınıflama performansına min-mak normalizasyon ve norm normalizasyon yönteminde olumlu bir etkisi olmuş ve en iyi performansa % 62.5 ile norm normalizasyon yöntemi ile ulaşılmıştır.

DVM sınıflandırma yönteminde normalizasyon yöntemlerinin performansı artırmadığı hatta ondalık ölçekleme normalizasyon yöntemi ve norm normalizasyon yönteminde olumsuz etkilediği görülmüştür.

YSA sınıflandırma yönteminde sadece min-mak normalizasyon yönteminde sınıflama doğruluğunun % 72.75 ile daha iyi olduğu görülmüştür.

Naive Bayes sınıflandırma yönteminde min-mak ve ondalık ölçekleme normalizasyon yönteminde sınıflama doğruluğunu arttırdığı görülmüştür. En iyi başarı % 75.25 ile ondalık ölçekleme normalizasyon yöntemine ait olmuştur. Diğer normalizasyon yöntemlerinin olumlu bir etkisinin olmadığı görülmüştür.

Sonuç olarak normalizasyon yöntemlerinin karaciğer hastalığı verilerinin sınıflandırma performansına da pek az olumlu bir etkisinin olduğu görülmüştür.

Karaciğer hastalığı ham verisine ve 4 faklı normalizasyon yöntemlerine k-kat çaprazlamanın etkisini görmek için Çizelge 5.15'ten Çizelge 5.19'a kadar ortalama sınıflama doğrulukları alınarak Çizelge 5.21'de toplu olarak gösterilmiştir.

Çizelge 5.21 Karaciğer hastalığı verilerine k-kat çaprazlamanın etkisinin değerlendirilmesi

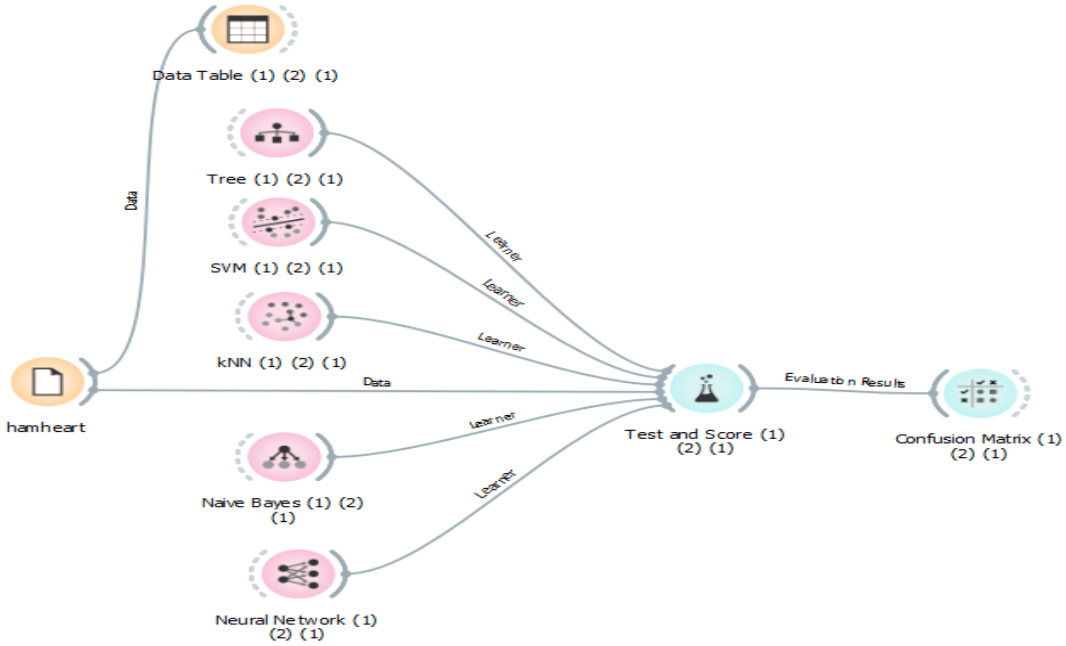
Karaciğer hastalığı veri tipi	**Ortalama sınıflama doğruluğu (%)**			
	k-kat çaprazlama			
	2	**5**	**10**	**20**
Ham veri	65.8	67.2	**69**	**67**
Minimum maksimum normalizasyon	68.4	68	**68.6**	68.4
Ondalık ölçekleme normalizasyon	67	67	**67.4**	65.8
Z-skor normalizasyon	68	67	**68.2**	**67**
Norm normalizasyon	68	67.4	**69.4**	67.2

Çizelge 5.21'e göre, karaciğer hastalığı veri seti doğru k-kat çaprazlama seçiminin sınıflama doğruluklarına bakıldığında en yüksek sınıflama doğruluğu 10-kat çaprazlamada olmuştur.

5.4 Kalp Hastalığı Sınıflandırma Performans Sonuçları

Kalp hastalığının sınıflandırma performansını değerlendirmek için 50 sağlıklı ve 50 hasta verisi olmak üzere toplam 100 kişiden değerler alınmıştır. Sınıflandırma işlemi ham verilere, minimum maksimum normalizasyon yöntemi uygulanmış verilere, ondalık ölçekleme normalizasyon yöntemi uygulanmış verilere, z-skor normalizasyon yöntemi uygulanmış verilere ve norm normalizasyon yöntemi uygulanmış verilere olmak üzere 5 farklı durum için değerlendirilmiştir. Kalp hastalığı verisinin ham veri ve normalize edilmiş durumu DVM, YSA, k-NN, KA ve Naive Bayes gibi sınıflandırma yöntemleri ile 4 farklı k-kat çaprazlamada (2,5,10,20) sınıflama doğruluğu değerlendirilmiştir.

Şekil 5.16'da kalp hastalığı ham veri setinin ORANGE programında KA, DVM, k-NN, Naive Bayes ve YSA gibi sınıflandırma algoritmaları kullanılarak sınıflandırma işlemine tabi tutulması gösterilmiştir.



Şekil 5.16 Kalp hastalığı ham verisinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Kalp hastalığı ham verisine 5 faklı sınıflandırma yöntemi kullanılarak yapılan sınıflandırma işlemlerinin sonucu 2,5,10 ve 20-kat çaprazlamada elde edilen sınıflama doğruluğu Çizelge 5.22'de gösterilmiştir.

Çizelge 5.22 Ham kalp hastalığı veri setinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Sınıflandırma Yöntemi	Sınıflama doğruluğu (%)				
	k-kat çaprazlama				
	2	**5**	**10**	**20**	**Ortalama**
k-NN	61	**64**	**64**	**64**	63.25
KA	60	74	**79**	74	71.75
DVM	77	78	**81**	76	78
YSA	80	80	79	**81**	**80**
Naive Bayes	77	79	**81**	80	79.25
Ortalama	71	75	**76.8**	75	

Çizelge 5.22'ye göre;

k-NN sınıflandırma yönteminde, 2-kat çaprazlama kriteri hariç diğer çaprazlama değerlerinde % 64 sınıflama doğruluğu elde edilmiştir.

KA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 79 olarak elde edilmiştir.

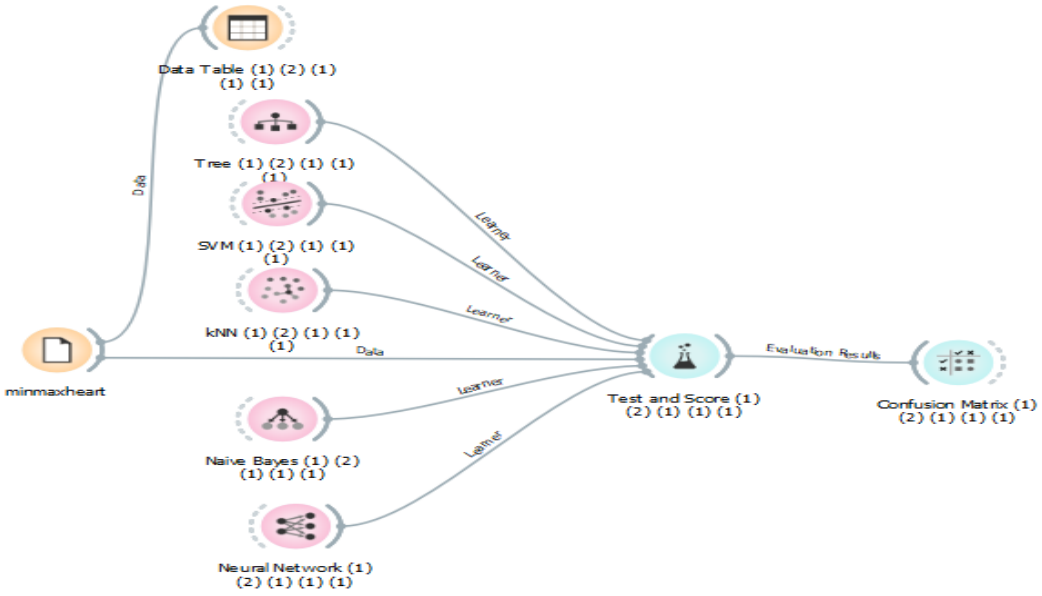
DVM sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 81 olarak bulunmuştur.

YSA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 20-kat çaprazlamada % 81 olarak bulunmuştur. 5 sınıflandırma yöntemi içinde ortalamalara bakıldığında ise % 80 sınıflama doğruluğuyla en yüksek sonuca YSA sınıflandırma yönteminde ulaşılmıştır.

Naive Bayes sınıflandırma yöntemindeyse, 10-kat çaprazlamada % 81 olarak en yüksek sınıflama doğruluğu elde edilmiştir.

k-kat çaprazlamanın kalp hastalığı ham verilerine etkisine bakıldığında; en yüksek sınıflama doğruluğu 10-kat çaprazlamada ortalama % 76.8 olmuştur.

Şekil 5.17'de kalp hastalığı veri setinin minimum maksimum normalizasyon yöntemi uygulanması sonrası ORANGE programında KA, DVM, k-NN, Naive Bayes ve YSA sınıflandırma algoritmaları kullanılarak sınıflandırma işlemine tabi tutulması gösterilmiştir.



Şekil 5.17 Kalp hastalığı minimum maksimum normalizasyon yöntemi uygulanmış verisinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Kalp hastalığı verisine minimum maksimum normalizasyon yöntemiyle 5 farklı sınıflandırma yöntemi kullanılarak yapılan sınıflandırma işlemlerinin sonucu 2,5,10 ve 20-kat çaprazlamada elde edilen sınıflama doğruluğu Çizelge 5.23'te gösterilmiştir.

Çizelge 5.23 Minimum maksimum normalizasyon yöntemi uygulanmış kalp hastalığı veri setinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Sınıflandırma Yöntemi	Sınıflama doğruluğu (%)				
	Çaprazlama Sayısı				
	2	**5**	**10**	**20**	**Ortalama**
k-NN	**79**	75	76	76	76.5
KA	60	74	**79**	74	71.75
DVM	77	78	**81**	76	78
YSA	80	80	79	**81**	**80**
Naive Bayes	77	79	**81**	80	79.25
Ortalama	74.6	77.2	**79.2**	77.4	

Çizelge 5.23'e göre;

k-NN sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 2-kat çaprazlamada % 79 olarak bulunmuştur.

KA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 79 olarak elde edilmiştir.

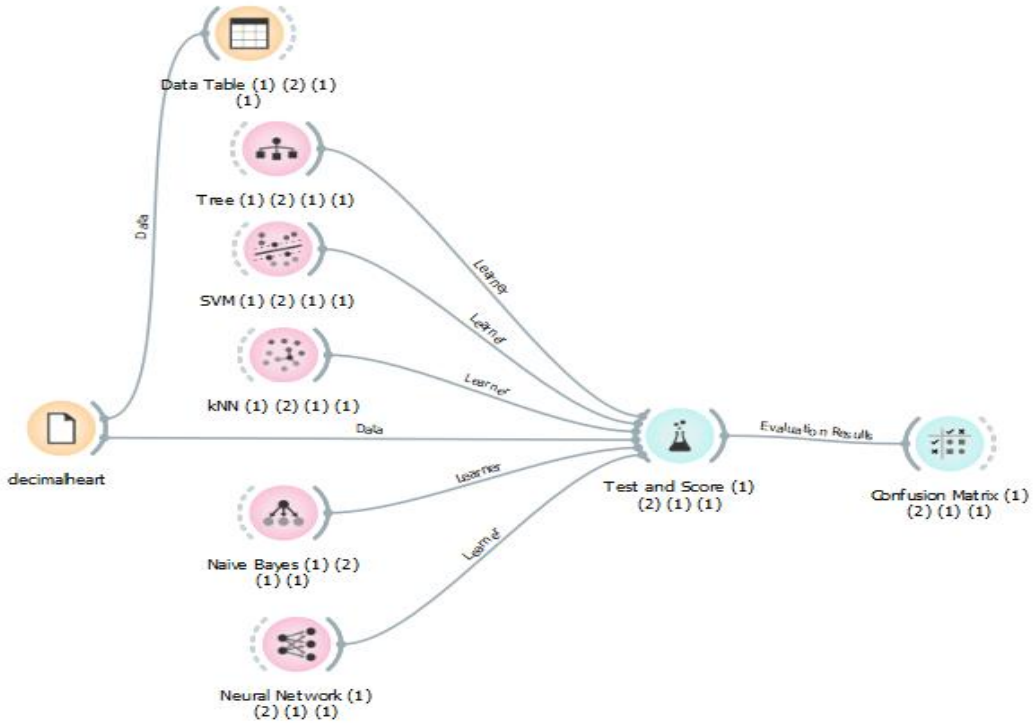
DVM sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 81 olarak elde edilmiştir.

YSA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 20-kat çaprazlamada % 81 olarak bulunmuştur. 5 sınıflandırma yöntemi içinde ortalamalara bakıldığında ise % 80 sınıflama doğruluğuyla en yüksek sonuca YSA sınıflandırma yönteminde ulaşılmıştır.

Naive Bayes sınıflandırma yöntemindeyse, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 81 olarak elde edilmiştir.

k-kat çaprazlamanın kalp hastalığı minimum maksimum normalizasyon verilerine etkisine bakıldığında; en yüksek sınıflama doğruluğu 10-kat çaprazlamada ortalama % 79.2 olmuştur.

Şekil 5.18'de kalp hastalığı veri setinin ondalık ölçekleme normalizasyon yöntemi uygulanması sonrası ORANGE programında KA, DVM, k-NN, Naive Bayes ve YSA sınıflandırma algoritmaları kullanılarak sınıflandırma işlemine tabi tutulması gösterilmiştir.



Şekil 5.18 Kalp hastalığı ondalık ölçekleme normalizasyon yöntemi uygulanmış verisinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Kalp hastalığı verisine ondalık ölçekleme normalizasyon yöntemiyle 5 farklı sınıflandırma yöntemi kullanılarak yapılan sınıflandırma işlemlerinin sonucu 2,5,10 ve 20-kat çaprazlamada elde edilen sınıflama doğruluğu Çizelge 5.24'te gösterilmiştir.

Çizelge 5.24 Ondalık ölçekleme normalizasyon yöntemi uygulanmış kalp hastalığı veri setinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Sınıflandırma Yöntemi	Sınıflama doğruluğu (%)				
	k-kat çaprazlama				
	2	**5**	**10**	**20**	**Ortalama**
k-NN	74	78	78	**79**	77.25
KA	60	74	**79**	74	71.75
DVM	77	78	**81**	76	78
YSA	80	80	79	**81**	**80**
Naive Bayes	77	79	**81**	80	79.25
Ortalama	73.6	77.8	**79.6**	78	

Çizelge 5.24'e göre;

k-NN sınıflandırma yönteminde, 20-kat çaprazlamada % 79 olarak en yüksek sınıflama doğruluğu elde edilmiştir.

KA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 79 olarak elde edilmiştir.

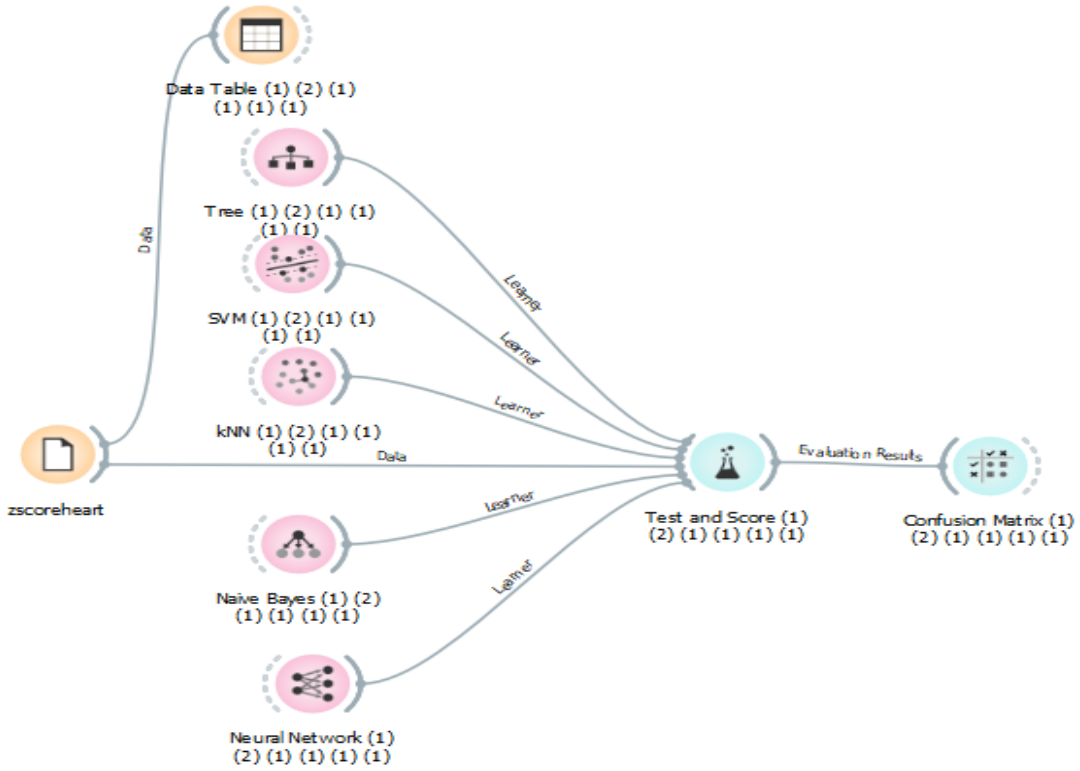
DVM sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 81 olarak elde edilmiştir.

YSA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 20-kat çaprazlamada % 81 olarak bulunmuştur. 5 sınıflandırma yöntemi içinde ortalamalara bakıldığında ise % 80 sınıflama doğruluğuyla en yüksek sonuca YSA sınıflandırma yöntemi ile ulaşılmıştır.

Naive Bayes sınıflandırma yöntemindeyse, 10-kat çaprazlamada % 81 olarak en yüksek sınıflama doğruluğu elde edilmiştir.

k-kat çaprazlamanın kalp hastalığı ondalık ölçekleme verilerine etkisine bakıldığında; en yüksek sınıflama doğruluğu 10-kat çaprazlamada ortalama % 79.6 olmuştur.

Şekil 5.19'da kalp hastalığı veri setinin z-skor normalizasyon yöntemi uygulanması sonrası ORANGE programında KA, DVM, k-NN, Naive Bayes ve YSA sınıflandırma algoritmaları kullanılarak sınıflandırma işlemine tabi tutulması gösterilmiştir.



Şekil 5.19 Kalp hastalığı z-skor normalizasyon yöntemi uygulanmış verisinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Kalp hastalığı verisine z-skor normalizasyon yöntemiyle 5 farklı sınıflandırma yöntemi kullanılarak yapılan sınıflandırma işlemlerinin sonucu 2,5,10 ve 20-kat çaprazlamada elde edilen sınıflama doğruluğu Çizelge 5.25'te gösterilmiştir.

Çizelge 5.25 Z-skor ölçekleme normalizasyon yöntemi uygulanmış kalp hastalığı veri setinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Sınıflandırma Yöntemi	Sınıflama doğruluğu (%)				
	k-kat çaprazlama				
	2	**5**	**10**	**20**	**Ortalama**
k-NN	80	**81**	80	79	**80**
KA	60	74	**79**	74	71.75
DVM	77	78	**81**	76	78
YSA	**82**	78	80	79	79.75
Naive Bayes	77	79	**81**	80	79.25
Ortalama	75.2	78	**80.2**	77.6	

Çizelge 5.25'e göre;

k-NN sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 5-kat çaprazlamada % 81 olarak bulunmuştur. 5 sınıflandırma yöntemi içinde ortalamalara bakıldığında ise % 80 sınıflama doğruluğuyla en yüksek sonuca ulaşılmıştır.

KA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 79 olarak elde edilmiştir.

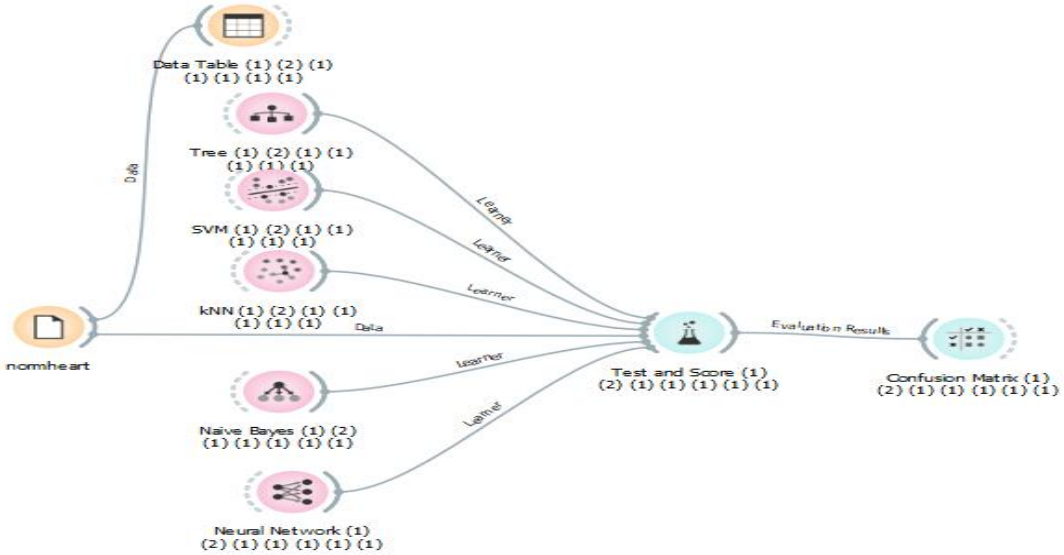
DVM sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 81 olarak elde edilmiştir.

YSA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 2-kat çaprazlamada % 82 olarak bulunmuştur.

Naive Bayes sınıflandırma yöntemindeyse, 10-kat çaprazlamada % 81 olarak en yüksek sınıflama doğruluğu elde edilmiştir.

k-kat çaprazlamanın kalp hastalığı z-skor normalizasyon verilerine etkisine bakıldığında; en yüksek sınıflama doğruluğu 10-kat çaprazlamada ortalama % 80.2 olmuştur.

Şekil 5.20'de kalp hastalığı veri setinin norm normalizasyon yöntemi uygulanması sonrası ORANGE programında KA, DVM, k-NN, Naive Bayes ve YSA sınıflandırma algoritmaları kullanılarak sınıflandırma işlemine tabi tutulması gösterilmiştir.



Şekil 5.20 Kalp hastalığı norm normalizasyon yöntemi uygulanmış verisinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Kalp hastalığı verisine norm normalizasyon yöntemiyle 5 farklı sınıflandırma yöntemi kullanılarak yapılan sınıflandırma işlemlerinin sonucu 2,5,10 ve 20-kat çaprazlamada elde edilen sınıflama doğruluğu Çizelge 5.26'da gösterilmiştir.

Çizelge 5.26 Norm ölçekleme normalizasyon yöntemi uygulanmış kalp hastalığı veri setinin 5 farklı sınıflandırma yöntemi ile değerlendirilmesi

Sınıflandırma Yöntemi	Sınıflama Doğruluğu (%)				
	Çaprazlama Sayısı				
	2	**5**	**10**	**20**	**Ortalama**
k-NN	72	76	**79**	76	75.75
KA	60	74	**79**	75	72
DVM	77	78	**81**	76	78
YSA	**81**	80	79	**81**	**80.25**
Naive Bayes	77	80	80	**81**	79.5
Ortalama	73.4	77.6	**79.6**	77.8	

Çizelge 5.26'ya göre;

k-NN sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 79 olarak bulunmuştur.

KA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 79 olarak elde edilmiştir.

DVM sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 10-kat çaprazlamada % 81 sınıflama doğruluğu elde edilmiştir.

YSA sınıflandırma yönteminde, en yüksek sınıflama doğruluğu 2 ve 20-kat çaprazlamada % 81 olarak bulunmuştur. 5 sınıflandırma yöntemi içinde ortalamalara bakıldığında ise % 80.25 sınıflama doğruluğuyla en yüksek sonuca ulaşılmıştır.

Naive Bayes sınıflandırma yöntemindeyse, 20-kat çaprazlamada % 81 olarak en yüksek sınıflama doğruluğu elde edilmiştir.

k-kat çaprazlamanın kalp hastalığı norm normalizasyon verilerine etkisine bakıldığında; en yüksek sınıflama doğruluğu 10-kat çaprazlamada ortalama % 79.6 olmuştur.

Normalizasyon yöntemlerinin sınıflandırma performansına etkisini değerlendirmek için ayrı ayrı değerlendirilen k-kat çaprazlamaların sınıflama doğruluklarının ortalamaları Çizelge 5.27'de gösterilmiştir.

Çizelge 5.27 Normalizasyon yöntemlerinin kalp hastalığı veri setinin sınıflandırma performansına etkisinin karşılaştırması

Sınıflandırma Yöntemleri	**Ham Veri Sınıflama Doğruluğu (%)**	**Sınıflama Doğruluğu (%)**			
		Minimum Maksimum	**Ondalık Ölçekleme**	**Z-skor**	**Norm Yöntemi**
k-NN	63.25	76.5	77.25	**80**	75.75
KA	71.75	71.75	71.75	71.75	**72**
DVM	78	78	78	78	78
YSA	80	80	80	79.75	**80.25**
Naive Bayes	**79.25**	**79.25**	**79.25**	79.25	**79.5**
Ortalama	74.45	77.1	77.25	**77,75**	77.1

Çizelge 5.27'ye göre,

k-NN sınıflandırma yönteminde normalizasyon yöntemlerinin sınıflama performansına doğrudan etkisi olmuştur. En iyi performans artışı z-skor normalizasyon yönteminde % 80 sınıflama doğruluğu elde edilmiştir.

KA sınıflandırma yönteminde sadece norm normalizasyon yönteminin %72 başarı performansı ile olumlu etkisi olmuştur.

DVM sınıflandırma yönteminde normalizasyon yöntemlerinin performansı değiştirmediği görülmüştür.

YSA sınıflandırma yönteminde sadece norm normalizasyon yönteminde sınıflama doğruluğu az da olsa arttığı (% 80.25) görüldü. Diğer normalizasyon yöntemlerinde z-skor normalizasyon yöntemi hariç performansı değişmemiştir.

Naive Bayes sınıflandırma yönteminde sadece norm normalizasyon yönteminde sınıflama doğruluğunu arttırdığı (% 79.5) görülmüştür. Diğer normalizasyon yöntemlerinin bir etkisinin olmadığı görülmüştür.

Sonuç olarak normalizasyon yöntemlerinin kalp hastalığı verilerinin sınıflandırma performansına pek az bir etkisinin olduğu görülmüştür.

Kalp hastalığı ham verisine ve 4 faklı normalizasyon yöntemlerine k-kat çaprazlamanın etkisini görmek için Çizelge 5.22'den Çizelge 5.26'ya kadar ortalama sınıflama doğrulukları alınarak Çizelge 5.28'de toplu olarak gösterilmiştir.

Çizelge 5.28 Kalp hastalığı verilerine k-kat çaprazlamanın etkisinin değerlendirilmesi

Kalp hastalığı veri tipi	**Ortalama sınıflama doğruluğu (%)**			
	k-kat çaprazlama			
	2	**5**	**10**	**20**
Ham veri	71	75	**76.8**	75
Minimum maksimum normalizasyon	74.6	77.2	**79.2**	77.4
Ondalık ölçekleme normalizasyon	73.6	77.8	**79.6**	78
Z-skor normalizasyon	75.2	78	**80.2**	77.6
Norm normalizasyon	73.4	77.6	**79.6**	77.8

Çizelge 5.28'e göre, kalp hastalığı veri seti doğru k-kat çaprazlama seçiminin sınıflama doğruluklarına bakıldığında en yüksek sınıflama doğruluğu 10-kat çaprazlamada olmuştur.

5.5 Öneriler

Yapılan bu çalışmada Pima Hintlilerinin diyabet hastalığı verisi, göğüs kanseri hastalığı verisi, karaciğer hastalığı verisi ve kalp hastalığı verisine min-mak normalizasyon yöntemi, ondalık ölçekleme normalizasyon yöntemi, z-skor normalizasyon yöntemi ve norm normalizasyon yöntemleri uygulanmıştır. Normalizasyon işlemi ardından veriler DVM, YSA, KNN, KA ve Naive Bayes gibi sınıflandırma yöntemleri ile 4 farklı k-kat çaprazlama (2,5,10,20) kriterinde doğruluk performansı olarak değerlendirilmiştir.

Sonuç olarak:

Min-mak normalizasyon yönteminin, ondalık ölçekleme normalizasyon yönteminin, z-skor normalizasyon yönteminin ve norm normalizasyon yönteminin sınıflandırma doğruluk performansını artırabileceği görülmüştür. Farklı normalizasyon

yöntemlerinin de sınıflandırma performansına olumlu etki edebileceği düşüncesi gelişmiştir.

Sınıflandırma performansını farklı k-kat çaprazlama (2,5,10,20) değerlerinde daha iyi doğruluk performansı verebileceği açıkça görülmüştür. 4, 25, 50, 100 gibi 100 örnek veriyi tam bölen faklı k-kat çaprazlama değerlerinin de sınıflandırma performansını artırabileceği düşüncesi gelişmiştir.

KAYNAKLAR

Abdi H., Normalizing Data, In Neil Salkind (Ed.), Encyclopedia of Research Design, Thousand Oaks, CA: Sage. 2010

Abdulkareem A. H., Kasapbaşlı M. C., Enhancıng detectıon method of breast cancer usıng coımbra dataset, Teknoloji ve Uygulamalı Bilimler Dergisi, Cilt 03, No 01, s. 51-59

Ahidha M, Premalatha K. An application of fuzzy normalization in miRNA data for Novel feature selection in cancer classification. Biomedical Research 2017, 28 (9): 4187-4195

Akdemir B., Tahmin uygulamalarında performans geliştirmek için kullanılan normalizasyon metotlarına yeni bir yaklaşım, Selçuk Üniversitesi Fen Bilimleri Enstitüsü, Elektrik Elektronik Mühendisliği Anabilim Dalı, Doktora Tezi, 2009, Konya.

Akdoğan E., Mekatronik Mühendisliği Uygulamalarında Yapay Zekâ, Ders 3- Yapay Sinir Ağları.

Atomi V. H., The effect of data preprocessing on the performance of artificial neural networks techniques for classification problem, Master Thesis, Faculty of Computer Science and Information Technology, University Tun Hussein Onn Malaysia, December 2012.

Basheer I.A, Hajmeer M. Artificial neural networks: fundamentals, Computing, design, and application, Journal of Microbiological Methods 43 (2000) 3-31

Bolandraftar M., Imandoust S. B., Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background, S B Imandoust et al. Int. Journal of Engineering Research and Applications, Vol. 3, Issue 5, Sep-Oct 2013, pp.605-610

Borkin D, Nemethova A, Michal'conok G, Maiorov K. (2019) Impact of data normalization on classification model accuracy. Research Papers Faculty Materials Science and Technology in Trvana.Slovak University of Techology in Bratislava, 2019, Volume 27, Number 45, DOI 10.2478/rput-2019-0029

Çalışkan S B, Soğukpınar İ, k×knn: k-means ve k en yakın komşu yöntemleri ile ağlarda nüfuz tespiti ,2008)

Çalış A, Kayapınar S, Çetinyokuş T, Veri madenciliğinde karar ağacı algoritmaları ile bilgisayar ve internet güvenliği üzerine bir uygulama. Endüstri Mühendisliği Dergisi Makale Cilt: 25 Sayı: 3-4 Sayfa: (2-19)

Dener M, Dörterler M, Orman A, Açık kaynak kodlu veri madenciliği programları: weka'da örnek uygulama, Akademik Bilişim' 09- XI. Akademik Bilişim Konferansı Bildirileri 11-13 Şubat 2009 Harran Üniversitesi, Şanlıurfa

Eesa A.S., Arabo W. K., Normalization methods for backpropagation: A comparative study, Science Journal of University of Zakho, Vol. 5, No. 4, Dec.-2017, 314 – 318.

Gautam R., Vanga S., Ariese F., Umapathy S. Review of multidimensional data processing approaches for Raman and infrared spectroscopy. Gautam et al. EPJ Techniques and Instrumentation (2015) 2:8 DOI 10.1140/epjti/s40485-015-0018-6

Gör İ., Çok katmanlı algılayıcı yapay sinir ağı ile lineer diferansiyel denklem sisteminin çözümü. Adnan Menderes Üniversitesi, Matematik Bölümü, Aydın

https://kod5.org/yapay-sinir-aglari-ysa-nedir/ [Ziyaret Tarihi: 02 Mart 2020]

http://www.derinogrenme.com/2017/03/04/yapay-sinir-aglari/ [Ziyaret Tarihi: 02 Mart 2020]

https://medium.com/@k.ulgen90/makine-%C3%B6%C4%9Frenimi b%C3%B6l%C3%BCm-3-4b160df1f4c8 [Ziyaret Tarihi: 07 Mart 2020]

https://e-abm.com/how-to-establish-quality-and-correctness-of-classification-models-part-3-confusion-matrix/ [Ziyaret Tarihi: 12 Nisan 2020]

https://kodedu.com/2014/05/naive-bayes-siniflandirma-algoritmasi/[Ziyaret Tarihi: 27 Şubat 2021]

https://devhunteryz.wordpress.com/2019/12/02/naive-bayes-siniflandirici/[Ziyaret Tarihi: 27 Şubat 2021]

https://erdincuzun.com/makine_ogrenmesi/naive-bayes-classifier/ [Ziyaret Tarihi: 27 Şubat 2021]

https://medium.com/@Emreyz/y%C3%B6ntemler-3-naive-bayes899314be2018[Ziyaret Tarihi: 27 Şubat 2021]

https://medium.com/yapay-zeka-makine-%C3%B6%C4%9Frenmesi-derin %C3%B6%C4%9Frenme/denetimli-%C3%B6%C4%9Frenme-d7237c50b10b[Ziyaret Tarihi: 27 Şubat 2021]

https://aycaakcay.medium.com/k-en-yak%C4%B1n-kom%C5%9Fu-k-nearest-neighbour-algoritmas%C4%B1-s%C4%B1n%C4%B1flama-7c456f8e2b0d) [Ziyaret Tarihi: 27 Şubat 2021]

https://en.wikipedia.org/wiki/Orange_(software) [Ziyaret Tarihi: 6 Mart 2021]

https://orangedatamining.com/ [Ziyaret Tarihi: 6 Mart 2021]

https://orangedatamining.com/faq/# [Ziyaret Tarihi: 6 Mart 2021]

https://medium.com/@k.ulgen90/makine-%C3%B6%C4%9Frenimi-b%C3%B6l%C3%BCm-5-karar-a%C4%9Fa%C3%A7lar%C4%B1-c90bd7593010. [Ziyaret Tarihi: 6 Mart 2021]

https://veribilimcisi.com/2017/07/20/k-en-yakin-komsu-k-nearest-neighborsknn/[Ziyaret Tarihi: 7 Nisan 2021]

https://ec.europa.eu/jrc/en/coin/10-step-guide/step-5 [Ziyaret Tarihi: 13 Nisan 2021]

https://msatechnosoft.in/blog/artificial-neural-network-types-feed-forward-feedback-structure-perceptron-machine-learning-applications/ [Ziyaret Tarihi: 23 Nisan 2021]

https://medium.com/@ekrem.hatipoglu/machine-learning-prediction-algorithms-decision-tree-random-forest-part-5-2970905c021e [Ziyaret Tarihi: 24 Nisan 2021]

https://veribilimcisi.com/2017/07/19/destek-vektor-makineleri-support-vector-machine/ [Ziyaret Tarihi: 24 Nisan 2021]

Huang H.C., Qin L.X., Emprical evaluation of data normalization methods for molecular classification.Peer-REVIEWED Biochemistry, Biophysics and Molecular Biology Section, 2018, (doi: 10.7717/peerj.4584)

İleri S.C, Karabina A, Kılıç E. Comparison of different normalization techniques on speaker' gender detection, MAKÜ-Uyg.Bil.Der.,2(2),1-12,2018

Jadhav S. D., Channe H. P., Comparative study of k-nn, naive bayes and decision tree classification techniques. International Journal of Science and Research (IJSR). 2013

Jayalakshmi T, Santhakumaran J., Statistical Normalization and Back Propagation for Classification. International Journal of Computer Theory and Engineering (ISSN: 1793-8201), Vol.3, No.1, February 2011, 89-93.

Kabalcı E. (2014). Yapay Sinir Ağları. Ders Notları

Kavzoğlu T, Çölkesen İ, Destek vektör makineleri ile uydu görüntülerinin sınıflandırılmasında kernel fonksiyonlarının etkilerinin incelenmesi, Harita Dergisi Temmuz 2010 Sayı 144

Koç E., Güler S, Diyabet Diyabet hastalığı, Hacettepe Üniversitesi Tıp Fakültesi Halk Sağlığı Anabilim Dalı, Toplum İçin Bilgilendirme Sunumları,2015

Mohamed A. E., Comparative study of four supervised machine learning techniques for classification. International Journal of Applied Science and Technology. Vol. 7, No. 2, June 2017

Mustaffa Z., Yusof Y., A comparison of normalization techniques in predicting dengue outbreak. 2010 International conference on business and economics research. Vol.1 (2011) IACSIT Press Kuala Lumpur. Malaysia

Muthuselvan S., Rajapraksh S., Somasundaram K, KKarthik K., Classification of Liver Patient Dataset Using Machine Learning Algorithms, International Journal of Engineering & Technology, 7 (3.34) (2018) 323-326

Nandakumar K., Integration of multiple cues in biometrıc systems, Michigan State University in partial fulfillment of the requirements for the degree of MASTER OF SCIENCE, Department of computer science and engineering, 2005

Öğücü M. O., Yapay sinir ağları ile sistem tanıma, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2006, İstanbul

Özkan A. O, Durğun S, Normalizasyon tekniklerinin Romatoid Artrit hastalığı tanısı için YSA sınıflama performansına etkisi, EEB 2016 Elektrik-Elektronik ve Bilgisayar Sempozyumu, 11-13 Mayıs 2016, Tokat TÜRKİYE

Öztemel E., (2003). Yapay Sinir Ağları. İstanbul: Papatya, s.15-18.

Öztürk K, Şahin M. E, Yapay sinir ağları ve yapay zekaya genel bir bakış, Takvim-i Vekayi, ISSN:2148-0087, Basım Tarihi: 30 Aralık 2018 / 23 Rebiülahir 1440 Cilt: 6 No: 2 Sayfa: 25-36 (2018)

Peshawa J.M. A. ata normalization and standardization. department of software engineering, koya university, kurdistan region, Iraq.03,11,2015

Polat K, Biyomedikal sinyallerde veri ön-işleme tekniklerinin medikal teşhiste sınıflama doğruluğuna etkisinin incelenmesi. 2008, 151 sayfa. Doktora tezi

Singh B K, ThokeA. S ve Verma K. Investigations on ımpact of feature normalization techniques on classifier's performance in breast tumor classification, ınternational journal of computer applications (ISSN: 0975 – 8887) Volume 116 – No. 19, April 2015, 11-15.

Sun Y., Zhao Z., Yang Z., Xu F., Lu H., Zhu Z., Shi W., Jiang J., Yao P., Zhu H., Risk Factors and preventions of breast cancer. International journal of biological sciences. 2017; 13(11): 1387-1397. doi: 10.7150/ijbs.21635

tr.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/mean median-basics/a/mean-median-and-mode-review. [Ziyaret Tarihi: 26 Ağustos 2019]

Üstüner M, Şanlı F. B, Balçık F. B, Esetlili M. T, Destek vektör makineleri tekniği ile sınıflandırma: Rapideye örneği, Türkiye Ulusal Fotogrametri ve Uzaktan Algılama Birliği VII. Teknik Sempozyumu (TUFUAB'2013), 23-25 Mayıs 2013, KTÜ, Trabzon.

Välikangas T, Suomi T, and Elo L.L. A systematic evaluation of normalization methods in quantitative label-freeproteomics, Briefings in Bioinformatics, 19 (1), 2018, 1–11

Yavuz S. Deveci M., İstatiksel normalizasyon tekniklerinin yapay sinir ağın performansına etkisi, Erciyes Üniversitesi İktisadi ve İdari Bilimler Dergisi, Sayı 40, Haziran-Aralık 2012, 167-187.

Yazıcı A. C, Öğüş E, Ankaralı S, Canan S, Ankaralı H, Akkuş Z. Yapay sinir ağlarına genel bakış, Turkiye Klinikleri J Med Sci 2007, 27:65-71

Weinhaus A. J, Roberts K. P, Anatomy of the human heart

www.vertica.com/docs/9.2.x/HTML/Content/Authoring/AnalyzingData/Machine Learning/DataPreparation/NormalizingData.htm [Ziyaret Tarihi: 18 Ağustos 2019]

www.oreilly.com/library/view/regression-analysiswith/9781788627306/6bb0d820-6200-4bfe-aa91-e7b7ffa2a9c1.xhtml. [Ziyaret Tarihi: 26 Ağustos 2019]

www.codecademy.com/articles/normalization [Ziyaret Tarihi: 26 Ağustos 2019]