T.C.
VAN YÜZÜNCÜ YIL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI

KAYIP VERİ İÇEREN VERİ SETLERİNDE KÜMELEME UYGULAMALARI

DOKTORA TEZİ

HAZIRLAYAN: Serpil SEVİMLİ DENİZ
DANIŞMAN: Prof. Dr. H. Eray ÇELİK

VAN-2020

T.C.
VAN YÜZÜNCÜ YIL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI

KAYIP VERİ İÇEREN VERİ SETLERİNDE KÜMELEME UYGULAMALARI

DOKTORA TEZİ

HAZIRLAYAN: Serpil SEVİMLİ DENİZ

VAN-2020

KABUL VE ONAY SAYFASI

İstatistik Anabilim Dalı'nda Prof. Dr. H. Eray ÇELİK danışmanlığında, Serpil SEVİMLİ DENİZ tarafından sunulan "Kayıp Veri İçeren Veri Setlerinde Kümeleme Uygulamaları" isimli bu çalışma Lisansüstü Eğitim ve Öğretim Yönetmeliği'nin ilgili hükümleri gereğince 28/01/ 2020 tarihinde aşağıdaki jüri tarafından oy birliği / ~~oy çokluğu~~ ile başarılı bulunmuş ve doktora tezi olarak kabul edilmiştir.

Başkan: Doç. Dr. M. Nuri ALMALI	İmza:
Üye: Prof. Dr. H. Eray ÇELİK	İmza:
Üye: Doç. Dr. Çağdaş Hakan ALADAĞ	İmza:
Üye: Doç. Dr. Hamit MIRTAGIOĞLU	İmza:
Üye: Dr. Öğr. Üyesi: Sanem ŞEHRİBANOĞLU	İmza:

Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 07/02/2020 tarih ve 2020/9-I sayılı kararı ile onaylanmıştır.

İmza

.....

Enstitü Müdürü

Prof. Dr. [illegible]

Enstitü Müdürü

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

Serpil SEVİMLİ DENİZ

ÖZET

KAYIP VERİ İÇEREN VERİ SETLERİNDE KÜMELEME UYGULAMALARI

SEVİMLİ DENİZ, Serpil
Doktora Tezi, İstatistik Anabilim Dalı
Tez Danışmanı: Prof. Dr. H. Eray ÇELİK
Şubat 2020, 77 sayfa

Kayıp veriler yapılan çalışmalarda olası bir durumdur. Kümeleme analizi nesnelerin doğal gruplarını bulmak için kullanılan bir yöntemdir. Birçok alanda kümeleme analizi veri çözümlemesinde en çok kullanılan yaklaşımlardan biridir. Çözümlenen veri setlerinde çeşitli oranlarda kayıp veri olabilir. Sayısal verilerde tamamen rastgele olan kayıp veri içeren veri setlerinin analizi için kullanılabilecek kümeleme yöntemleri içinde hangisinin en iyi olduğu ile ilgili kesin bir bilgi bulunmamaktadır. Veri sayısına ve veri yapısına göre her bir yöntemin birbirine üstünlükleri ve eksiklikleri vardır. Bu çalışmada sürekli tam ve kayıp veri içeren verilerin kümelenmesi incelenmiştir. Bölmeli kümeleme tekniklerinden k-ortalamalar ve yapay sinir ağı tabanlı kümeleme tekniklerinden öz düzenleyici haritalar (SOM) ve doğrusal vektör parçalama (LVQ) yöntemleri kullanılarak kümeleme analizleri yapılmış ve elde edilen sonuçlar karşılaştırılmıştır. Nitelikli bir karşılaştırma yapmak için literatürde bu tür karşılaştırmaların yapılmasında yaygın olarak kullanılan on bir gerçek veri setlinden yararlanılmıştır. Analiz sonuçlarına göre tüm yöntemlerde kayıp veri oranı arttıkça doğru sınıflandırma oranlarının düştüğü ve yedi veri setinde SOM kümeleme yönteminin k-ortalamalar ve LVQ yöntemlerine göre daha iyi performans gösterdiği görülmektedir. Dört veri setinde ise verilerin yapısına bağlı olarak LVQ'nun daha iyi performans gösterdiği tespit edilmiştir.

Bu çalışmada, ifade edilen sınırlılıklar dahilinde kayıp veri ile çalışma alternatifi sunularak en iyi yöntem önerilerinde bulunulmuştur.

Anahtar kelimeler: Kayıp veri, Kümeleme, LVQ, SOM.

ABSTRACT

CLUSTERING APPLICATIONS IN DATA SETS WITH MISSING DATA

SEVİMLİ DENİZ, Serpil
Ph.D. Thesis, Department of Statistics
Supervisor: Prof. Dr. H. Eray ÇELİK
Fabruary, 2020, 77 pages

Missing data is a possible case in studies. Clustering analysis is a method that using to find natural groups of objects. Cluster analysis in many areas is one of the most used approaches in data analysis. In the analyzed datasets, there may be missing data at various rates. There is no definite information about which is the best in the clustering methods that can be used for the analysis of data sets containing lost data, which is completely random in numerical data. According to the number of data and data structure, each method has its advantages and deficiencies. In this study, clustering of data with continuous full and lost data is examined. Clustering analyzes were performed using k-ortalama clustering from division clustering techniques, self organization map (SOM) and learning vector quantization (LVQ) methods from artificial neural network-based clustering techniques, and the results obtained were compared. In order to make a qualitative comparison, eleven sets of real data, which are widely used in making such comparisons, were used in the literature. According to the results of the analysis, it is seen that as the loss data rate increases in all methods, the correct classification rates decrease and the SOM clustering method performs better than the k-averages and LVQ methods in seven data sets. In four data sets, it was determined that LVQ performed better depending on the structure of the data.

In this study, the best method has been proposed by presenting an alternative to working with missing data within the stated limitations.

Key words: Clustering, LVQ, Missing data, SOM.

ÖN SÖZ

Bu tez çalışmasında, her türlü ilgi ve yardımlarını esirgemeyen danışmanım, hocam Prof. Dr. H. Eray ÇELİK'e teşekkürlerimi sunarım. Tez çalışmamı tamamlamam için desteklerini esirgemeyen sayın Doç. Dr. Çağdaş Hakan ALADAĞ hocama teşekkür ederim. Ayrıca bu süreçte sabırla beni destekleyen kızlarım Şevval ve Asya'ya, eşim Suat DENİZ'e ve annem Ayşe SEVİMLİ'ye teşekkür ederim.

Varlığını hayatımın her aşamasında hissettiğim sevgili dayım A. Kadir SÖNMEZ'e ve babama…

2020

Serpil SEVİMLİ DENİZ

İÇİNDEKİLER

Sayfa

ÖZET i
ABSTRACT iii
ÖN SÖZ v
İÇİNDEKİLER vii
ÇİZELGELER LİSTESİ ix
ŞEKİLLER LİSTESİ xi
SİMGELER VE KISALTMALAR xiii
1. GİRİŞ 1
1.1. Genel Bilgiler 3
1.1.1. Kayıp veri analizi 3
1.1.1.1. Kayıp veri mekanizmaları 3
1.1.1.2. Rastgeleliğin incelenmesi 4
1.1.1.3. Kayıp veri analizinde kullanılan başlıca yöntemler 5
1.1.2. Kümeleme analizi 10
1.1.2.1. Veri türleri ve uzaklık ölçüleri 10
1.1.2.2. Kümeleme yöntemlerinin sınıflandırılması 11
1.1.3. Yapay sinir ağları ve kümeleme analizinde kullanımı 15
1.1.3.1. Nöron 17
1.1.3.2. Yapay sinir ağları türleri 18
1.1.4. Model değerlendirme 23
1.1.4.1. İçsel değerlendirme 24
1.1.4.2. Dışsal değerlendirme 26
2. KAYNAK BİLDİRİŞLERİ 31
3. MATERYAL VE YÖNTEM 35
3.1. Tezde Kullanılan Veri Setleri 35
3.2. Tezde Kullanılan Yazılımlar 36
3.3. Yöntemler 38
3.3.1. k-ortalamalar 38
3.3.2. Kümelemede kullanılan yapay sinir ağları 40
4. BULGULAR 49

Sayfa

4.1. Kümeleme Yöntemlerinin Performanslarının Gerçek Veri Setleri Üzerinde İncelenmesi 49
4.1.1. Kayıp veri içeren veri setlerinin oluşturulması 49
4.1.2. Kümeleme yöntemlerinin karşılaştırılası 50
4.1.2.1. Küçük veri setleri 50
4.1.2.2. Orta veri setleri 54
4.1.2.3. Büyük veri seti 59
5. TARTIŞMA VE SONUÇ 65
KAYNAKLAR 69
ÖZ GEÇMİŞ 77

ÇİZELGELER LİSTESİ

Çizelge **Sayfa**

Çizelge 1.1. Kümeleme yöntemlerine genel bir bakış 12

Çizelge 1.2. İstatistik terminolojisi kavramları ile yapay sinir ağı terminolojisinde kavramları 17

Çizelge 1.3. İç değerlendirme ölçütleri 25

Çizelge 1.4. Dış Değerlendirme Ölçüleri 26

Çizelge 1.5. Confusion Matrisi 28

Çizelge 1.6. Kappa İstatistiğininin Yorumlanmasına İlişkin Değer Aralıkları 29

Çizelge 3.1. Tezde kullanılan veri setleri 35

Çizelge 4.1. Tüm veri setlerine ait tüm oranlarda p değerleri 49

Çizelge 4.2. İris verilerine ait doğru sınıflandırma oranları 50

Çizelge 4.3. Göğüs kanseri verilerine ait doğru sınıflandırma oranları 51

Çizelge 4.4. Kan transfüzyon verilerine ait doğru sınıflandırma oranları 52

Çizelge 4.5. Diabet verilerine ait doğru sınıflandırma oranları 52

Çizelge 4.6. Şarap verilerine ait doğru sınıflandırma oranları 53

Çizelge 4.7. Abalone verilerine ait doğru sınıflandırma oranları 54

Çizelge 4.8. Kredi kartı verilerine ait doğru sınıflandırma oranları 55

Çizelge 4.9. Ionosphere verilerine ait doğru sınıflandırma oranları 56

Çizelge 4.10. Sensör verilerine ait doğru sınıflandırma oranları 57

Çizelge 4.11. Segment verilerine ait doğru sınıflandırma oranları 58

Çizelge 4.12. Gaz veri setine ait doğru sınıflandırma oranları 59

ŞEKİLLER LİSTESİ

Şekil **Sayfa**

Şekil 1.1. Yapay nöron yapısı 18

Şekil 1.2. Tek tabakalı YSA 19

Şekil 1.3. Çok tabakalı YSA 19

Şekil 1.4. İleri beslemeli YSA 20

Şekil 1.5. Geri beslemeli YSA. 21

Şekil 3.1. Farklı topolojiler 41

Şekil 3.2. Kohonen SOM sinir ağı 41

Şekil 3.3. Gauss fonksiyonu grafiği 44

Şekil 3.4. LVQ şeması 45

Şekil 4.1. İris verilerine ait doğru sınıflandırma oranları. 50

Şekil 4.2. Göğüs kanseri verilerine ait doğru sınıflandırma oranları. 51

Şekil 4.3. Kan transfüzyon verilerine ait doğru sınıflandırma oranları. 52

Şekil 4.4. Diabet verilerine ait doğru sınıflandırma oranları. 53

Şekil 4.5. Şarap verilerine ait doğru sınıflandırma oranları. 54

Şekil 4.6. Abalone verilerine ait doğru sınıflandırma oranları. 55

Şekil 4.7. Kredi kartı verilerine ait doğru sınıflandırma oranları. 56

Şekil 4.8. Ionosphere verilerine ait doğru sınıflandırma oranları. 57

Şekil 4.9. Sensör verilerine ait doğru sınıflandırma oranları. 58

Şekil 4.10. Segment verilerine ait doğru sınıflandırma oranları. 59

Şekil 4.11. Gaz veri setine ait doğru sınıflandırma oranları. 60

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış bazı simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Kısaltmalar	Açıklama
YSA	Yapay Sinir Ağı
SOM	Kendini Düzenleyen Haritalar
SOFM	Kendini Düzenleyen Özellik Haritaları
HKOK	Hata kareler ortalamasının karekökü
OMH-MSE	Ortalama mutlak hata
ÇKA	Çok katmanlı ağ
MCAR-TROK	Tamamen rastgele kayıp veri
MAR	Rastgele kayıp veri
NMAR	Rastgele olmayan kayıp veri
KNN	k- komşuluk
ABD	Amerika Birleşik Devletleri
BMU	Kazanan Nöron
LVQ	Doğrusal Vektör Nicemleme
MPA	Çok Katmanlı Ağ
SSE	Hata Kareler Toplamı
XOR	Veya Problemi
CPU	Merkezi İşlem Birimi
EM	Beklenti maksimizasyonu

1. GİRİŞ

Kayıp veri, veri setlerinde bir veya daha fazla değerin elde edilememesi ya da eksik olması durumudur. Doğru ve güvenilir analizler için bütün verilerin tam olması istenen bir durumdur fakat ölçülebilen-gözlemlenebilen tüm sistemlerde az veya çok kayıp veri varlığı kaçınılmazdır. Veri kaybının birçok nedeni olabilir. Kayıp değerli veri setleri pek çok istatistiksel analizde problemler oluşturmaktadır. Kayıp değerli gözlemlerin veri setinden çıkarılması, örneklem hacminin küçülmesine ve yapılan analizlerin istatistiksel gücünün azalmasına neden olmaktadır. Bu nedenle kayıp değerli gözlemlerin veri setinden çıkarılması yerine alternatif çözümler aranmıştır. Kayıp değerleri tahmin etmek için farklı yöntemler geliştirilerek kayıp değerler yerine yeni değer ataması yapılabilmektedir. Bunlara örnek olarak; tekil atama, çoklu atama (Bal, 2003), benzerlik ve tahmin fonksiyonu yöntemleri, Bayesci ve çoklu atama (Burns ve Burns, 2008), k en yakın komşu algoritması gösterilebilir (Harrington, 2012). Bu yöntemlerin hepsinde kayıp değerlerin tamamen rastgele kayıp veya rastgele kayıp olması kısıtı vardır ve kayıp değerlerin tahmin edilmesi amaçlanmıştır. Kayıp veriyi göz ardı etmek de bir alternatiftir ancak çoğu zaman önemli bir bilgi kaybına neden olacağından çok tercih edilmez.

Kümeleme analizinin amacı verileri benzerliklerine göre sınıflayarak araştırmacıya özet bilgi sağlamak ve çok fazla olan veri sayısını gruplayarak daha az sayıya indirgemektir. Kümeleme analizi veri setindeki nesneler arasında ilişki ve anlamlı şekiller bulmaya odaklandığı için bir veri modelleme problemi olarak da düşünülebilir. Kümeleme analizi ayrıca gruplanacak veriler ile ilgili önceden tanımlanan sınıflarla ve veri setiyle ilgili başlangıç bilgisine sahip olmadığı için denetimsiz öğrenme yöntemidir. Değişkenler sayısal ve kategorik olmak üzere ikiye ayrılır. Sayısal değişkenlerle kümeleme yapabilmek için genellikle uzaklık ölçüleri kullanılır. Bunlardan en sık kullanılan Öklid uzaklık ölçüsü ve Manhattan City Block uzaklık ölçüsüdür. Bölmeli temelli kümeleme yöntemleri veri setini kullanıcı tarafından belirlenmiş küme sayısına böler. Nesnel bir kriter fonksiyonunun optimize edilmesiyle oluşan kümelerin algoritması olan bölmeli temelli algoritma küme içindeki nesneler arası benzerliği maksimize eder veya uzaklığı minimize eder. K-ortalamalar algoritması

iyi bilinen bölmeli temelli algoritmalara örnek olarak verilebilir (Nemalhabib ve Shiri, 2006). Kümeleme, bir dizi öğeyi bir ölçüt temelinde bir dizi farklı gruba ayıran güçlü bir analiz aracıdır. Sayısal verilerde tamamen rastgele olan kayıp veri içeren veri setlerini modelleyen, kullanılan hangi kümeleme yönteminin en iyi olduğu ile ilgili kesin bir bilgi bulunmamaktadır. Veri sayısına ve veri yapısına göre her bir yöntemin birbirine üstünlükleri ve eksiklikleri vardır.

Bu tez çalışmasında, bölmeli kümeleme tekniklerinden k-ortalamalar ve yapay sinir ağı tabanlı kümeleme tekniklerinden öz düzenlemeli haritalar (SOM) ve doğrusal vektör parçalama (LVQ) yöntemleri kullanılarak, yöntemlerin kayıp değerli veri setlerinde uygulanması durumunda performanslarının karşılaştırılması amaçlanmıştır. Nitelikli bir karşılaştırma yapmak için literatürde bu tür karşılaştırmaların yapılmasında yaygın olarak kullanılan gerçek veri setlerinden yararlanılmıştır. İlk aşamada nümerik ve nominal veriler içeren beş küçük, beş orta ve bir büyük veri seti seçilerek tam veri setlerinde kümeleme yöntemlerinin doğru sınıflandırma oranlarına göre performansları incelenmiştir. İkinci aşamada bu veri setlerinden yüzde beş, yüzde on, yüzde on beş, yüzde yirmi, yüzde yirmi beş ve yüzde otuz oranlarında veriler tamamen rastgele eksiltilerek yeni veri setleri oluşturulmuş, bu veri setlerinde de kümeleme yöntemlerinin doğru sınıflandırma oranlarına göre performansları incelenmiştir. En iyi kümelemeyi yapan yöntemlerin sıralanması amacıyla ilk adımda elde edilen sonuçlar, doğru sınıflandırma oranlarına göre değerlendirilmiştir. İkinci adımda ise yine doğru sınıflandırma oranları eşleştirilmiş-t testi kullanılarak kümeleme yöntemlerinin birbirlerine göre durumları değerlendirilmiştir.

Çalışmanın genel bilgiler bölümünde kayıp veri analizi, kümeleme analizi ve küme doğrulama yöntemleri tüm ayrıntılarıyla anlatılmıştır. İkinci bölümde ilgili literatür incelemesi verilmiştir. Üçüncü bölümde tezde kullanılan yazılımlar ve veri setleri tanıtılarak uygulanan yöntem açıklanmış, tezde kullanılan kümeleme yöntemleri incelenmiştir. Dördüncü bölüm bulgular bölümüdür ve gerçek veri setleri ile çalışılarak k-ortalamalar, SOM ve LVQ yöntemlerinin tam ve kayıp verili veri setlerinde kümeleme performansları değerlendirilmiştir. Son olarak beşinci bölümde analiz sonuçları yorumlanarak önerilerde bulunulmuştur.

1.1. Genel Bilgiler

1.1.1. Kayıp veri analizi

Kayıp verilerin ele alınması için mümkün olan en iyi yöntem, araştırmayı iyi planlayarak ve verileri dikkatle toplayarak sorunu önlemektir (Wisniewski ve ark., 2006). Veri kaybı birçok tanımlanmış veya tanımlanmamış nedenlerden kaynaklanıyor olabilir. Analiz yapılacak veri kümesinin tam olması doğru ve güvenilir analiz için çok önemlidir. İstatistiksel analizde kayıp veri problemi önemli bir tartışma alanıdır. Bu tartışmaların başlangıç noktasını, araştırmacılar tarafından kullanılan ve yirminci yüzyılın başlarında geliştirilen analitik süreçlerin kayıpsız veri kümeleriyle yapılandırılmasına bağlamaktadır. Kayıp veri sorununun analizi için değişken sayısı ve birim sayısı çok önemlidir. Çok boyutlu verilerin analizinde, kayıp verilerin değerlendirme dışında tutulması yerine, regresyon yöntemleri ile tahmin edilmeleri daha iyi bir seçenektir (Baygül, 2007).

Kayıp verilerin varlığında, kullanılabilecek birçok istatistiksel yöntem geliştirilmiştir. Kullanılacak yöntemin seçiminde;

1) Kayıp verilerden kaynaklanan önyargıyı minimize etmek
2) Mevcut bilgilerin kullanımını maksimize etmek
3) Standart hata, güven aralığı ve olasılık değerlerinin titiz tahminleri sağlaması beklenmektedir (Allison, 2009).

1.1.1.1. Kayıp veri mekanizmaları

Kayıp verinin ortaya çıkmasında birçok etken rol oynamaktadır. Kayıp verinin minimize edilebilmesi bu etkenlerin bilinmesine, kayıp veriye yol açan nedenlerin tanımlanmasına ve anlaşılmasına bağlıdır. Bu etkenlerin en önemlileri;

- Hatalı veya eksik tasarım: Kayıp verinin bir nedeni araştırma tasarımında yapılan hatalardır. Tüm verinin sağlıklı toplanabilmesi her zaman mümkün olmayabilir. Etkin tasarımlar daha az kayıp veri oluşturur.
- Cevaplayıcının kişisel özellikleri: Cevaplayıcının soruyu algılayamaması veya cevaplamayı reddetmesi kayıp veri nedenidir.
- Ölçüm araçlarının özellikleri: Ölçüm cihazlarının hatalı çalışması veya ölçüm aracını kullanan kişinin yeteneksizliği kayıp veri için bir etkendir.

- Veri toplama ortamlarından kaynaklı olumsuzluklar: Cevaplayıcıların bulunduğu ortamdaki hava şartları, gürültü, tehlike, o anki ruh hali verilecek cevapları etkiler.
- Veri yönetiminin başarısızlığıdır: Zayıf veri yönetimi sonucunda veri depolama araçlarının kısmen veya tamamen bozulması ile ham veri kaybedilebilir. Kullanılan yazılımlardan ve kullanıcı kaynaklı hatalardan dolayı da kayıp veri olabilir (Akpınar, 2014).

1.1.1.2. Rastgeleliğin incelenmesi

Little (1988) tarafından önerilen MCAR testi, bağımsız iki örneklem t testi ve Pearson korelasyon katsayısı genellikle rastgelelik analizi için kullanılır.

1-Bağımsız iki örneklem t testi

Kayıp verilerin başka değişkenden etkilenip etkilenmediğini test etmek için bağımsız iki örneklem t testi kullanılır. Bu testte önce, kayıp gözlemlere sahip birimlere ait olan değişkenlerin değerleri birinci grup ve diğer değişken değerler ikinci grup olarak kabul edilir. Gruplar arasında istatistiksel olarak anlamlı fark bulunursa, kayıp gözlemlerin etkisinin incelenen değişken ve kayıp veri işleminden kaynaklandığına, rastgele olmadığına kararı verilir (Little,1988; Somasundaram ve Nedunchezhian, 2011).

2-Pearson Korelasyon Katsayısı

Rasgelelik araştırmasında kullanılan ikinci yöntem, değişkenlerdeki kayıp değerleri sıfıra ve kesin değerleri bire kodlayarak değişkenler arasındaki korelasyonu hesaplamaktır. Hesaplanan korelasyon matrisindeki ilişkiler yüksekse, kayıp veri mekanizmasının rastgele olmadığı söylenir. Little (1988)'de rastgeleliği incelediğinde t testi yaklaşımının çok değişkenli formunu önermiştir (Sinharay ve ark., 2001).

3- Little'ın MCAR testi (Little's MCAR test)

Veri setindeki kayıp verimliliğini belirlemek için az sayıda çalışma yapılmıştır. Bu, Little tarafından geliştirilen en yaygın kullanılan MCAR testidir. Testin varsayımı, veri kümesinin MCAR yapısına sahip olmasıdır. Bu varsayım altında, Little çok değişkenli ki-kare testi geliştirdi. Bu testin gerçekleştirilmesi üç aşamada özetlenebilir (Somasundaram ve Nedunchezhian, 2011).

1) Tahmin edilen ortalama ve varyans- kovaryans matrisleri EM (Beklenti-Maksimizasyon) yaklaşımı kullanılarak belirlenir (Allison, 2009).
2) Kayıp veriye göre gözlemler farklı yapılar olarak gruplandırılmakta ve her bir grup için gözlemlerin ortalaması hesaplanmaktadır (Baygül, 2007).
3) Gözlenen ve tahmini ortalama değerler arasındaki fark alınır. Bu fark, tahmin edilen varyans-kovaryans matrisi ve gruplardaki gözlem sayısı kullanılarak ağırlıklandırılmıştır. Sonuç olarak, yaklaşık orana karşılık gelen bir istatistik elde edilir. Son olarak, varsayım elde edilen istatistiklere karşı test edilir (Bal, 2003).

1.1.1.3. Kayıp veri analizinde kullanılan başlıca yöntemler

Kayıp veri mekanizmasına baktıktan sonra, kayıp verilerin nasıl değiştirileceğine karar verebilirsiniz. Bu amaçla kullanılan silme ve atama yöntemleri vardır (Carpenter ve ark., 2015).

1) Silme yöntemleri

a) Listwise silme yöntemi

Bugüne kadar kayıp verilere en yaygın yaklaşım, bu vakaları veri setinden çıkarmak ve kalan verileri analiz etmektir. Bu yaklaşım, tam vaka (veya mevcut vaka) analizi veya listesel olarak silinmesi olarak bilinir. Listwise silme, kayıp verilerin işlenmesinde en sık kullanılan yöntemdir ve bu nedenle çoğu istatistik yazılım paketlerinde analiz için varsayılan seçenek haline gelmiştir. Bazı araştırmacılar, parametrelerin tahmininde yanlılık gösterebileceği konusunda ısrar ediyorlar. Ancak, MCAR varsayımı karşılanırsa, listesel olarak silinmenin tarafsız tahminler ve muhafazakâr sonuçlar üreteceği bilinmektedir. Veri MCAR varsayımı yerine olmadığında, listwise silme parametreleri tahminlerinde yanlılığa neden olabilir (Donner, 1982; Rahman ve Davis, 2013).

Yeterli büyüklükte bir örnek varsa ve MCAR'ın kabul edildiği durumlarda, listenin silinmesi makul bir strateji olabilir. Ancak, büyük bir örneklem yoksa veya MCAR'ın varsayımı karşılanmadığında, listesel olarak silme işlemi optimal strateji değildir (Rahman ve Davis, 2013).

Bu yöntemin negatif yönü, kayıp gözlemlere sahip birimlerin veya değişkenlerin kaldırılmasından kaynaklanan bilgi kaybıdır. Tamamlanmamış gözlemlerin bir değişken

veya değişken bir grupta yoğunlaşmadığı durumlarda ve silme işleminin birim bazında gerçekleştirildiği ünitelerde, örnek boyutu azalacak ve silme işlemi yetersiz kalacaktır (Steyerberg, 2008).

b) Çiftler temelinde silme yöntemi

Çift yönlü silme, bilgileri yalnızca belirli bir varsayımı test etmek için gereken belirli veri noktası kayıp olduğunda ortadan kaldırır. Veri kümesinde başka yerde kayıp veri varsa, var olan değerler istatistiksel testlerde kullanılır. Bir ikili silme gözlemlenen tüm bilgileri kullandığı için, herhangi bir kayıp verilerle davayı silebilir listwise silinmesi, daha fazla bilgi korur. Bu yaklaşım, aşağıdaki problemleri sunmaktadır (Carpenter ve ark., 2015).

1) Modelin parametreleri, örnek büyüklüğü ve standart hatalar gibi farklı istatistiklerle farklı veri kümeleri üzerinde duracaktır
2) Pozitif analiz olmayan bir inter korelasyon matrisi üretebilir, bu da ileri analizleri önleyebilir (Kim ve ark., 1977).

Çift yönlü silme MCAR veya MAR verileri için daha az yanlı olarak bilinir ve uygun mekanizmalar ortak değişkenler olarak dahil edilir. Ancak, birçok kayıp gözlem varsa, analiz kayıp olacaktır. Çiftler temelinde silme yöntemi, eldeki tüm olası verilerin kullanılması yöntemi olarak da bilinir. Bu yöntemde kayıp veriler dikkate alınmayarak verinin mümkün olan kısmıyla analiz yapılır. Veri setindeki değişkenler silinmez. Sadece kayıp veriler dikkate alınmaz (Durant, 2005).

2) Atama yöntemleri

Kayıp değerleri tahmin etmek için alternatif yöntemler vardır. Bu yöntemlerden biri de atama yöntemleridir. Bu yöntemler ilk bakışta uygun metotlar olmasına rağmen, analiz sonuçları üzerindeki etkileri nedeniyle dikkatli kullanılmalıdırlar. Atama yöntemleri gözlemlerin alt kümelerinden elde edilen bilgiler yardımıyla, kayıp verilerin tahmini olarak tanımlanabilir (Batista ve Monard, 2003).

a) Ortalama değer atama

Ortalama bir değişimde, bir değişkenin ortalama değeri, aynı değişken için kayıp veri değerinin yerine kullanılır. Bu, araştırmacıların toplanan verileri kayıp bir veri setinde kullanmalarına izin verir (Durant, 2005). Ortalama atamanın teorik arka planı, ortalamanın, normal dağılımdan rastgele seçilen bir gözlem için makul bir tahmin olmasıdır. Ancak, kesin olarak rastgele olmayan, özellikle farklı değişkenler için kayıp

değerlerin sayısında büyük bir eşitsizlik varlığında kayıp olan değerlerle, ortalama atama metodu tutarsız yanlılığa yol açabilir. Dahası, bu yaklaşım yeni bilgi eklemez, ancak örneklem büyüklüğünü artırır ve hataların tahmin edilmesine yol açar. Böylece, ortalama atama genellikle kabul edilmez (Sijtsma ve ark., 2003). Bu yöntemin esası, kayıp gözlemlerin yerine gözlenen değerlerinin ortalamasını kullanmaktır. Parametre tahminindeki hatalar için dezavantajlı olmasına rağmen, standart sapma değeri azalacaktır. Bu değişkenler arasında daha düşük korelasyon katsayıları demektir (Sinharay ve ark., 2001; Allison, 2009).

b) En Çok Benzer Birime Benzetme

Kayıp gözlemin bulunduğu üniteye en çok benzeyen üniteler belirlenir ve bu ünitenin bu değişken için aldığı değer, kayıp gözlemin tahmini olarak kabul edilir. Bu yöntemin kullanılması için, kayıp gözleme benzeyen ünite dikkate alınmalıdır. Kayıp gözlem yerine, birimin en fazla gözlemlenme değerinin kullanılması durumunda aynı verinin birden fazla kullanılması söz konusu olabilir (Carpenter ve ark., 2015).

c) Regresyon

Regresyon analizi ile iki veya daha fazla değişken arasındaki ilişki ölçülür. Analiz, tek bir değişken kullanılarak gerçekleştirilirse, tek değişkenli regresyon analizi; birden fazla değişken kullanılıyorsa, çok değişkenli regresyon analizi olarak adlandırılır (Alpar, 2011). Kayıp değeri olan herhangi bir veriyi silmek yerine, bu yaklaşım, kayıp verileri diğer mevcut bilgiler tarafından tahmin edilen olası bir değerle değiştirerek tüm verileri korur. Tüm kayıp değerler bu yaklaşımla değiştirildikten sonra, veri seti kayıpsız bir veri için standart teknikler kullanılarak analiz edilir. Regresyon atamada, mevcut değişkenler bir tahmin yapmak için kullanılır ve daha sonra tahmin edilen değer, gerçek bir elde edilen değerle değiştirilir (Karabulut ve Alpar, 2011). Bu yaklaşımın birtakım avantajları vardır, çünkü bu uygulama, liste halinde veya çift yönlü silme üzerinde büyük miktarda veriyi muhafaza eder ve standart sapmayı veya dağılımın şeklini önemli ölçüde değiştirmekten kaçınır. Regresyon analizinin temel amacı, bağımlı değişken değerlerini bir veya daha fazla bağımsız değişken aracılığıyla tahmin etmektir. Regresyon ile veri atama metodunda, kayıp gözlemleri olan değişken bağımlı değişken olarak kabul edilirken, kayıp gözlemleri ve kesin gözlemleri tahmin etmek için kullanılacak olan veri setindeki diğer değişkenler bağımsız değişkenler olarak kabul edilir (Rahman ve Davis, 2013).

d) EM (Expactation-Maksimizasyon- Maksimum olabilirlik)

Kayıp verileri işlemek için maksimum olasılık yöntemini kullanan birkaç strateji vardır. Bunlarda gözlemlenen verilerin çok değişkenli normal dağılımdan alınan bir örnek olduğu varsayımı vardır (Topuz ve Çakır, 2003). Parametreler mevcut veriler kullanılarak tahmin edildikten sonra, kayıp veriler, daha önce tahmin edilen parametrelere dayanarak tahmin edilir. Kayıp veri olduğunda, değişkenler arasındaki ilişkileri açıklayan istatistikler, maksimum olabilirlik metodu kullanılarakhesaplanabilir. Yani, kayıp olan veriler diğer değişkenlerin koşullu dağılımı kullanılarak tahmin edilebilir (Czepiel, 2002).

Beklenti-Maksimizasyon (EM), tüm kayıp değerlerin maksimum olabilirlik metotları ile tahmin edilen değerlerle ifade edildiği yeni bir veri seti oluşturmak için kullanılabilecek maksimum olasılık yönteminin bir türüdür (Dempster ve ark., 1997). Bu yaklaşım, parametrelerin (örneğin, varyanslar, kovaryanslar ve araçlar) tahmin edildiği, belki de listesel olarak silinmeyi kullanan beklenti adımı ile başlar. Butahminler daha sonra kayıp verileri tahmin etmek için bir regresyon denklemi oluşturmak için kullanılır. Maksimizasyon adımı, kayıp verileri doldurmak için bu denklemleri kullanır. Beklenti adımı yeni parametrelerle tekrarlanır, burada yeni regresyon denklemleri kayıp verileri ”dolduracak” şekilde belirlenir (Fiona ve ark., 2006). Ancak, beklenti maksimizasyonunun bazı dezavantajları vardır. Bu yaklaşımın, özellikle de kayıp verilerin büyük bir kısmı olduğu zaman, yakınsak olmak için uzun zaman alabilir. Bu yaklaşım, önyargılı parametre tahminlerine yol açabilir ve standart hatayı hafife alabilir. Beklenti maksimizasyon yöntemi için, her bir durum için mevcut olan değişkenlere dayanan tahmin edilen bir değer, kayıp verilerin yerine konur (Topuz ve Çakır, 2003).

e) Çoklu Atama Yöntemi

Kayıp verilere atama yapmak için farklı bir stratejidir. Birden fazla denemede, kayıp olan her veri için tek bir değerin yerini almak yerine, kayıp değerler bir dizi makul değer ile değiştirilir (Bal, 2003).

Bu yaklaşım, diğer değişkenlerde var olan verileri kullanarak kayıp verilerin tahminiyle başlar. Kayıp değerler daha sonra tahmin edilen değerlerle değiştirilir ve atanan veri kümesi adı verilen tam bir veri kümesi oluşturulur (Burns ve Burns, 2008). Çoklu atamanın yararı, kayıp değerlerin doğal değişkenliğini geri getirmenin

yanı sıra, kayıp verilerden kaynaklanan belirsizliği de içermesidir ki bu da geçerli bir istatistiksel çıkarımla sonuçlanır. Çoklu atamanın, kayıp verilerin tahmini ile ilgili belirsizliği yansıtan geçerli istatistiksel çıkarsama ürettiği gösterilmiştir. Çoklu atamanın, normallik varsayımlarının ihlaline karşı sağlam olduğu ve küçük bir örneklem büyüklüğünün veya çok sayıda kayıp verinin varlığında bile uygun sonuçları ürettiği ortaya çıkmaktadır (Demir, 2013). Bu yöntem ile iki veya daha fazla atama yöntemiberaber kullanılarak kayıp gözlemler tahmin edilmeye çalışılır. Bu tahminleme metodu elde edilen tahmin değerlerinin ortalaması alınarak kullanılır. Çoklu atama yönteminin bir takım avantajları vardır. En önemlisi, anlaşılabilir olmasıdır. Aynı zamanda alan değişkenlerinin normalliği ihmal edildiğinde analiz değeri güçlü sonuçlar verir. Dezavantajı ise, atama sürecinin uzun sürmesidir (Elmas, 2010).

f) Son Gözlemin İleriye Taşınması

En yaygın kullanılan yöntemlerden biri de, ileriye taşınan son gözlemdir. Bu yöntem, her kayıp değeri, aynı konuda en son gözlemlenen değerle değiştirir (Hamer ve Simpson, 2009).

Bu yöntemin, farklı zamanlarda elde edilen sonuçlara göre bir veri seti oluşturulmasında belirli bir mantığı vardır. Bununla birlikte, sonuç açısından tartışmalı bir yaklaşımdır (Little, 1987).

g) Naive Bayes Yöntemi ile Değer Atama

Naive Bayes metodu ile değer atama, Bayes karar teorisine dayalı basit bir olasılık sınıflandırıcıdır. Her bir sınıf için olasılıklar hesaplanır ve her örnek için olasılığı en yüksek bulma eğilimi vardır. İyi bir performans, basit yapı, yüksek hesaplama hızı ve kayıp verilere karşı duyarsız olması bu metodu popüler kılmaktadır (https://www.saedsayad.com/naive_bayesian.htm).

h) K En Yakın Komşu (KNN) Metodu ile Kayıp Değer Atama Yöntemi

KNN algoritması, en yakın komşu (KNN) ile kayıp değer atama yönteminin temelidir. Bu algoritmada; bir gözlemdeki gözlemlerden biri, gözlemlerin her birinin belirli bir gözlem değerine göre mesafesini hesaplayarak ve en küçük k gözlemlerini seçerek elde edilir (Harrington, 2012). KNN algoritmasında, mesafe hesaplamaları için farklı fonksiyonlar belirlenmiştir. Örnekler arasında Öklid (Öklid), Manhattan ve Minkowski uzaklık fonksiyonları bulunmaktadır. Bu, mesafe fonksiyonlarında en yaygın Öklid mesafe fonksiyonudur. Veri kümesinin ikiden fazla değişken içerdiği

durumlarda, standart Öklid uzaklık fonksiyonu kullanılır (Bridge, 2013). Her değişken, z dönüşümü uygulanarak standardize edilir. Böylece değişkenler arasındaki ölçüm farkları ortadan kaldırılır. Bu algoritmayı kurarken, k değeri seçimine dikkat edilmelidir. Eğer komşuluk değeri küçük bir sayı ise, baskın gözlemlerin aşırı vurgulanması nedeniyle atamayönteminin tahmin performansında bir bozulma meydana gelebilir (Xia, 2013). Öte yandan, büyük bir komşuluk değeri, tahmin sürecindeki kayıp değerli gözlemden oldukça farklı olan ve sonuç olarak tutarsız değerler üretecek gözlemlere yol açabilir. Sonzamanlarda, KNN algoritmasını kullanan yeni bir yöntem, kayıp veri sorunlarının çözümünde çok etkili olmuştur (Bal, 2003; Xia, 2013).

1.1.2. Kümeleme analizi

Kümeleme analizi, nesnelerin veya gözlemlerin benzerlik esas alınarak homojen gruplara ayrıldığı istatistiksel veri bölümleme yöntemlerini ifade eder (Mooi, 2011). Çok değişkenli bir veri kümesinden (gözlemler, nesneler) küme grupları oluşturmak için kullanılan çok değişkenli bir yöntemdir (Alkarkhi, 2019). Bu nedenle kümeleme analizi, doğal olarak benzerlik kavramına bağlıdır. Her gruptaki gözlemler birbirine benzerdir, ancak kümelerin kendileri birbirinden farklıdır. Verileri benzerliğe göre gruplara ayırmak için birçok algoritma vardır ve bunlar önemli ölçüde farklıdır. Kümeleme analizi, bir tür veri azaltma tekniğidir (Alkarkhi, 2019). Kümeleme, sınıflandırmadan farklı olarak bilinen bir sınıf etiketi olmadan verileri analiz eder. Veriler arasında küme içi benzerliği en üst düzeye çıkarmak ve kümeler arası benzerliği en aza indirmek amaçlanır (Grabmaier ve Rudolph, 2002).

1.1.2.1. Veri türleri ve uzaklık ölçüleri

Kümeleme analizinde sıklıkla kullanılan bazı uzaklık ölçülerinin formülleri şöyledir;

Öklid uzaklığı: Öklid uzaklığı standartlaştırılmış veriler yerine, işlenmemiş ham verilerde kullanılır. Bu uzaklık türü, kümeleme analizinde aykırı değerlerden çok etkilenmezken, verilerin farklı ölçeklerden elde edilmiş olmasından önemli ölçüde etkilenmektedir. En büyük değerlere sahip olan değişken Öklid uzaklığını maksimize etme eğilimindedir.

İki birim arasındaki uzaklık; p değişken sayısı olmak üzere: i ve j biriminin birbirine olan uzaklığı;

$$d(i,j)=\left[\sum_{k=1}^{p}(x_{ik}-x_{jk})^2\right]^{\frac{1}{2}} \tag{1.1}$$

formülü ile hesaplanır.

Minkowski uzaklığı:

$$d(i,j)=\left[\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right|^m\right]^{\frac{1}{m}} \tag{1.2}$$

formülü ile hesaplanır ve genel bir formdur.

Mahalanobis uzaklığı:

Bir çeşit Öklid uzaklığı ölçüsüdür. Standart sapma değerlerini dikkate alarak standardizasyon olanağı sağlar. Grup içi varyans- kovaryanslar toplanarak değişkenler arasındaki korelasyon ayarlanır.

$$d(i,j)=(x_i-x_j)'S^{-1}(x_i-x_j) \tag{1.3}$$

Burada S kovaryans matrisini göstermektedir.

Pearson korelasyon ölçüsü

En eski bilinen uzaklık ölçüsüdür. Sürekli iki değişken arasındaki doğrusal ilişkiyi gösterir.

$$d(i,j)=\sqrt{\frac{\sum_{k=1}^{p}(x_{ik}-x_{jk})^2}{S_k^2}},\quad i,j=1,2,...,n \tag{1.4}$$

1.1.2.2. Kümeleme yöntemlerinin sınıflandırılması

Bazı kümeleme yöntemleri, birkaç kümeleme yönteminin fikirlerini bütünleştirir. Verilen bir algoritmayı sadece bir kümelenmeye özgü olarak sınıflandırmak zordur.

Çizelge 1.1. Kümeleme yöntemlerine genel bir bakış (Han ve Kamber, 2001)

Metot	Genel Karakteristikleri
Bölümleme Yöntemleri	- Mesafeye dayalı - Küme merkezini temsil etmek için ortalama veya medoid (vb.) kullanabilir - Küçük ve orta boy veri kümeleri için etkili
Hiyerarşik Yöntemler	-Kümeleme hiyerarşik bir ayrıştırmadır - Hatalı birleştirme veya bölmeleri düzeltemiyorum - Mikro kümeleme veya başka teknikler içerebilir.
Yoğunluk Bazlı Yöntemler	-Keyfi şekilli kümeleri bulabilir - Kümeler, uzayda düşük yoğunluklu bölgelere ayrılmış olan nesnelerin yoğun bölgeleridir. - Küme yoğunluğu: Her nokta minimum sayıda olmalıdır - Aykırı değerleri filtreleyebilir
Izgara tabanlı Yöntemleri	-Çok çözünürlüklü bir ızgara veri yapısı kullanır - Hızlı işlem süresi (genellikle sayısından bağımsız veri nesneleri, ancak ızgara boyutuna bağlı)
Model Tabanlı Yöntemler	-Her bir kümeyi model sayar. Optimım modeli arar.

a) Bölümleme yöntemleri

Bölümleme yöntemleri doğrudan tek bir bölümünü elde etmeyi amaçlar. Bu yöntemlerin çoğu veri ile bölüm arasındaki "anlaşmayı" yansıtan bir kriter fonksiyonunun yinelemeli optimizasyonuna dayanır (MacQueen, 1967).

-k-ortalamalar

Kümeleme algoritmalarının çoğunluğu, her bir gözlemin sadece bir kümeye ait olabileceği anlamına gelen özel kümeler üretir. En basit ve en etkili örtüşen kümeleme yöntemlerinden biri, geleneksel k-ortalamalar algoritmasının bir uzantısı olan örtüşen k-ortalamalar olarak bilinir (Khanmohammadi ve ark., 2017).

-k- medoid kümeleme yöntemi

k -ortalamalar yöntemine çok benzemekle birlikte bu yöntemin aykırı değer ve gürültülere karşı olan hassaslığını azaltmak için küme merkezleri yerine, kümede ortaya en yakın noktada konumlanmış olan nesne anlamındaki medoid kullanılmaktadır. k-medoid kümeleme yönteminde süreç, k kümeleri temsil etmek için başlangıçtaki medoidlerin rasgele seçilmesi ile başlar. Geride kalan diğer tüm nesneler, medoidleri kendilerine en yakın olan bir kümeye dâhil edilir. Bundan sonra kümeyi daha iyi temsil edebilen yeni bir medoid belirlenir. Geriye kalan tüm nesneler, yine en yakın medoide sahip kümelere atanır. Her yinelemede, medoidler yerlerini değiştirir. Metot, her bir nesne ile ilgili medoid arasındaki farklılıkların toplamını en aza indirir. Bu döngü, medoidin yerleşimini değiştirinceye kadar tekrarlanır. Bu işlemin sonunu gösterir ve sonuçta ortaya çıkan nihai kümelenmelerin medoidleri tanımlanır. Kümeler medoidler

etrafında toplanmış ve tüm nesneler en yakın medoide dayalı olarak uygun kümeye yerleştirilmiş olarak tamamlanır (Işık ve Çamurcu, 2007).

b) Hiyerarşik kümeleme yöntemleri

Hiyerarşik kümelemede, veriler birkaç adımda daha küçük sınıflara bölünür veya tersine küçük sınıflar aşamalı olarak daha büyük olana kadar birleştirilir. Başlangıçta veri setinde kaç grup bulunduğu bilinmediği durumlarda uygun bir yöntemdir. Hiyerarşik kümeleme prosedürleri, analiz sırasında kurulan ağaç benzeri yapı ile karakterize edilir. "Ayırıcı ve gruplayıcı" olmak üzere iki yöntemi vardır.

-Ayırıcı hiyerarşik kümeleme yöntemi

Bu yöntem tüm nesnelerden oluşan büyük bir setle başlar. Daha küçük kümeler oluşturmak için benzer kümeler çıkarılır. Bu parçalanma süreci, her küme tutarlı bir alt kümeye bölünene kadar devam eder.

-Gruplayıcı hiyerarşik kümeleme yöntemi

Kümelerdeki veriler ilk olarak, en benzer iki küme (başka bir deyişle, aralarındaki en küçük mesafe olanlar), hiyerarşinin en altında yeni bir küme oluşturmak üzere birleştirilir. Bir sonraki adımda, başka bir çift küme birleştirilir ve daha üst düzey bir hiyerarşiye bağlanır ve algoritma böylece devam eder. Bu yaklaşımdaki kümeleme analizi için her aşamada hangi kümelerin birleştirileceğini belirlemek için kullanılan çeşitli yöntemler vardır.

Temel yöntemler aşağıda özetlenmiştir:

Tek Bağlantı Kümeleme (Single Linkage Method, Nearest Neighbours Method): Bu yöntemde, birbirine en yakın iki birim birleştirilir ve birleştirme işlemi bütün birimler herhangi bir kümede toplanıncaya kadar devam eder. Küme yapısını hesaba katmaz ve zincirleme olarak adlandırılan, kümelenmelerin uzun ve seyrek olmasıyla sonuçlanan bir probleme neden olabilir.

Tam Bağlantı Kümeleme (Complete Linkage Method, Furthest Neighbours Method): Birimler arasında en uzak mesafe değerleri esas alındığından tek bağlantılı yöntemin tam tersi bir yöntem olarak bilinir. Benzer boyutta küçük kümeler üretme eğilimindedir; ancak, en yakın komşu yönteminde olduğu gibi, küme yapısını da hesaba katmaz. Aykırı değerlere de oldukça hassastır.

Ortalama Bağlantı Kümeleme (Average Linkage Method): Tek ve tam bağlantılı tekniklerin aksine kümedeki tüm benzerliği ele alan alternatif yöntemdir (Özdamar, 2002).

Küresel Ortalama Bağlantı Kümeleme (Centroid Method): Ortalama bağlantı kümeleme yönteminin özel bir hali olan küresel ortalama bağlantı kümeleme yönteminde, her kümenin centroidi (her değişkene ilişkin ortalama değer) hesaplanır ve centroidler arasındaki mesafe kullanır.

Ward Bağlantı Kümeleme (Ward's Method): Hiyerarşik kümelemede yaygın olarak kullanılan bir diğer yaklaşım da Ward'ın yöntemidir. Temel olarak, mesafe veya ilişki ölçümlerini kullanmak yerine, küme analizine "varyans analizi" analizi olarak bakmaktadır. Bu yaklaşım, en benzer iki değişkeni birbiri ardına birleştirmez. Bunun yerine, birleşmesi genel küme içi varyansı mümkün olan en küçük seviyede yükselten nesneler birleştirilir (Alpar, 2011).

c) Yoğunluk tabanlı metotlar

Kümeleme analizinde önemli dar boğazlardan birisi, geliştirilen algoritmaların sadevce küresel biçimlerde dağılan veri noktaları için etkin sonuçlar verebilmesidir. Ancak sayısız gerçek dünya olayı, mekânsal verinin yönetimini ve dolayısı ile keyfi biçime sahip kümelerin keşfedilmesini gerektirmektedir. Bu amaçla geliştirilen algoritmalardan bir kısmı veri noktalarının belirli mekanlarda yoğunlaşması üzerine odaklanılmıştır. Bu yaklaşımla geliştirilen algoritmalar yoğunluk temelli kümeleme olarak isimlendirilmişlerdi (Brecheisen ve ark., 2003).

d) Fuzzy clustering

Klasik kümelemede, farklı kümelerin sınırları nettir. Her bir model tam olarak bir sınıfa atanır. Öte yandan, kümeler arasındaki sınırlar, gerçek hayatta kesin olarak tanımlanamayabilir, öyle ki, bazı veriler farklı derecelere sahip birden fazla kümeye ait olabilir. Bu durum net kümelenme yerine bulanık kümelenme ile temsil edilmektedir (Hoppner ve ark.,1999; Dumitrescu ve ark., 2000). Fuzzy kümeleme temelde bir nesne ile bir küme arasındaki bağlantı veya üyelik düzeyi değeri hesaplanmasına dayanır (Akpınar, 2014).

e) Izgara tabanlı metotlar

Izgara tabanlı yöntemlerde çalışma uzayı sonlu sayıda hücrelere bölünerek, çok boyutlu bir ızgara yapısı oluşturulur. Mekansal veri analizlerinde kullanılan bir

algoritmadır. STING ızgara temelli bir kümeleme algoritmasıdır (Wang ve Yang, 1997). Bu algoritmaya göre mekânsal alan dikdörtgensel hücrelere bölünür. Bu hücreler farklı tabakalarda ve büyüklüklerde olup, hücreler tabakalarda hiyerarşik bir yapı oluşturmaktadır. Bu algoritma ızgara stilinde tanımlanan hücrelerden sağlanan bilginin sorgulanmasına olanak sağlamaktadır (Akpınar, 2014).

f) Model tabanlı yöntemler

Model tabanlı yöntemler, kümelerin her biri için bir model varsaymakta ve en iyisini bulmaktadır. Model tabanlı bir algoritma veri noktalarının uzamsal dağılımını yansıtan bir yoğunluk fonksiyonu oluşturur ve küme sayısını otomatik olarak belirler. Model tabanlı kümeleme yöntemlerinde iki ana yaklaşım vardır bunlar istatistiksel yaklaşım ve yapay sinir ağı yaklaşımıdır. İstatistiksel yaklaşım örnekleri arasında COBWEB yer almaktadır. Yapay sinir ağı yaklaşımına da özdüzenleyici haritalar (SOM) ve doğrusal vektör parçalama (LVQ) yöntemleri örnek verilebilir.

1.1.3. Yapay sinir ağları ve kümeleme analizinde kullanımı

Yapay sinir ağları, biyolojik sinir ağlarını taklit eden sentetik ağlar olarak tanımlanabilir (Günay ve ark., 2007). Biyolojik sinir sistemlerinden ilham alan yapay sinir ağları, veriler arasındaki ilişkiyi öğrenme, saklama ve açığa çıkarma yeteneğine sahiptir. Biyolojik sinir sistemleri nöronlardan oluşur. Yapay sinir ağları da benzer şekilde yapay nöronlardan oluşur. Ağdaki nöronlar bir veya daha fazla girdi alır ve tek bir çıktı üretir. Bu çıktı, yapay sinir ağında bir çıktı olabilir veya diğer nöronlara girdi olarak kullanılabilir. Geliştirilen modellerde farklılıklar olsa da, genel olarak her işlem elemanı beş ana bileşenden oluşur: girdiler, ağırlıklar, toplama fonksiyonu, aktivasyon fonksiyonu ve çıktı (Öztemel, 2003).

Milyonlarca nöron beynin bağlantılar yoluyla birleşmesinden oluşur. Başka bir deyişle, tabakalar aynı yönde bir araya gelen nöronlar tarafından oluşturulur ve bu birkaç tabakanın birleştirilmesiyle yapay sinir ağları oluşturulur. Bu yapının bazı nöronları, girişleri almak için dış alana ve çıkışları iletmek için diğerlerine bağlanır. Diğer tüm nöronlar gizli tabakalardır. Bu durumda, tabakaların farklı biçimlerde birbirine bağlanması farklı ağ mimarilerinin ortaya çıkmasına neden olur. Yapay sinir ağlarının ilk yıllarında, bazı araştırmacılar rastgele nöronlar arasındaki bağlantıları kurmuş ve olumsuz sonuçlarla karşılaşmışlardır (Yurtoğlu, 2005).

Bir yapı tasarlamanın en kolay yolu, elemanları yerleştirmektir. Nöronların kombinasyonu rastgele olamaz. Genel olarak, giriş tabakası, gizli tabaka veya ara tabaka ve çıkış tabakası olan üç tabakada bir araya gelirler. Ağı oluşturmak için paralel olarak bağlanırlar (Öztemel, 2003).

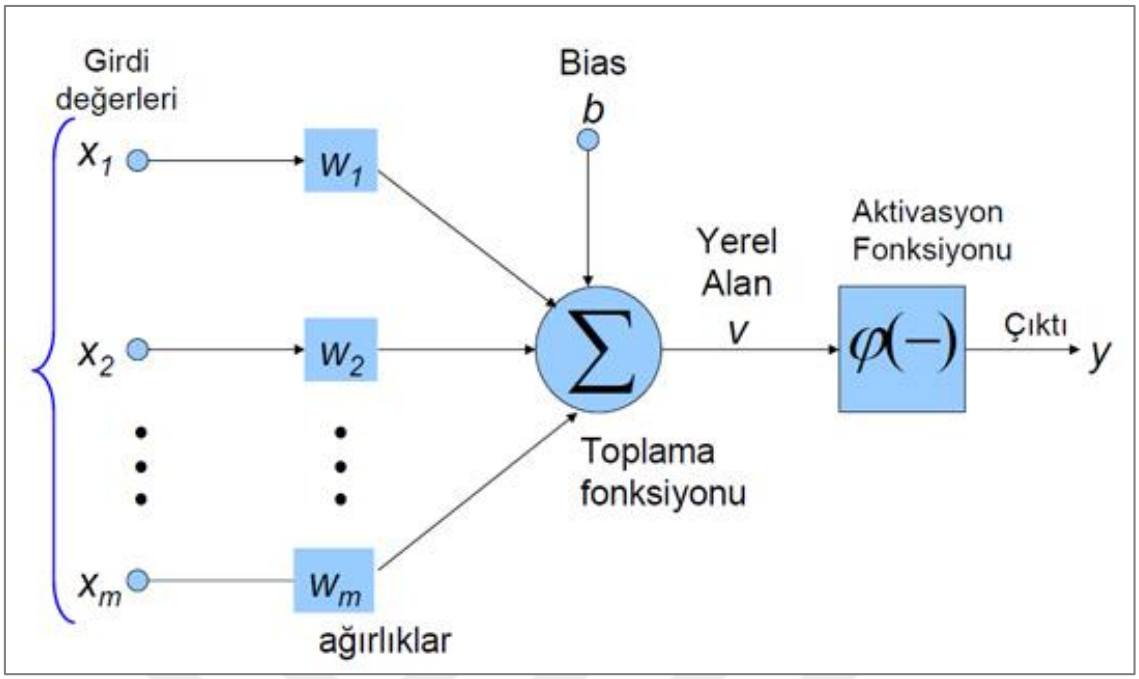
Klasik istatistik metotlarda kullanılan matematik modeller ile yapay sinir ağlarında kullanılan öğrenme algoritmaları birbirine benzer (White, 1989). İki alanda aynı temalar farklı isimler altında ele alınmaktadır. Bir nörobilimci istatistiksel bilgiye ne kadar çok ihtiyaç duyarsa, bir istatistikçi de sinir ağları bilgisine ihtiyaç duyar (Cheng ve Titterington, 1994). Birçok yapay sinir ağı modeli istatistiksel yöntemlerle uyumludur. Gizli tabakası olmayan ileri beslemeli ağlar doğrusal modellere, gizli tabakalı ileri beslemeli ağlar doğrusal olmayan regresyon modellerine, Kohonen SOM ağları küme analizine, Hebbian öğrenme temel bileşenler analizine karşılık gelir (Sarle, 1994). Geri yayılım algoritması denetimli öğrenme algoritmasıdır. İstatistiksel yöntemlerin çoğu verilerle ilgili dağılım varsayımları ve değişkenlerle ilgili varyasyon ve korelasyon varsayımları gerektirir (Neter ve ark., 1989). Yapay sinir ağları yaklaşımının klasik istatistiksel yöntemlere göre avantajı, verilerin dağılım varsayımları ve değişkenlerle ilgili varsayımlar gerektirmemesidir. Yapay sinir ağları bazı değişkenlerin kayıp verilerini tolere edebilmektedir. Bununla birlikte, sinir ağlarının eğitilmesi çok sayıda veriye ihtiyaç duyulması ve bir ağ modeli oluşturmadaki zorluklar dezavantaj olarak görülebilir (Öztemel, 2003). Çizelge1.2'de istatistik terminolojisi ile yapay sinir ağı terminolojisinde aynı kavramları belirten bazı terimler sunulmaktadır (Gunay ve ark., 2007).

Çizelge 1.2. İstatistik terminolojisi kavramları ile yapay sinir ağı terminolojisinde kavramları(Gunay ve ark., 2007)

İstatistik Terminolojisi	Yapay Sinir Ağı Terminolojisi
Model	Yapay Sinir Ağı
Parametre	Ağırlık
Bağımsız Değişken	Girdi
Tahmin Değeri	Çıktı
Bağımlı Değişken	Hedef
Artık	Hata
Güven Aralığı	Hata Çizgisi
Temel Fonksiyon	Nöron
Bağımsız Değişkenler Kümesi	Girdi Tabakası
Temel Fonksiyonlar Kümesi	Gizli Tabaka
Tahmin Değerleri Kümesi	Çıktı Tabakası
Gözlem	Örüntü
Kestirim ya da Optimizasyon	Eğitim, Öğrenme ya da Adaptasyon
Örneklem Adaptasyonu	Çevrimiçi Öğrenme
Grup Adaptasyonu	Çevrimdışı Öğrenme
Kestirim Ölçütü	Hata, Maliyet ya da Lyapunov Fonksiyonu
Dönüşüm	Fonksiyonel Bağlantı
Diskriminant Analizi	Sınıflama
Regresyon	Eşleme, Yaklaşım ya da Denetimli Öğrenme
Veri İndirgeme	Denetimsiz Öğrenme ya da Şifreleme
Genelleştirme	İnterpolasyon ya da Ekstrapolosyon

1.1.3.1. Nöron

İnsan nöronları gibi dış dünyadan gelen girdileri alıp işleyen ve çıktı olarak sunan bir yapıya sahiptir. Yapay sinir ağlarının temel bileşenleri mimari yapı, öğrenme algoritması ve aktivasyon fonksiyonudur (Aladağ, 2009). Çıktı değerlerini oluştururken tıpkı insan nöronları gibi öğrenirler. YSA'nın bütün işlemleri yapay nöronlarda gerçekleşir. Nöronların bu fonksiyonları yerine getirebilmeleri için bir takım işlem elemanları vardır. Bunlar Şekil 2.1'de gösterilmektedir.



Şekil 1.1. Yapay nöron yapısı (Güzeller ve Aksu, 2018).

1. Girdiler: Çalışmada kullanılacak verilerdir.
2. Ağırlıklar: Girdilerin her birinin bir önemi(ağırlığı) vardır. Bu önem ile girdi değerleri çarpılarak nöron üzerindeki etkisi hesaplanır. Her girdiye farklı bir değer verilir.
3. Toplama Fonksiyonu: Önem düzeyi ile çarpılarak yeniden hesaplanmış girdilerin nöron bazında toplanarak net girdinin hesaplandığı fonksiyondur.
4. Aktivasyon Fonksiyonu: Toplama fonksiyonu ile elde edilen toplam net girdi aktivasyon fonksiyonu sayesinde çıktıya dönüşür. Aktivasyon fonksiyonunun seçimi tamamen çalışmanın türüne bağlıdır.
5. Nöron Çıktısı: Aktivasyon fonksiyonu yardımıyla elde edilen çıktı değeridir. Her bir nöronun sadece bir çıktısı vardır (Güzeller ve Aksu, 2018).

1.1.3.2. Yapay sinir ağları türleri

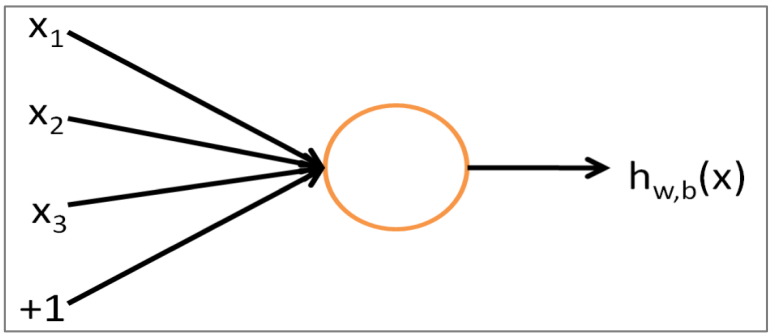
Yapay sinir ağları tabaka sayılarına, tipine ve öğrenme algoritmalarına göre üç gruba ayrılır.

1) Tabaka Sayılarına Göre YSA

Benzer özellikteki nöronların oluşturduğu yapıya tabaka denir. Tabaka sayısına göre YSA tek tabakalı ve çok tabakalı olarak sınıflandırılabilir.

-Tek Tabakalı YSA

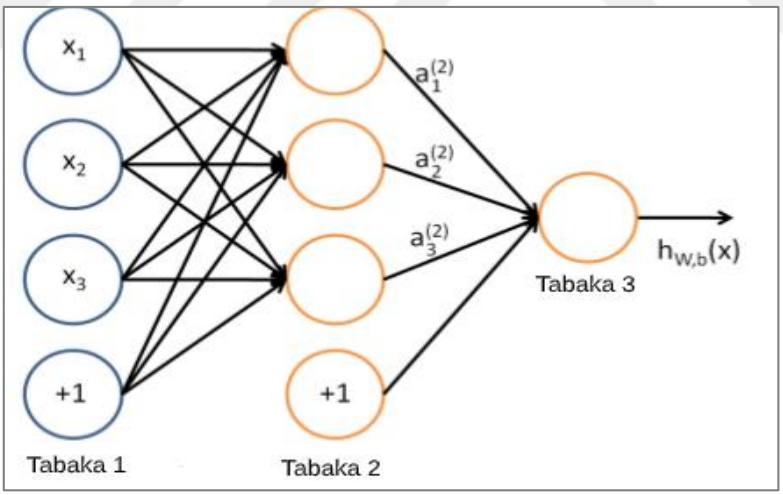
Girdi tabakası ile çıktı tabakasından oluşan şekil 1.2'deki gibi yapılara tek tabakalı YSA denir. Gizli tabaka yoktur girdi ve çıktı tabakaları vardır.



Şekil 1.2. Tek tabakalı YSA(https://tr.sciencewal.com/10226-understanding-neural-networks-from-neuron-to-rnn-cnn-and-deep-learning-cd88e90e0a90-42).

-Çok Tabakalı YSA

Çalışmanın amacına göre oluşturulan YSA girdi, gizli ve çıktı tabakalarından oluşuyorsa çok tabakalı olarak tanımlanır.



Şekil 1.3. Çok tabakalı YSA(https://tr.sciencewal.com/10226-understanding-neural-networks-from-neuron-to-rnn-cnn-and-deep-learning-cd88e90e0a90-42).

Girdi Tabakası: Girdilerin ağa sunulduğu tabakadır. Veriler herhangi bir işlem uygulanmadan sonraki tabakaya iletilir.

Gizli Tabaka: Girdi tabakasından gelen veriler gizli tabakaya iletilir. Burada aktivasyon fonksiyonu uygulanarak çıktı tabakasına iletilir. Gizli tabaka sayısı ağın yapısına göre farklılık gösterir.

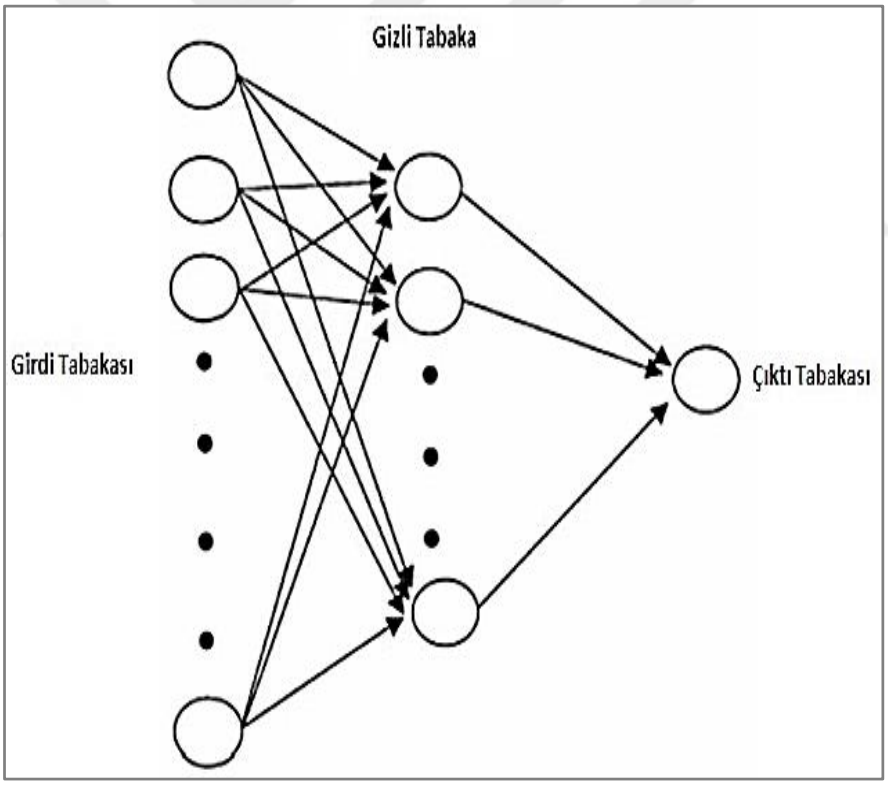
Çıktı Tabakası: Ağların yapısına bağlı olarak çıktı değerleri hesaplanarak ortama verildiği tabakadır.

2) Tipine Göre YSA

Nöronlar arası bağlantı yapılarına göre YSA ileri beslemeli ve geri beslemeli olmak üzere iki gruba ayırılmaktadır.

- *İleri Beslemeli YSA*

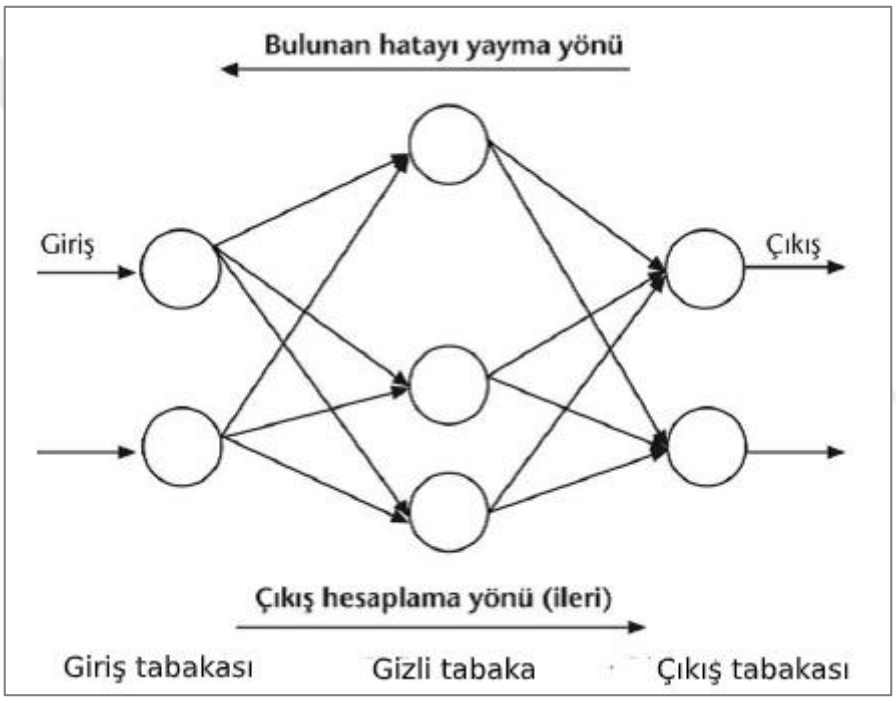
İleri beslemeli bir ağda, nöronlar tabakalara ayrılır. Her tabakadaki nöronların sonraki tabakadaki nöronlarla bağlantısı ağırlıklar ile sağlanır. Ancak, tabakalar arasında bağlantı yoktur. İleri besleme ağlarında bilgi akışı geri bildirim olmadan, girdi tabakasından çıktı tabakasına tek yönde yapılır (Hamzaçebi, 2007). Bu aktivasyon yönü olarak da tanımlanır. Bu ağlar danışmanlı öğrenme tekniği kullanılarak eğitilir. Şekil 1.4 ileri beslemeli YSA örneği görülmektedir.



Şekil 1.4. İleri beslemeli YSA (http://veribilimci.org/yapay-sinir-aglari-ysa-nedir-bolum-2).

-Geri Beslemeli YSA

Bu ağ yapısı, çıkış tabasının ve gizli tabakadaki çıkışların giriş birimlerine veya önceki ara tabakalara geri beslendiği bir ağ yapısıdır. Bu ağda, nöronların dönüş veya geri besleme bağlantıları vardır ve bu nedenle dinamik belleğe sahip oldukları söylenir. Geri bildirim ağlarında, herhangi bir nöronun çıktısı doğrudan girdi tabakasına döndürülebilir (Zhang, 2014). Geri besleme tabakalardaki nöronlar ile tabakadaki nöronlar arasında olabilir. Doğrusal olmayan dinamik davranış sergileyen geri beslemeli ağlara Hopfield ve Elman ağları örnek verilebilir. Şekil 1.5, geri bildirim yapay sinir ağına sahip bir yapıyı göstermektedir.



Şekil 1.5. Geri beslemeli YSA (Öztemel, 2003).

3) Öğrenme Algoritmalarına Göre YSA

a. Danışmanlı Öğrenme

Danışmanlı öğrenme biçiminde bir danışman, verilerin olayı öğrenmesine yardımcı olur. Danışman, sisteme giriş / çıkış seti olarak örnekler verir. Diğer bir deyişle, her bir örnek için üretilmesi gereken her iki girdi ve çıktı sisteme gönderilir. Sistemin görevi, girdileri danışman tarafından belirlenen çıktılara göre yönetmektir. Bu şekilde, olayın girdileri ve çıktıları arasındaki ilişkiler öğrenilir (Hamzaçebi, 2007; Langley, 1996).

Bu öğrenmede, ağ tarafından üretilen sonuçlar ve amaçlanan sonuçlar arasındaki farka hata denir. Bu hatanın en aza indirilmesi hedeflenir. Bu işlem, kabul edilebilir bir doğruluğa ulaşana kadar sürekli test edilir. Bu deneylerde, bağlantıların ağırlıkları en uygun çıktıyı vermek için değiştirilir (Güzeller ve Aksu, 2018). Belirtilen çıkış verilen girişe cevap vermiyorsa ağın çıkışındaki hatayı en aza indirmek için bağlantı ağırlıklarının değiştirilmesi gerekir. Bu nedenle, danışmanlı öğrenmenin bir danışmana ihtiyacı vardır. Danışman, kontrolün performansını değerlendirir ve öğrenme sürecini, performansı kademeli olarak artıracak şekilde yönlendirir. Danışmanlı öğrenme algoritmalarında, ortalama mutlak hata (OMH) ve hata karelerinin ortalamasının karekökü (HKOK) performans kriterleri kullanılır (Çanakçı, 2006).

b. Danışmansız Öğrenme

Unsupervised ya da eğiticisiz öğrenme de denilen bu öğrenmede, sistemin öğrenmesini sağlayan bir danışman yoktur. Sisteme sadece giriş değerleri gösterilmiştir. Örneklerdeki parametreler arasındaki ilişkilerin sistem tarafından öğrenilmesi beklenir. Sadece sistem tamamlandıktan sonra çıktıların ne anlam ifade ettiğini ve etiketlemeyi öğrenmek kullanıcı tarafından yapılmalıdır. Bu tür öğrenme örnekleri Hebb, Hopfield ve Kohonen öğrenme algoritmaları ve ART ağlarıdır (Mohri ve ark., 2012; Hebb, 1949; Hopfield, 1990).

Öğrenme danışmansız olduğunda çıktı değişkeni mevcut değildir. Fakat model oluşturmak için kullanılan tüm değişkenler bağımsız değişkenlerdir yani girdilerdir. Danışmansız öğrenmede amaç girdi değişkenleri ile çıktı değişkenlerini tahmin etmektir (Gorunescu, 2011). Danışmanlı öğrenmeden farklı olarak danışman olmadan model oluşturulur. Yani bir denetleme mekanizması yoktur. Kullanıcı kendi başına modeli kurar ve değerlendirir. Gelecekteki veya mevcut durumun tahmini için kullanılır. Danışmansız öğrenmeye örnek kümeleme ve birliktelik kurallarıdır (Hastie ve ark., 2008).

c. Destekleyici Öğrenme

Bu eğitim türünde, ağdaki bir eğitici yardımcı olur. Bu öğrenme yaklaşımında, ağın her yinelemesinin sonucunun iyi ya da kötü olup olmadığı hakkında bilgi verilmektedir. Ağ bu bilgilere göre kendini yeniden düzenlemektedir. Bu şekilde ağ, herhangi bir giriş sırasına sahiptir ve hem giriş hem de çıkış ile çalışmaya devam eder (Çanakçı, 2006). Destekleyici öğrenmeye örnek LVQ verilebilir.

1.1.4. Model değerlendirme

Kümeleme analizi çalışmalarında elde edilen sonuçların yeterli olup olmadığına karar vermek için yapılan değerlendirme çalışmaları genel olarak küme geçerliliği olarak adlandırılır (Akpınar, 2014). Küme geçerliliği, kümelenme kalitesini doğrulamanın bir yoludur. Algoritmalar ve doğal yapıyı keşfetme araçları veri setleri ölçümleri karşılaştırmak için kullanılır. Kümelenme algoritmalarının sonuçları ve ayrıca farklı kümelenmelere sahip iki kümelenme sonucu aynı olmayabilir. Bu nedenle, küme geçerliliği veri kümesindeki doğru küme sayısını belirlemek için kullanılabilir. Aslında, kümeleme prosedürü ve küme geçerliliği tavuk ve yumurta ilişkisine sahiptir. İyi bir kümelenme kriterinin nasıl tanımlanacağının bilinmesi verilerin anlaşılmasını gerektirir. Kümeleme, veriyi anlamada kullanılan temel araçlardan biridir (Caruana ve ark., 2006).

Farklı özelliklere sahip farklı kümeleme algoritmaları birbirinden farklı çözümler üretme eğilimindedir ve tüm olası veri kümeleri için tek bir "en iyi" kümeleme yöntemi yoktur. Kümeleme algoritmalarında önemli bir adım da, optimum çözümü belirlemek için kümelenme çözümlerini değerlendirmektir. Veri kümesi için küme yapısı, genellikle kümelerin sayısının değerlendirilmesine bağlıdır. Verilen veri kümesine en uygun kümeleme çözümünü bulmak için kümelenme sonuçlarının değerlendirilmesi veya küme doğrulaması yapılması gerekir. Kümeleme analizinin birçok yönü olduğu için, genellikle kümelenme görevini gerçekleştirmek için zaman alıcı bir iştir. Veri ön işleme, benzerlik ölçütleri, kümelerin sayısı, kümelenme parametreleri gibi dikkatli bir şekilde ele alınmalıdır. Algoritmalar, geçerlilik indisleri, kümeleme çözümlerinin değerlendirilmesi kümelenme süreçlerinde önemli faktörlerdir. Bu nedenle, bir kümelenme görevini daha iyi gerçekleştirmek için, bir ölçüt kullanılabilir. Kümelenme analizinde en önemli konulardan biri, bölümlemeyi bulmak için kümelenme sonuçlarının değerlendirilmesidir (Halkidi ve ark., 2001). Kümeleme çözümlerinin değerlendirilmesinde ve kümelenme sonuçlarının kalitesinin ölçülmesinde geçerlilik indeksleri kullanılır. İki çeşit geçerlilik göstergesi vardır: dış indeksler ve iç endeksler (Dudoit ve ark., 2002).

Kümeleme analizinde verilere uygun sayıda küme bulunması önemli bir sorundur. Özellikle k-ortalamalar gibi bazı kümeleme teknikleri, analizin başlangıcında küme sayısının belirlenmesini gerektirir. Diğer taraftan, kümelerin kalitesini sorgulamak çok önemlidir çünkü bir veri kümesi için farklı kümeleme algoritmaları

seçilerek farklı kümeleme algoritmaları oluşturulabilir. Bu bağlamda, aynı veri kümesinden elde edilebilecek farklı küme yapılarının anlamlı olup olmadığını test etmek için kümenin geçerliliği doğrulanır ve değerlendirme için çeşitli geçerlilik kriterleri geliştirilir. Küme kalitesi değerlendirilir. Böylece, eldeki gözlemler veya değişkenler için doğal kümelenme yapısı ve küme sayısı belirlenir ve alınan kararlar kümelenme geçerliliği kriterleri ile desteklenebilir (Bolshakova ve Azuaje, 2003).

İç ve dış kriterler olarak iki tür küme geçerlilik kriteri vardır. Dış ölçütlerde (Rand, Jaccard, Hubert, vb.) verilerle ilgili ön bilgiler kümelenme algoritmasının sonunda elde edilen kümelenme yapısı ile karşılaştırılır. Kümeleme sonunda elde edilen gözlemlerin küme etiketleri, daha önce bildirilmiş olan gözlemlerin kategori etiketleri ile karşılaştırılır. Diğer taraftan, iç ölçütler (Calinski-Harabasz, Hartigan, Shilhouette, vb.), kümelenme sonuçlarını, veri kümesi ile kümelenme yapısı arasındaki uyumu belirlerken, veri kümesindeki niceliksel değerleri ve yalnızca doğal yapıyı dikkate alarak değerlendirir. Çoğu iç ölçüt, kümeler içindeki kareler toplamı veya kümeler arasındaki kareler toplamı temelinde değerlendirilir (Theodoridis ve Koutroumbas, 2006).

Kümelenme çalışmalarının sonuç bölümünü kümelerin yorumlanması oluşturur. Bu nedenle kümeleme doğrulama yöntemleri, kümeleme algoritmasının sonuç değerlendirme sürecinde kullanılmaktadır. Küme doğrulama yöntemleri kullanılarak oluşturulan kümelerin kalitesi ölçülür.

1.1.4.1. İçsel değerlendirme

Elde edilen kümeleme sonucunun geçerliliğinin değerlendirilmesinde, bu kümelerin oluşturulmasında kullanılan veri dizisinden yararlanılması durumunda gerçekleştirilen değerlendirmeler içsel değerlendirme olarak adlandırılır. Bu yöntemlerde en iyi skoru küme içerisinde en yüksek benzerliğe ve kümeler arasındaki en düşük benzemezliği sağlayan algoritma içerir. İçsel değerlendirme yöntemleri aşağıdaki tabloda gösterilmiştir.

Çizelge 1.3. İç değerlendirme ölçütleri

Rootmeansquare	$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(E_{tahmin} - E_{gerçek})^2}$
R-squared	$RS = \frac{(\sum_{x\in D}\|x-c\|^2 - \sum_i \sum_{x\in c_i}\|x-c_i\|^2)}{\sum_{x\in D}\|x-c\|^2}$
ModifiedHubert Γ statistic	$\Gamma = \frac{2}{n(n-1)}\sum_{x\in D}\sum_{y\in D} d(x,y) d_{x\in c_i, y\in c_j(c_i,c_j)}$
Calinski-Harabasz index	$CH = \frac{\sum_i n_i d^2(c_i,c)/(NC-1)}{\sum_i \sum_{x\in c_i} d^2(x,c_i)/(n-NC)}$
I index	$I = (\frac{1}{NC}\frac{\sum_{x\in D} d(x,c)}{\sum_i \sum_{x\in C_i} d(x,c_i)} \max_{i,j} d(c_i,c_j))^p$
Dunn's index	$D = \min_i \left[\min_j (\frac{\min_{x\in c_i, y\in c_j} d(x,y)}{\max_k (\max_{x,y\in c_k} d(x,y))}) \right]$
Xie-Beni index	$XB = \frac{\sum_i \sum_{x\in c_i} d^2(x,c_i)}{\left[n \min_{i,j\neq i} d^2(c_i,c_j) \right]}$
Silhouette index	$\frac{1}{NC}\sum_i \frac{1}{n_i}\sum_{x\in c_i} \frac{b(x)-a(x)}{\max[b(x),a(x)]}$

D: data set; n: number of objects in D; c: center of D; P: attributesnumber of D; NC: number of clusters; Ci: the i–thcluster; ni: number of objects in Ci; ci: center of Ci; $\sigma(Ci)$: variancevector of Ci; $d(x, y)$: distancebetween x and y. (Yanchi ve ark., 2010).

Root Mean Square: Kümeler arasındaki farkın derecesini ölçmek için toplanmış örneklem varyansının karekökü alınarak hesaplanır (Sharma, 1996).

Hubert Γ istatistiği (Γ): herhangi iki matris arasındaki korelasyon katsayısını kullanarak kümeler arasındaki farkı değerlendirir (Hubertand ve Arabie, 1985).

Calinski-Harabasz indeksi (CH): Varyans oran kriteri olarak da adlandırılan bu indeks kümelenme geçerliliği ile kümeleşme arasındaki ortalamaya dayanarak geçerlilik kareler toplamını değerlendirir. İndeks küme arasındaki maksimum mesafeye bağlı olarak merkezler ve toplamı temel alan kompaktlığı ölçer (Harabasz ve ark., 1974).

Dunn'in endeksi (D): Yoğun ve iyi ayırılmış kümeleri hedeflemektedir. Küme içi uzaklıklarla, kümeler arası uzaklıkların oranını temel alan bir indekstir. Kümeler arası uzaklık iki kümenin merkezleri arasındaki uzaklık gibi herhangi bir uzaklık ölçüsü olabilir. En uygun küme numarası için bu endekslerin değerini maksimize etmek

gerekir. Yani birçok kümeleme algoritması denendiğinde, en büyük endeks değerine sahip olan algoritma bu uygulama için en iyi olarak seçilecektir (Dunn, 1974).

Silhouette endeksi (S): Bir gözlemin ne kadar iyi kümelendiğini kümeler arasındaki ortalama mesafeyi tahmin ederek ölçer. Ek olarak, optimum küme sayısı, bu endeksin değeri maksimize edilerek belirlenir (Rousseeuw, 1987).

Davies-Bouldin endeksi (DB): Küme içi uzaklıkların düşük, kümeler arası uzaklıkların yüksek olduğu kümeler bulan algoritmaların düşük bir DB indeksine sahip olması gerekir (Davies ve Bouldin, 1979).

Xie-Beni indeksi (XB) yoğunluk ve ayrılma geçerlilik fonksiyonu olarak da bilinen indeksin küçük olması önemlidir (Xie ve Beni, 1991).

1.1.4.2. Dışsal değerlendirme

Kümeleme sonuçları, bilinen sınıf etiketleri ve dışsal kıyas kriterleri gibi kümeleme için kullanılmayan veri temel alınarak gerçekleştirilir. Bu kıyas kriterleri önceden sınıflandırılmış ve genellikle uzmanlar tarafından oluşturulmuş eleman dizilerini içerir. Değerlendirme yöntemlerinin bu tipleri kümeleme sonuçlarının önceden belirlenmiş kıyas kriteri sınıflarına ne kadar yakın olduğunu ölçmektedir.

Çizelge 1.4. Dış Değerlendirme Ölçüleri

Entropy	$E = -\sum_i p_i \left(\frac{\sum_j p_{ij}}{p_i (\log(p_{ij} / p_i))}\right)$
Purity	$P = \sum_i p_i (\max_j p_{ij} / p_i)$
F Ölçüsü	$F = \sum_j p_j (\max_i \left[2 \frac{p_{ij} p_{ij}}{p_i p_j} / (\frac{p_{ij}}{p_i} + \frac{p_{ij}}{p_j}) \right]$
Variation of Information	$VI = -\sum_i p_i \log p_i - \sum_j p_j \log p_j - 2\sum_i \sum_j p_{ij} \log \frac{p_{ij}}{p_i p_j}$
Mutual Information	$MI = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{p_i p_j}$
Rand indeksi	$RI = \frac{\left[\binom{n}{2} - \sum_i \binom{n_{i.}}{2} - \sum_j \binom{n_{.j}}{2} + 2\sum_{ij} \binom{n_{ij}}{2}\right]}{\binom{n}{2}}$
Jaccard İndeksi	$J = \frac{\sum_{ij} \binom{n_{ij}}{2}}{\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2} - \sum_{ij} \binom{n_{ij}}{2}\right]}$

Entropy: Bir rastgele değişkenin entropisi, bir rastgele değişkenin öngörülemezliğini karakterize etmeye çalışan bir fonksiyondur. Entropi kümelenme sınıfının saflığını ölçer ve etiketler. Bir kümedeki nesne aynı sınıf etiketine sahipseentropy değeri sıfır olur. Bir kümedeki nesnelerin sınıf etiketleri daha fazla olduğunda entropi değeri artar.

Purity: Saflık entropiye çok benzer. Saflık, bir kümenin yalnızca bir bölümden varlıkları ne ölçüde içerdiğini belirler. Başka bir deyişle, nasıl olduğunu ölçer. Her küme "saf" dır. Bu nedenle, K kümelenmesinin saflığı küme-bilgili saflık değerlerinin ağırlıklı toplamı olarak tanımlanır.

Rand İndeksi: Kümelerin kıyas kriteri sınıflandırmalarına ne kadar benzer olduğu hesaplanmaktadır. Rand Endeksi, bir ölçüde algoritma tarafından bulunan doğru kararların yüzdesi olarak da düşünülebilir.

Jaccard İndeksi: İki veri dizisi arasındaki benzerliği sayısallaştırmak için kullanılır ve 0 ile 1 arasında değer alır. Değerin 1 olması iki veri dizisinin aynı olduğu, 0 olması da iki dizi arasında hiç ortak eleman olmadığı anlamına gelir (Akpınar, 2014).

Mutual information: Aynı anda örneklenen iki rastgele değişken arasındaki ilişkiyi ölçen bir niceliktir. Karşılıklı olarak ortalama rastgele bir değişkende ne kadar bilginin iletildiğini ölçer. İki değişken arasındaki karşılıklı bilginin sadece iki değişkenin istatistiksel olarak bağımsız olması durumunda 0 olduğu anlamına gelir.

Rand Endeksi:

Kümelerin kıyas kriteri sınıflandırmalarına ne kadar benzer olduğu hesaplanmaktadır. Rand endeksi, bir ölçüde algoritma tarafından bulunan doğru kararların yüzdesi olarak da düşünülebilir (Nizam ve ark., 2014).

Confusion Matrix:

Bir Confusion Matrisi, sınıflandırma modeli tarafından yapılan doğru ve yanlış tahminlerin sayısını, verilerdeki gerçek sonuçlara (hedef değer) kıyasla gösterir. Matris *N* x *N'dir*, burada *N* hedef değerlerin (sınıflar) sayısıdır. Bu tür modellerin performansı genellikle matristeki veriler kullanılarak değerlendirilir. Aşağıdaki çizelgede iki sınıf için 2x2 Confusion matrisi gösterilmektedir (Pozitif ve Negatif) (Flach, 2004; Ertosun ve ark., 2009).

Çizelge 1.5. Confusion Matrisi (Akpınar, 2014).

Confusion Matrix		HEDEF			
		Pozitif	Negatif		
Model	Pozitif	a	b	Pozitif tahmin	a/(a+b)
	Negatif	c	d	Negatif tahmin	d/(c+d)
		Duyarlılık $\frac{a}{a+b}$	Özgünlük $\frac{d}{b+d}$	Doğruluk $\frac{a+d}{a+b+c+d}$	

a: TP (true positive), b: FN (false negative), c: FP (false positive), d: TN (true negatif)

$$TPR(sensitivity) = \frac{TP}{TP + FN} \quad TNR(specificity) = \frac{TN}{TN + FP}$$

$$p = \frac{a}{a+c}, \qquad r = \frac{a}{a+d}$$

$$F = \frac{2rp}{r+p}$$

F daha küçük hassasiyete daha yakın olma eğilimindedir (Flach, 2004).

Kappa Test İstatistiği

Kappa testi, iki veya daha fazla gözlemci arasındaki uyumun güvenilirliğini ölçen istatistiksel bir yöntemdir (Gwet, 2010). Cohen'in kappa katsayısı yalnızca iki gözlemci arasındaki uyumluluğu ele almaktadır. Sonuç yalnızca "Kappa katsayısı Sonuç olarak ifade edilirse, bu değer" Cohen'in kappa katsayısı Sonuç olarak ifade edilir. Uygunluğun değerlendirildiği değişken kategorik bir değişken olduğundan, uygulanan istatistikler parametrik olmayan istatistiklerdir. Cohen'in kappa testine göre, gözlemciler arasındakiuyumun bir şans olabileceği ve iki gözlemci arasındaki uyumdan daha güçlü bir sonuç verdiği düşünülmektedir (Cohen, 1960). Kappa katsayısı hesaplanırken iki farklı olasılık hesaplanır. Bunlar Pr (a) ve Pr (e). Pr (a) iki değerlendirici için gözlemlenen toplam gözlem oranıdır ve Pr (e) şans olasılığıdır. 'Cohen'in kappa katsayısı' için kullanılacak formül aşağıdaki gibidir (Glas ve ark., 2003).

$$\kappa = \frac{pr(a) - pr(e)}{1 - pr(e)} \tag{1.11}$$

Kappa değeri (-1) ile (+1) arasında değer alabilir ve bulunan değer şu şekilde yorumlanır (Dawson ve Trap, 2004):

κ= +1 İki gözlemcinin sonuçları tümüyle birbiri ile uyumludur.

κ= 0 İki gözlemci arasındaki uyum sadece şansa bağlıdır.

κ= -1 İki gözlemci tümüyle birbirinin tersini değerlendirmektedir.

Çizelge 1.6. Kappa İstatistiğininin Yorumlanmasına İlişkin Değer Aralıkları (Goodwin, 2001).

κ	Gücü
< 0,00	Zayıf
0,00 – 0,20	Önemsiz
0,21 – 0,40	Düşük
0,41 – 0,60	Orta
0,61 – 0,80	Önemli
0,81 – 1,00	Çok Yüksek

2. KAYNAK BİLDİRİŞLERİ

Kayıp veri probleminin tarihsel geçmişi en az veri analizi kadar eskidir. Kayıp veri ile ilgili ilk çalışmalar 1900'lü yılların başına dayanmaktadır. Rubin'in (1976) kayıp veri mekanizmalarını açıkladığı çalışması bu konuda dönüm noktası olarak gösterilebilir.

Greenlees ve ark. (1982)'na göre; "birkaç istisna dışında parametre tahmini problemlerindeki kayıp değerlerin rastgele kayıp olduğu durumlar ele alınmaktadır ve önemli ilerlemeler kaydedilmesine rağmen, rastgele kayıp şartı geçerli olmaya devam etmektedir. Little ve Rubin, (1987); Navarro ve Losilla, (2000); Schafer, (1997) ve Simonoff, (1988) çalışmalarında kayıp verilerin analizi için farklı yöntemlerin gerekliliklerini değerlendirmişlerdir. Maksimum olabilirlik (EM) algoritması, Yapısal Eşitlik Modellemesi (SEM), Raw Maksimum Olabilirlik (RML) ve çoklu atama (MI) metotları, genelleştirilebilir sonuçlar için, kayıp değerlerin tamamen rastgele veya en azından rastgele kayıp olmasını gerektirmektedir.

Gold ve Bentler (2000) ve Graham ve ark., (1996) açıklayıcı bir doğrusal regresyon modelinin tahmininde kayıp verilerin tamamen rastgele veya rastgele olarak kayıp olduğu çalışmalarda, kayıp verilerin işlenmesi için en yüksek olasılık tahmini ve çoklu atama yöntemlerini uygulamışlardır.

Graham ve arkadaşları (1996), planlanan kayıp değer desenleri ile elde edilen kayıp olmayan değerlerin analiz edilmesinde Maksimum olabilirlik ve çoklu atama üstünlüğünü göstermiştir. Wothke (1998), bu tekniklerin, kayıp değerler rastgele olmasa bile, diğer geleneksel yaklaşımlara göre daha az taraflı sonuçlar verdikleri için kullanılmasını önermektedir.

Kromrey ve Hines (1994), doğrusal olmayan bir regresyon modelinde açıklayıcı iki değişkenden birisinde, kayıp olmayan verilerin etkilerini araştırmıştır.

Hippel (2004), normal olarak dağıtılan değerler rastgele olarak kayıp olduğunda kayıp değer analizindeki önyargıları araştırmıştır. Özellikle klinik çalışmalar söz konusu olduğunda, kayıp veri içeren araştırmalarda Son gözlemi ileri taşıma gibi atama yöntemleri kullanılır. Birçok yazarın eleştirisine rağmen bu yöntemlerin varsayımları sonuca etkisi göz önünde bulundurulmadan kabul edilir. Bu konuyla ilgili göreceli

olarak ilk çalışmalar Heyting, Tolboom ve Essers tarafından (1992) yapılmıştır. Mallinckrodt ve ark. (1995) direkt-benzerlik ve çoklu atama yöntemlerini önermişlerdir.

Siddiqui ve Ali (1998) direkt-benzerlik ve son gözlemin ileri taşınması yöntemlerini karşılaştırmıştır.

Little ve Rubin (2002)'e göre eğer veri kaybı, gözlenmiş ve gözlenmemiş çıktıların her ikisinden de bağımsız ise bu işlem Tamamen Rastgele Kayıp (TROK) olarak tanımlanır. Kayıp veri, gözlenmemiş değişkenden bağımsız ise Rastgele Kayıp (ROK) olarak adlandırılır. Aksi takdirde Rastgele Olmayan Kayıp (MNAR) olarak tanımlanabilir. MNAR durumunda yeterli bir veri analizi yapılamaz. Kayıp sürekli veriler için Piantadosi (1997), Clayton ve Hills (1993), Green ve ark. (1997), Friedman ve ark. (1998), çalışmalar yapmışlardır. Kayıp kategorik veriler için ise Stasny (1986), Baker ve Laird (1988), Baker ve ark. (1992) ve Conaway (1992,1993) tarafından, önemli çalışmalar yapılmıştır.

Enders (2010) kayıp veri mekanizmalarının ve kayıp veri yapılarının birbirinden farklı kavramlar olduğunun altını çizmiştir. Kayıp veri yapıları kayıp veriler ile gözlenmiş verilerin veri seti içerisinde nasıl yapıldığı ile ilgiliyken, kayıp veri mekanizmaları ölçülen değerler ile kayıp veri olasılığı arasındaki ilişkiyi açıklar. Kayıp veri yapıları kayıp verinin nedeni ile değil nerede bulunduğu ile ilgili sorulara cevap verir.

Juhola ve Laurikkala (2013) sınıflandırma sonuçlarını etkileyen kayıp veri miktarını ölçmek için çalışmalar yapmışlardır. %0, %10, %20 ve %30 kayıp veri bulunduran veri setlerini en yakın komşu yöntemi, diskriminant analizi ve Bayes yöntemle sınıflandırmışlardır. İkiden fazla değişken bulunan %20- %30'lara varan ölçüde kayıp veri bulunduran veri setleri üzerinde yapılan analizler verimli sonuçlar verirken, ikiden fazla değişken olan durumlarda %10- %20 kayıp veri olduğu durumların problem yaratabileceğini söylemişlerdir.

Kayıp verilerle ilgili yapılan çalışmalarda en çok %5, %10 ve %15 (Chen, Wang, Chen, 2011; Enders ve Bandalos 2001; Fiona ve ark., 2006; Sijtsma ve ark., 2003) kayıp veri oranlarının kullanıldığı belirlenmiştir. Bu çalışmalarda kayıp veriler farklı atama yöntemleri denenerek atmalar ve karşılaştırmalar yapılmıştır.

Chi ve Yang (2006); Wong ve ark., (2006); Brugger ve ark.,(2008); Chi ve Yang (2008); Gorgonio ve Costa, (2008) ; Yen ve Wu (2008); Cottrell ve ark., (2009); Arous

ve Ellouze, (2010); Tasdemir ve ark., (2011); Ayadi ve ark., (2012) kendini düzenleyen harita temelli veri sınıflandırmasındaki kayıp değerlerin tespit edilmesi ve görselleştirmeleri için model önermektedirler. İstatistiksel kümeleme metotlarının uygulanması zor olduğunda, kendini düzenleyen haritaların görselleştirilmiş kümeleme analizinde yararlı olduğu görülmektedir.

Orczyk ve Porwik, 2013'te yaptıkları Kayıp veri doldurma yönteminin tıbbi verilerin sınıflandırma doğruluğuna etkisi adlı çalışmada kayıp verileri - özellikle tıbbi verileri - doldurmanın tehlikelerini göstermektedirler. Veri atamasından kaynaklanan tehlikeler, seçilen kayıp veri doldurma algoritmalarının sınıflandırma doğruluğu üzerindeki etkisini gösteren makalede gösterilmektedir.

Zhu ve Shi, 2018'de yaptıkları bir çalışmadakayıp veriler için yeni bir destek vektör makinesi algoritması önerdikleri çalışmalarında kayıp verilerin nedenlerini göz önünde bulundurarak, sorunu nispeten büyük oranda kayıp veri ile çözmek için yeni bir destek vektör makinesi yöntemi önermişlerdir. Bu yöntem, çok sayıda kayıp değerin doldurulmasından kaynaklanan hatayı azaltmak için gözlenen verilerin tam olarak kullanılmasını önerir. Yöntem, UCI Makine Öğrenimi Deposu'ndan 4 veri seti üzerinde doğrulanmış ve performansı değerlendirmek için doğru sınıflandırma, F skoru, Kappa istatistikleri kullanılmıştır.

3. MATERYAL VE YÖNTEM

3.1. Tezde Kullanılan Veri Setleri

Bu çalışmada kümeleme yöntemlerinden k-ortalamalar, SOM ve LVQ yöntemleri kullanılarak gerçek veri setleri üzerinde çalışıldı. Çeşitli oranlarda veriler tamamen rastgele eksiltilerek tüm oranlarda doğru sınıflandırma oranları kullanılarak yöntemlerin kümeleme performansları değerlendirildi. Kullanılan veri setlerine Machine Learning Repository web sitesinden erişilmiştir. Bu analizler WEKA'da yapılmıştır. Kullanılan 11 veri seti küçük, orta ve büyük olarak sınıflandırılmıştır. Veri setleri ve özellikleri Çizelge 3.1'de verilmiştir.

Çizelge 3.1. Tezde kullanılan veri setleri

Veri Seti	Boyut	Küme Sayısı	İçerik	Kaynak	Veri Türü
Göğüs kanseri	10*699	2	Göğüs kanseri verileri Institute of Oncology University Medical Center Ljubljana, Yugoslavia'dan elde edilmiştir.M. Zwitter ve M. Soklic tarafından paylaşılmıştır Tarih: 11 Temmuz 1988	https://archive.ics.uci.edu/ml/datasets/breast	Nümerik ve nominal
Abalone	9*4177	3	Fiziksel ölçümlerden abalone yaşını tahmin etme amacıyla oluşturulmuştur. Marine Research Laboratories – Taroona tarafından paylaşılmıştır.	https://archive.ics.uci.edu/ml/datasets/Abalone	Nümerik
Diabet	8*768	2	Pima Hintliler Diyabet Veritabanı National Institute of Diabetes and Digestive and Kidney Diseases Vincent Sigillito tarafından 09.05.1990 tarihinde paylaşılmıştır.	https://archive.ics.uci.edu/ml/datasets/diabetes	Nümerik ve nominal
Ionosphere	35*351	2	Johns Hopkins Üniversitesi İyonosfer veritabanı Vince Sigillito 1989 yılında Uygulamalı Uzay Fiziği Grubu Uygulamalı PhysJohns Hopkins Üniversitesi tarafından paylaşıma açılmıştır.	https://archive.ics.uci.edu/ml/datasets/ıonosphere	Nümerik
Iris	5*150	3	İris Veritabanı R.A. Fisher ve Michael Marshall tarafından Temmuz 1988 de paylaşılmıştır.	https://archive.ics.uci.edu/ml/datasets/iris	Nümerik
Kredi kartı	21*1000	2	Alman Kredisi verileri Institut f"ur Statistik und "Okonometrie Universit" Hamburg tarafından Eylül 1987 de paylaşılmıştır.	https://archive.ics.uci.edu/ml/datasets/germancreditcard	Nümerik ve nominal
Kan Transfüzyon	5*748	2	Tayvan'daki Hsin-Chu Şehrindeki Kan Transfüzyon Servis Merkezinin donör veritabanın paylaştığı verilerdir.	https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center	Nümerik ve nominal

Çizelge 3.1. Tezde kullanılan veri setleri (devam)

Veri Seti	Boyut	Küme Sayısı	İçerik	Kaynak	Veri Türü
Segment	20*1500	7	Resim Bölümlendirme verileri Vision Group, Massachusetts Üniversitesi Vizyon Grubu (Carla Brodley, brodley@cs.umass.edu) Tarih: Kasım 1990 Örnekler rastgele 7 dış mekan görüntüsünün çizildiği bir veritabanıdır. Görüntüler her piksel için bir sınıflandırma oluşturmak üzere ele alınmıştır. Her örnek 3x3 bir bölgedir.	https://archive.ics.uci.edu/ml/datasets/segment	Nümerik
Sensör	13*2212	3	Sensör verileri, fiziksel ortamdan bir tür girdiyi algılayan ve yanıt veren bir cihazın çıktısıdır. Çıktı, başka bir sisteme bilgi veya girdi sağlamak veya bir işlemi yönlendirmek için kullanılabilir.	https://internetofthingsagenda.techtarget.com/definition/sensor-data	Nümerik
Şarap	15*178	3	Veriler, İtalya'da aynı bölgede yetişen ancak üç farklı çeşitten elde edilen şarapların kimyasal analizlerinin sonuçlarıdır. Analiz, üç şarap türünün her birinde bulunan 15 bileşendir.	https://archive.ics.uci.edu/ml/datasets/şarap	Nümerik
Gaz	128*13980	6	Özel olarak tasarlanmış sensörlerin çalışma sıcaklık kontrolü ve sinyalleri toplanmıştır.	https://archive.ics.uci.edu/ml/datasets/Twingassensorarrays	Nümerik

3.2. Tezde Kullanılan Yazılımlar

Kutools For Excel: Bir Excel eklentisi olan araç veri setlerinden rastgele veri çıkarmada kullanılmıştır.

SPSS 16.0 For Windows-Statistical Package Social Sciences: SPSS 16.0, her türlü veriyi düzenlemek ve analiz etmek için kullanılan lisanslı bir yazılımdır. Veriler temel olarak herhangi bir kaynaktan alınabilir. MS Excel veya OpenOffice'den e-tablolar; düz metin dosyaları (.txt veya. csv); ilişkisel (SQL) veritabanları; Stata ve SAS. Tezde kullanılan SPSS Van Yüzüncü Yıl Üniversitesi tarafından sağlanan pakettir.

Veri setlerinden veriler eksiltildikten sonraki adımda yeni veri setlerinin tamamen rastgele kayıp veri seti olup olmadığını test etmek için kullanılmıştır.

KNIME Yazılımı: Knime, data mining suit ismi verilir. Yani veri madenciliği yapılan bir yazılımdır. Bu yazılımlar genellikle bir akış mantığı ile çalışır. Tamamen ücretsizdir. Basitçe amaç bir veri kaynağından bir hedefe/bilgiye akış sağlamaktır. Amaç, müşteri analizi, kampanya analizi, tahminler, farklı kaynaklardan gelen veri birleştirilmesi vb. olabilir. www.knime.org sitesinden indirilebilir. Veri setlerini WEKA'da kullanabilmek için arff formatına çevirmek için kullanılmıştır (Şeker ve Erdoğan, 2017).

WEKA Yazılımı: Weka, Yeni Zelanda‟daki Waikato Üniversitesi tarafından geliştirilmiş olup “Waikato Environment for Knowledge Analysis” kelimelerinin baş harflernin kısaltmasıdır. Weka başta Yeni Zelanda’da tarımsal verinin işlenmesi amacıyla geliştirilmiştir. Bununla birlikte sahip olduğu makine öğrenme metotları ve veri mühendisliği kabiliyeti öyle hızlı ve köklü bir şekilde gelişmiştir ki, şimdi veri madenciliği uygulamalarının tüm formlarında yaygın olarak kullanılmaktadır (Frank ve ark., 2002).

Weka, bir öğrenen makinalar algoritmaları koleksiyonu olduğu gibi yeni algoritmaların geliştirilmesi için de çok uygundur. GNU (General Public License) altında yayınlanmış, Java dilinde kodlanmış, açık kaynaklı bir yazılımdır (Frank ve ark., 2002). Ayrıca WEKA, Windows, Linux ve Mac-os gibi farklı işletim sistemleri üzerinde çalışabilen bir programdır (Witten ve ark., 2005). Weka Grafiksel Kullanıcı Arayüzü, WEKA‟nın grafiksel çevresine erişim için kullanılmaktadır. www.weka.orgsitesinden indirilebilir.

Weka pencerinin alt kısmında ise dört adet seçenek bulunmaktadır:

1. Simple CLI: WEKA komutlarının direkt olarak işlenmesine olanak sağlayan basit bir komut satırı arayüzü sağlar.

2. Explorer: Verinin WEKA ile keşfi için bir arayüzdür. Bu arayüzde sınıflandırma, kümeleme ve birliktelik kuralı uygulamaları kolaylıkla gerçekleştirilmektedir. Weka Explorer ile, Bayes sınıflayıcısı, karar ağaçları, karar kuralları, regresyon, yapay sinir ağları gibi sınıflandırma algoritmaları; K-ortalama, Cobweb gibi kümeleme algoritmaları; Apriori gibi birliktelik kuralları kolaylıkla uygulanabilmektedir. Weka Explorer‟ da önişleme, sınıflama, kümeleme, birliktelik kuralları, özellik seçme ve görselleştirme panelleri bulunmaktadır.

Önişleme: Veri dosyalarının yüklendiği, veri tabanının seçildiği ve verinin çeşitli yollarla değiştirildiği keşif sürecinin ilk adımıdır(Han ve Kamber, 2006).

Sınıflama: Sınıflandırma ve regresyon algoritmalarının uygulanıp değerlendirildiği paneldir. Sınıflandırma fonksiyonları, kuralları, karar ağaçları, Bayes ağları, sinir ağları gibi sınıflandırma algoritmaları bu panelde yer almaktadır (Somasundaram ve Nedunchezhian, 2011).

Kümeleme: K-ortalama, cobweb gibi kümeleme algoritmalarının yer aldığı paneldir. Birliktelik kuralları: Verilerden birliktelik kurallarının çıkarıldığı paneldir.

Özellik şeçme: Veri kümesindeki ilişkili verilerin seçildiği paneldir.

Görselleştirme: Özellikler arasındaki ilişkiler iki boyutlu grafiklerle izlenebildiği paneldir.

3. Experimenter: Deneylerin gerçekleştirilmesi ve öğrenme planları arasındaki istatistiksel testleri yürüten bir arayüzdür. Bir veri setine farklı teknikleri uygulayarak ya da aynı tekniği farklı parametrelerle tekrarlayarak, tek seferde birden fazla deneyin gerçekleştirilmesine izin veren bir araçtır.

4. Knowledge Flow: Weka veri madenciliği paketi ile sağlanan fonksiyonerliğin alternatif bir arayüzüdür. Bu arayüz temel olarak Explorer ile aynı işlevleri sürükle-bırak arayüzü ile yerine getirmektedir. Experimenter tarafından desteklenmeyen ek özellikleri ve experimenter de bulunan bazı eksik özellikleri ile gelişmekte olan bir bölümdür.

WEKA, algoritmalarının sonuçlarını değerlendirirken şu çıktıları sunmaktadır:

Düzensizlik matrisi: Yakınsaklık matrisi olarak da adlandırılır. Doğru olarak sınıflandırılan örneklerin sayısı bu matrisin köşegeni üzerindeki elemanlarının toplamına eşittir. Doğru olarak sınıflandırılan kayıt yüzdesi bize madencilik algoritmalarını karşılaştırma imkanı sunmaktadır.

True Positive (TP): Sınıflandırma algoritması tarafından herhangi bir sınıfa atanan kayıtlardan gerçekte o sınıfa ait olanların oranını yüzdesel olarak gösterir.

FalsePositive (FP): Sınıflandırma algoritması tarafından herhangi bir sınıfa atandığı halde gerçekte o sınıfa ait olmayan kayıtların oranını gösterir.

Kesinlik: Gerçekte herhangi bir sınıfa ait olan kayıtların hangi oranda sınıflandırma algoritması tarafından o sınıfa atandığı gösterir.

Kappa istatistiği: Tahmin doğruluğunun ölçüsüdür.

3.3. Yöntemler

3.3.1. k-ortalamalar

k-ortalamalar yöntemi verileri tanımlamak için kümelerin merkezleri olan k prototiplerini kullanır. Hata kareler toplamını minimize ederek belirlenirler. k-ortalamalar (Macqueen, 1967), klasik ancak bölümlü kümeleme için en çok kullanılan yöntemlerden biridir. Veri setini k grupta gruplar. Gruplama, veri noktaları ve ilgili

küme merkezleri arasındaki mesafelerin karelerinin toplamını en aza indirgeyerek yapılır. Metodun mantığı, bu adımların yinelemelerine bağlıdır; ilk olarak merkezin koordinatının belirlenmesi, her cismin merkezlere olan uzaklığının değerlendirilmesi ve asgari mesafeye dayalı nesnelerin son gruplandırılmasına dayanır (Macqueen, 1967).

Küme içi benzerliği maksimum, kümeler arası benzerliklerin minimum olmasını sağlamak k-ortalamalar kümeleme yönteminin temel amacıdır (Han ve ark., 2012). Bu yöntemin değerlendirilmesinde en yaygın olarak hata kareler toplamı (Sum of Squared Error-SSE) kullanılır. En küçük SSE değerine sahip kümeleme, bu kümelemede merkezlerin (ortalamaların) kümeleri en iyi temsil eden noktalar olduğu anlamına gelir. Hata kareler toplamı (SSE) ise;

$$SSE = \sum_{j=1}^{k} \sum_{x \in c_j} \left\| x_i - c_j \right\|^2 \tag{1.5}$$

şeklinde ifade edilir. Burada x_i: veri noktalarını, c_j: merkezini k: küme sayısını göstermektedir.

k-ortalamalar süreci n adet birimin rasgele k tanesi seçilip ve her birinin k tane kümeye atanmasıyla başlar. Bu verilerin her biri, bir kümenin merkezini oluşturur. Bir sonraki aşamada diğer verilerin hepsi teker teker, kendine en yakın küme merkezine göre yeniden kümelenir. Ortalama, yeni oluşan her küme için tekrar hesaplanır. Bulunan yenideğer o kümenin yeni merkezidir. Merkezlerin yerleri değiştiği için yeni merkezleri esas alarak birimler yeniden en yakın küme merkezlerine atanırlar. Bu işlemler hiçbir kümenin merkezi değişmeyene kadar devam eder. Böylece her birim kendi kümesini bulmuş olur (Han ve ark., 2012).

Sadece kümenin ortalamasının tanımlanabildiği durumlarda k-ortalamalar kümeleme yöntemi kullanılır. Kullanıcının küme sayısını önceden bilmesi gerekliliği bir dezavantaj olarak görülebilir. Dahası, veri kümeleri boyutları, yoğunluğu, küresel olmayan şekiller ve aşırı noktalı farklı olduğunda k-ortalamalar yöntemi istenilen iyi sonucu üretemez. Asıl önemli olan dezavantaj ise aykırı değerlere ve gürültülere olan duyarlılıktır (Han ve ark., 2012).

3.3.2. Kümelemede kullanılan yapay sinir ağları

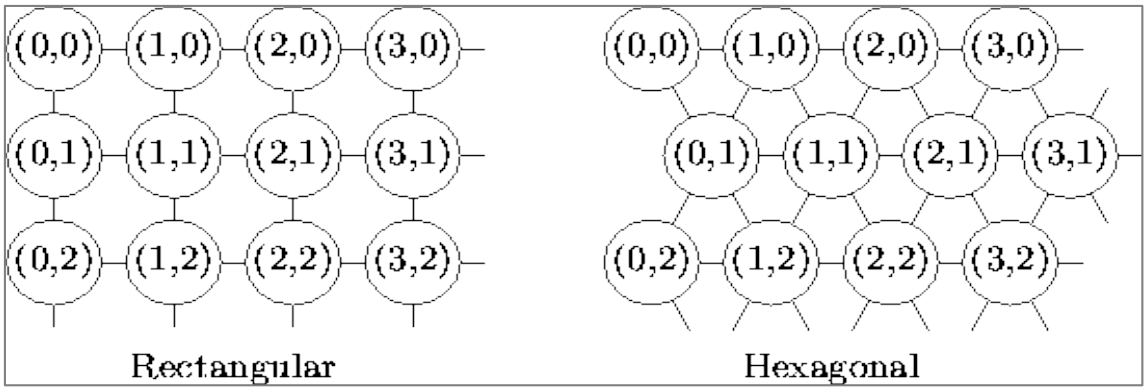
1) Öz düzenlemeli haritalar- Self Organization Map (SOM)

Öz düzenlemeli harita, en popüler sinir ağı modellerinden biridir. Rekabetçi öğrenme ağları kategorisinde sıralanır. Öz düzenlemeli harita (SOM) denetimsiz öğrenme modelidir. Bu da girdi verilerinin özellikleri hakkında çok az bilginin yeterli olduğu anlamına gelir. Girdi verilerinin sınıf üyeliğini bilmeden veri kümelemesi için SOM kullanabiliriz. SOM, problemin doğasındaki özellikleri tespit etmek için kullanılabilir ve literatürde Self Organization Future Map (SOFM), kendini organize eden özellik haritası olarak da adlandırılmaktadır. Öz düzenlemeli harita kavramı profesör Teuvo Kohonen tarafından geliştirilmiştir (Kohonen, 1993). SOM'un birçok uygulamada yararlı olduğu kanıtlanmıştır. SOM, denetimsiz, rekabetçi öğrenme algoritması kullanır. Yüksek boyutlu uzaydan harita birimlerine kadar bir topoloji koruma eşlemesi sağlar. Harita birimleri ya da nöronlar genellikle iki boyutlu bir kafes oluştururlar. Haritalama yüksek boyutlu alandan bir düzleme doğru bir haritalamadır. Topoloji koruma özelliği, haritalamanın noktalar arasındaki nispi mesafeyi koruduğu anlamına gelir. Giriş alanındaki birbirine yakın noktalar SOM'daki yakın harita birimlerine eşlenir. SOM, bu nedenle yüksek boyutlu verilerde küme analiz aracı olarak kullanılabilir. Ayrıca, SOM'un genelleme kabiliyeti vardır. Genelleştirme özelliği, ağın daha önce hiç karşılaşmadığı girdileri tanıyabileceği veya karakterize edebileceği anlamına gelir. Yeni bir giriş, eşlendiği harita birimi ile temsil edilir (Kohonen, 2001).

SOM haritası iki boyutlu nöron dizisidir:

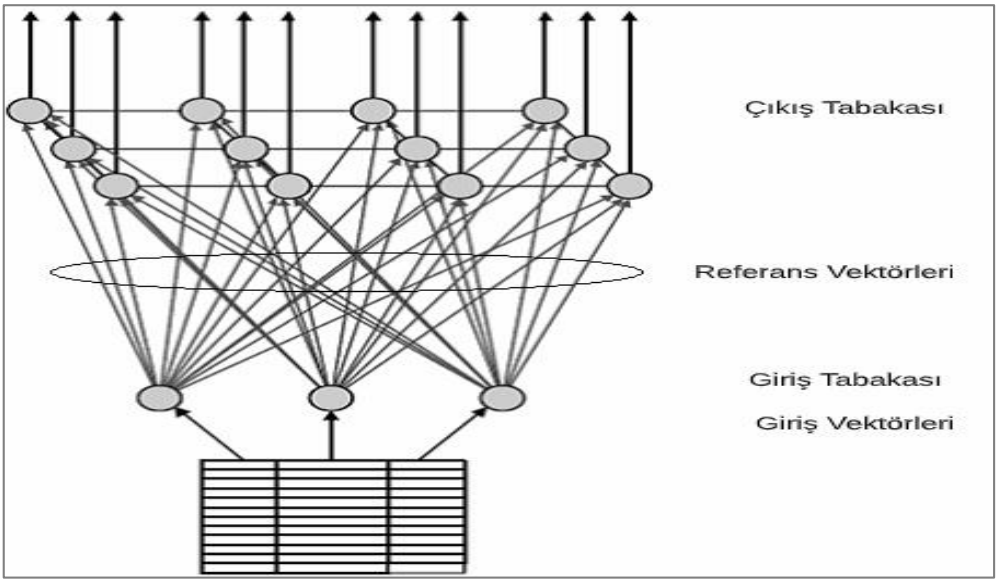
$$M = \{m_1, ..., m_{p*q}\} \tag{1.6}$$

Bu giriş vektörleri ile aynı boyuta sahiptir (n -boyutlu). Nöronlar komşu nöronlara bir komşuluk ilişkisi ile bağlanır. Bu haritanın topolojisini veya yapısını belirler. Genellikle, nöronlar dikdörtgen veya altıgen topoloji ile birbirine bağlanır. Şekil 3.1'de topolojik ilişkiler nöronlar arasındaki çizgilerle gösterilmiştir (Ultsch ve Siemon,1990).



Şekil 3.1. Farklı topolojiler ((Ultsch ve Siemon,1990).

SOM ağları giriş ve çıkış nöronlarından oluşan tek tabakalı bir yapay sinir ağıdır. Giriş nöronlarının sayısı, veri kümesindeki değişkenlerin sayısı tarafından belirlenir (Vatansever, 2008). Çıkış nöronları her biri bir kümeyi gösterir. Şekil 3.2 bir SOM ağını göstermektedir. SOM ağlarında, çıkış tabakasındaki nöronların sırası çok önemlidir. Bu dizi, doğrusal, dikdörtgen, altıgen veya kübik olabilir. Çoğunlukla dikdörtgen ve altıgen diziler kullanılır. Uygulamada, dikdörtgen hizalama genellikle karesel hizalama olarak uygulanır. Buradaki dizi topolojik komşuluklar için önemlidir. Çıkış nöronları arasında doğrudan bir bağlantı yoktur. Giriş nöronları ve her bir çıkış nöronu arasındaki bağlantı referans vektörlerini gösterir. Bu vektörler ayrıca bir katsayılı matrisin sütunları olarak da kabul edilebilir. SOM sinir ağları eğitilirken bu topolojik komşuluk referans vektörleri yenilenerek kullanılır (Zontul ve ark., 2004).



Şekil 3.2. Kohonen SOM sinir ağı (Vatansever, 2008).

SOM Öğrenme Algoritması

Bu ağlarda öğrenme algoritması olarak denetimsiz öğrenme kullanılır. Ağın eğitiminde bağımlı değişken kullanılmaz. Veri setindeki giriş vektörleri ağa girildikçe ağ kendi kendini düzenler ve referans vektörleri oluşur. Bu algoritma aşağıdaki gibidir (Zontul ve ark., 2004).

Bu algoritmada kullanılan semboller:

$x_n = x_{n1}, x_{n2}, ..., x_{nm}$: m özellik ve n kayıttan oluşan $m*n$' lik x veri matrisi için giriş vektörleri

$w_j = w_{1j}, w_{2j}, ..., w_{mj}$: m tane ağırlıktan oluşan j çıkış nöronları için referans vektörleri

$d(i, j)$: giriş vektörünün (i, j) koordinatındaki çıkış nöronuna olan Öklid uzaklığının karesi.

J : giriş vektörünün en yakın olduğu çıkış nöronları.

α : öğrenme katsayısı.

h : komşuluk fonksiyonu

c : kazanan nöron

Algoritma (Zontulve ark., 2004; Vatansever, 2008).

0.Adım

w_{ij} katsayılarına ilk değerleri ata.

Topolojik komşuluk (R) parametrelerini belirle

Öğrenme katsayısı (α) parametrelerini ayarla

1. Adım

Giriş vektörü x_n ve ağırlık vektörü w_j için

$$d(w_j, x_n) = \sqrt{\sum_i (w_{ij} - x_{ni})^2} \tag{1.7}$$

şeklinde Öklid uzaklıklarını hesapla.

2. Adım

Bütün nöronlar için $d(w_j, x_n)$' nin minimum olduğu j kazanan nöronu bul.

3. Adım

Komşuluk parametresi R için j kazanan nöronun J komşuluk nöronlarını bul.

4. Adım

t. iterasyonda j'nin belirtilen komşuluğundaki bütün çıkış nöronları (J) için aşağıdaki gibi referans vektörlerini güncelle

$$w_{ij}(t+1) = w_{ij}(t) + \alpha(t)h_{ci}(t)(x_{ni} - w_{ij}(t)) \tag{1.8}$$

5. Adım

Öğrenme katsayısını güncelle.

6. Adım

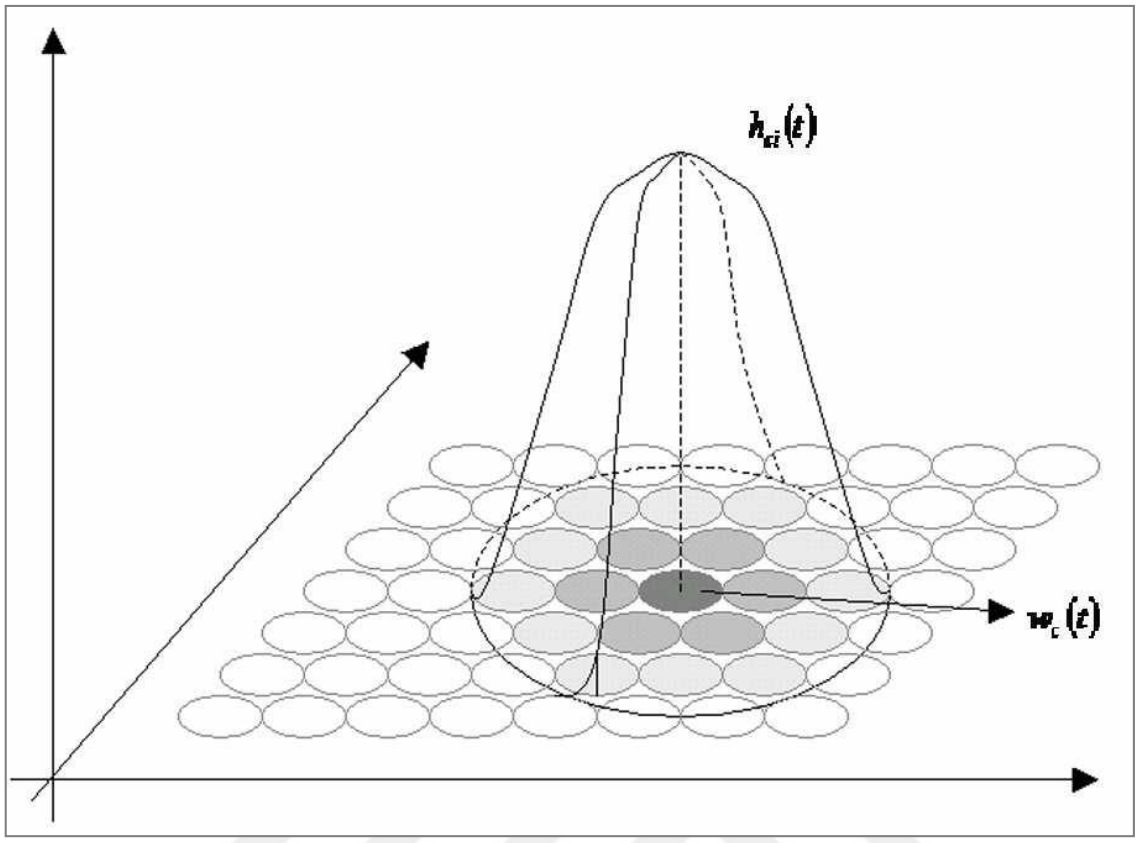
Belirtilen zamanlarda topolojik komşuluk parametresini azalt (Ultsch ve Siemon, 1990; Larose, 2005; Van Hulle, 2012).

Yukarıdaki algoritmayı açıklarsak, ilk olarak referans vektörlerine bir ilk değer atanır. Bu atama işlemi genellikle rastlantısal yapılır. Döngüye başlamadan önce öğrenme katsayısı (α) ve komşuluk değişkenine (R) yüksek bir değer atanır. α 'ya 0 ile 1 arasında bir değer verilir. Bu değerin 1'e yakın olması tercih edilir. Algoritma için bir döngü veri setindeki tüm satırların birer kere SOM ağına girdi olarak sunulmasıdır. Veri setinin bir satırı x_n vektörüdür. x_n vektörünün çıkış tabakasındaki her bir nörona olan öklid uzaklığının karesi bulunur. Çıkış tabakasındaki her bir nöronu bir referans vektörü w_j temsil eder. Dolayısıyla, bu uzaklık x_n vektörü ile w_j arasındaki uzaklıktır. Hesaplanan uzaklıklardan en küçüğü bulunur. Bu uzaklık hangi çıkış nöronuna ait ise o nöron kazanan (Winner neuron) nörondur. Literatürde kazanan nöron için BMU (Best Matching Units) kısaltması genellikle tercih edilir. Kazanan nöron aşağıdaki gibi ifade edilebilir (Kohonen, 1993).

$$c : w_c(t) = \min_t \|x(t) - w_i(t)\| \tag{1.9}$$

Komşuluk fonksiyonu, kazanan birimin konumu etrafında simetrik bir şekle sahip ve kazanandan uzaklaştıkça eşit şekilde azalan tek tepeli bir fonksiyondur. Gauss fonksiyonu komşuluk fonksiyonunu modellemek için kullanılabilir. Şekil 3.3'te SOM ağında kazanan birim üzerine konumlandırılmış Gauss fonksiyonun grafiği bulunmaktadır(Van Hulle, 2012).

$$h_{ci}(t) = \exp\left(-\frac{\|r_c - r_i\|^2}{2\alpha^2(t)}\right) \tag{1.10}$$

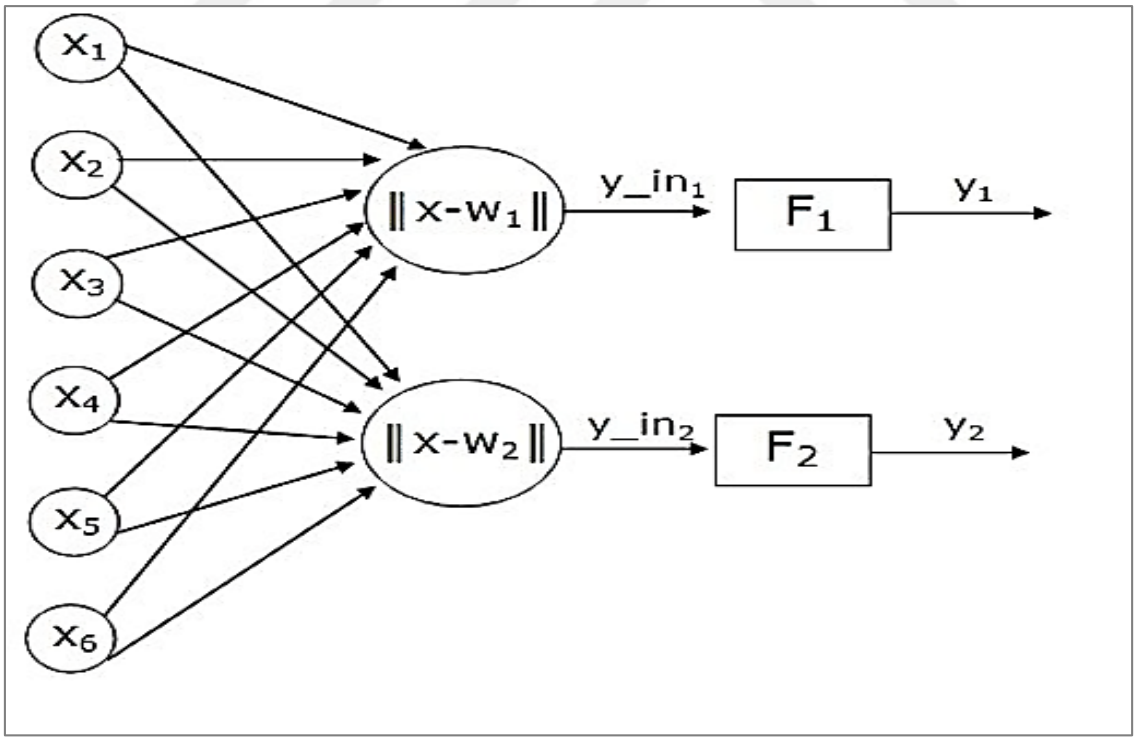


Şekil 3.3. Gauss fonksiyonu grafiği (Alpdoğan, 2007).

Denklem (1.10)'da r, i. biriminin ızgaradaki yerini gösteren iki boyutlu bir vektördür. Eşitlikteki $\|r_c - r_i\|$ ise aktif eğitim iterasyonundaki kazanan nöron c ile çıkış uzayındaki i birimi arasındaki uzaklığı göstermektedir. Etkileşimin uzaysal genişliği, zamanla değişen parametresi ile belirlenir. $h_{ci}(t)$ komşuluk fonksiyonu Gauss olarak seçilerek, kazanan nöronun ve çevresindeki komşularının ağırlığının güncellenmesi sağlanır. Ağırlık vektörlerinin hareketi ile giriş deseni ve ağırlık vektörü arasındaki Öklid mesafesi sürekli azalır ve sonuç olarak ağırlık vektörleri giriş modeline çok benzer hale gelir (Vatansever, 2008). Böylece, bir sonraki birim yineleme kazanma olasılığı artar. Benzer kalıpların mekansal kümelenmesi, sadece kazanan birim değil, aynı zamanda bu birime bitişik diğer birimler de kazanan ile birleştirilerek sağlanır. Bu nedenle, n boyutlu bir giriş alanındaki benzer giriş kalıpları, haritaları kendi kendine organize ederek iki boyutlu çıkış alanına dönüştürür. Çıktı alanındaki benzer kalıpların kümelenmesi, kendi kendini organize eden haritaların eğitim süreci ile sağlanır (Kohonen, 2001).

2) Doğrusal Vektör Parçalama Modeli (Learning Vectör Quantization-LVQ)

LVQ ağı 1984 yılında Tuevo Kohonen tarafından geliştirilmiştir. LVQ ağlarının temeli, Kohonen tarafından geliştirilen SOM modelindeki Kohonen tabakasıdır. LVQ modelinin sınıflandırma problemi için geliştirilen temel felsefesi, n boyutlu bir vektörü bir dizi vektörle eşlemektir. Hızlı sonuçlar vermesi ve yüksek performansları nedeniyle sık tercih edilen ağlardan biridir. LVQ ağının öğrenmesi, giriş vektörünün temsil ettiği vektör dizisi anlamına gelir. Bu vektör setine referans vektörleri denir. LVQ ağının görevi, bu referans vektörlerini öğrenerek tanımlamak yani, üye olabilecekleri giriş vektörlerinin vektör sınıflarını belirlemektir (Öztemel, 2003). LVQ ağları üç tabakadan oluşur. İlk tabaka olan giriş tabakası bilgi işlemez. Gelen bilgi giriş tabakasını oluşturur. İkinci tabaka, Kohonen tabakası olarak da adlandırılan ara tabakadır. Bu tabakada, giriş kümesine en yakın ağırlık vektörü SOM'daki gibi belirlenir. Bu tabakadaki her eleman bir referans vektörünü temsil eder. Giriş vektörü, giriş tabakası ve Kohonen tabakası arasındaki ağırlıkların oluşturduğu referans vektörleriyle eşleştirilir. Üçüncü tabaka olan çıktı tabakasında, girdinin ait olduğu sınıf belirlenir (Baş, 2006).



Şekil 3.4. LVQ şeması (Baş, 2006).

LVQ Ağının Yapısı

LVQ ağlarında giriş tabakasındaki her nöron, Kohonen tabakasındaki tüm nöronlarla ilişkilidir. Kohonen tabakasındaki nöronlar, çıktı tabakasındaki tek bir nöron ile ilişkilidir. Kohonen tabakası ve çıkış tabakası arasındaki ağırlıklar 1'e eşit, sabittir ve bu ağırlıklar değişmez. Ağın eğitimi sadece Kohonen tabakası ve giriş tabakası arasındaki ağırlıkları, yani referans vektörlerinin değerlerini değiştirerek gerçekleştirilir. Bu değişiklikler sayesinde, referans vektörleri, girişleri doğru sınıflara ayırmak içinbelirlenir. Ağı eğitirken, her yinelemede ağ tarafından üretilen çıkışın değer yerine sadece doğru olup olmadığı belirlenir. Sadece vektöre (kazanan vektör) giriş vektörüne yakın değerler (bu vektör için ağın ağırlıkları) değiştirilir (Kröse ve ark., 1996). Hem Kohonen tabakasındaki hem de çıkış tabakasındaki her nöron öğesinin çıktıları ikili değerleri alır ve yalnızca bir nöron öğesinin 1 çıktı değeri vardır ve diğerleri 0'dır. 1'in çıktı değeri, girdilerin girilen sınıfa ait olduğunu belirtir. LVQ ağlarının eğitiminde nicelleştirme algoritması kullanılmaktadır. Bir ağı eğitmenin amacı, her yinelemede giriş vektörüne en yakın referans vektörünü bulmak, yani küme merkezini ayarlamak ve egzersiz setindeki tüm giriş vektörlerinin nicelik hatalarını en aza indirmektir. Referans vektörleri, daha önce belirtildiği gibi, Kohonen tabakasındaki nöronları giriş tabakasındaki nöronlara bağlayan ağırlık değerleridir. Öğrenme sırasında sadece referans vektörlerin ağırlık değerleri değiştirilir. Bu Kohonen öğrenme kuralı kullanılarak yapılır. Kohonen öğrenme kuralı, Kohonen tabakasındaki nöron öğelerinin birbiriyle rekabet ettiği ilkesine dayanır. Rekabet kriteri, girdi vektörü ile ağırlık vektörleri (referans vektörleri) arasındaki Öklid mesafesi ile hesaplanır. Giriş vektörüne en yakın nöron rekabeti kazanır. Kazanan nöron için iki durum vardır (Cozart, 1996). İlk durumda, kazanan nöron doğru sınıfın bir üyesidir. Bu durumda, ilgili ağırlıklar giriş vektörüne yaklaştırılır. Bu, aynı örnek ağa tekrar gösterildiğinde aynı nöronnın tekrar kazanması için yapılır. Bu durumda, ağırlıklar değiştirilir. Öğrenme katsayısı monoton bir şekilde zaman içindeki 0 değerine düşürülür. Bunun nedeni, giriş vektörünün referans vektörüne çok yakın olduğunda durması ve bir daha geri çekilmemesidir. Bunu yapmamak daha fazla tersine çevrilmesine neden olacaktır. İkinci durumda, kazanan nöron yanlış sınıftır. Bu durumda, ağırlık vektörü giriş vektöründen çıkarılmalıdır. Bunun amacı, aynı nöron öğesinin, aynı örneğin bir dahaki sefere gelmesinde kazanmamasıdır. Ağırlıklar daha sonra değiştirilir. Öğrenme katsayısının zaman içinde

azaltılması da burada geçerlidir. Kohonen tabakası ile çıkış tabakası arasındaki ağırlıklar, eğitim sırasında değişmez. Kohonen tabakasındaki nöron elemanlarının çıkışları, ağın çıkışını hesaplamak için bu nöron elemanlarını çıkış tabakasına bağlayan ağırlık değerleri ile çarpılır. Ağın çıktıları belirlendiğinde, çıktının doğru bir şekilde sınıflandırılıp sınıflandırılmadığı sorgulanır. Cevap, nöronnı Kohonen tabakasındaki giriş tabakasına bağlayan ağırlıkları değiştirmektir. Bu nedenle, LVQ ağları güçlendirilmiş öğrenme sınıfındadır. Bu işlemlere eğitim setindeki tüm örnekler doğru bir şekilde sınıflandırılıncaya kadar devam edilir. Hepsi doğru bir şekilde sınıflandırıldığında, öğrenme gerçekleşir (Elmas, 2010).

4. BULGULAR

4.1. Kümeleme Yöntemlerinin Performanslarının Gerçek Veri Setleri Üzerinde İncelenmesi

Kümeleme yöntemlerinin performansları doğru sınıflandırma oranları kıllanılarak karşılaştırılabilir.

4.1.1. Kayıp veri içeren veri setlerinin oluşturulması

İlk olarak tam veri setlerinde analizleri yapılmış daha sonra Excel-Kutools araç kutusu yardımıyla tamamen rastgele kayıp (TROK) mekanizmasını sağlayan %5, %10, %15, %20, %25 ve %30 oranlarında kayıp veri setleri oluşturulmuştur.

Oluşturulan veri setlerinin tamamen rastgele olup olmadığının test edilmiştir. Rastgeleliğin sağlandığını görmek amacıyla verilere Little'ın TROK testi uygulanmıştır. Bunun için SPSS 16.0 programında bulunan EM algoritmasına bağlı olarak hesaplanan Little'ın TROK testi kullanılmıştır. Sonuçlar Çizelge 4.1'de özetlenmiştir.

Çizelge 4.1. Tüm veri setlerine ait tüm oranlarda p değerleri

Veri Seti	İris	Göğüs K.	Diabet	Şarap	Kan Tr.	Abalone	Ion.	Kredi K.	Sensör	Segment	Gaz
Kayıp Veri Oranı	p değeri	p değeri	p değeri	p değeri	p değeri	p değeri	p değeri	p değeri	p değeri	p değeri	p değeri
5%	0,99	0,99	0,74	0,13	0,67	0,74	0,65	0,95	0,58	0,96	0,75
10%	0,95	0,98	0,74	0,32	0,64	0,73	0,64	0,99	0,57	0,97	0,74
15%	0,94	0,98	0,75	0,38	0,45	0,7	0,62	0,97	0,54	0,94	0,72
20%	0,94	0,92	0,74	0,41	0,47	0,69	0,65	0,96	0,54	0,92	0,73
25%	0,93	0,84	0,73	0,54	0,43	0,66	0,63	0,95	0,56	0,92	0,74
30%	0,94	0,84	0,74	0,23	0,41	0,65	0,64	0,95	0,54	0,94	0,74

Çizelge 4.1 incelendiğinde tüm oranların p anlamlılık değerinin 0,05'ten büyük olduğu görülmektedir. Bu durum verilerin TROK mekanizmasına uyduğunun göstergesidir. Verilerin ilgili analizler için uygun olduğu görülmektedir

4.1.2. Kümeleme yöntemlerinin karşılaştırılası

Çalışmada kullanılan veri setleri kendi içinde küçük veri setleri, orta veri setleri ve büyük veri seti olarak sınıflandırılmışlardır.

4.1.2.1. Küçük veri setleri

Küçük veri setleri iris, göğüs kanseri, kan transfüzyon, diabet ve şarap very setleridir.

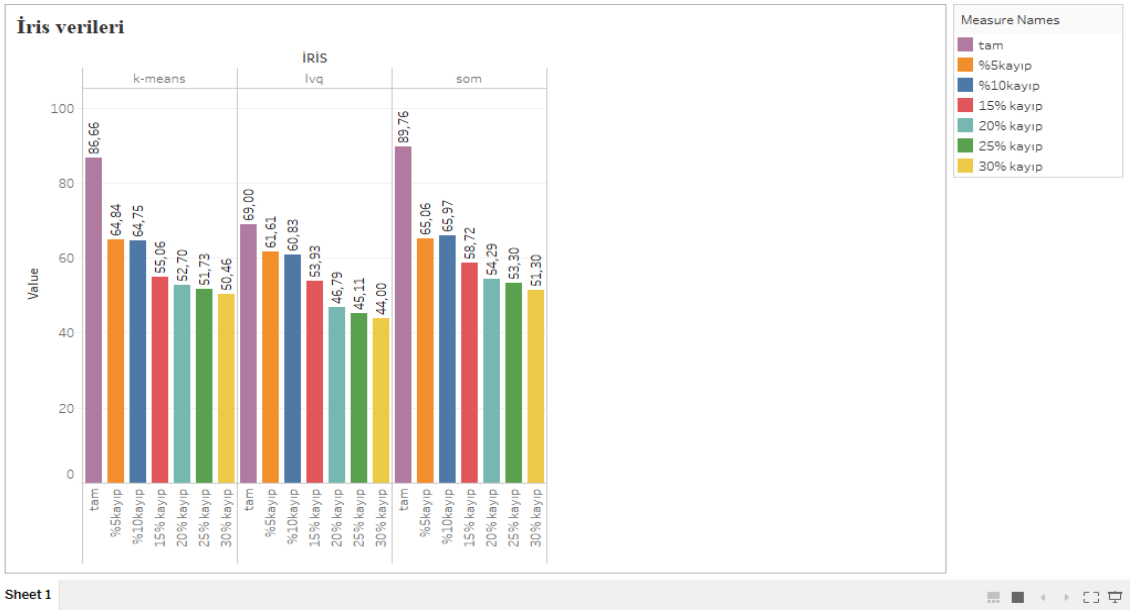
a) İris Verilerine Ait Değerlendirme

İris verileri, iris çiçeğine ait ölçümlerden oluşan, üç kümeli, nümerik veriler içeren bir veri setidir. Çizelge 4.2'de İris verilerine ait doğru sınıflandırma oranları görülmektedir. Kayıp veri oranı arttıkça doğru sınıflandırma oranlarının düştüğü görülmektedir.

Çizelge 4.2. İris verilerine ait doğru sınıflandırma oranları

İRİS	tam	%5 kayıp	%10 kayıp	15% kayıp	20% kayıp	25% kayıp	30% kayıp
k-ortalama	86,66	64,84	64,75	55,06	52,7	51,73	50,46
som	89,76	65,06	65,97	58,72	54,29	53,3	51,3
lvq	69	61,61	60,83	53,93	46,79	45,11	44

Şekil 4.1'de İris veri setine ait doğru sınıflandırma oranları oranları görülmektedir.



Şekil 4.1. İris verilerine ait doğru sınıflandırma oranları.

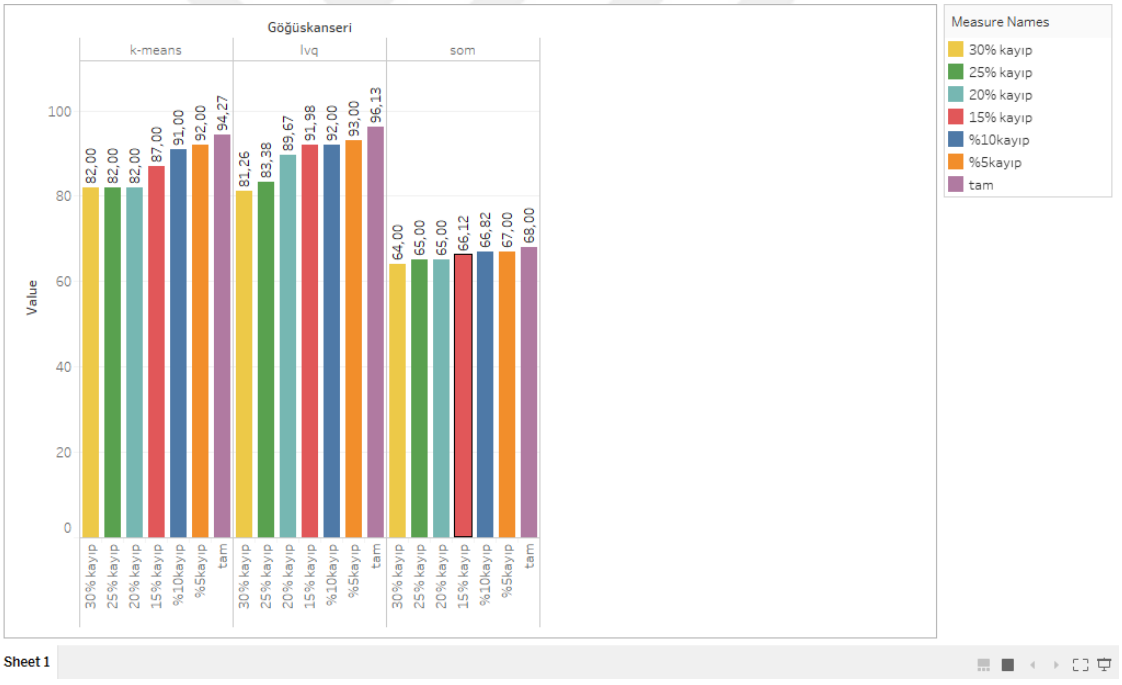
b) Göğüs Kanseri Veri seti için Değerlendirme

Göğüs kanseri verileri, göğüs kanseri hastalarına ait iki kümeli, nümerik ve nominal veriler içeren bir veri setidir. Çizelge 4. 3'te göğüs kanseri verilerine ait doğru sınıflandırma oranları görülmektedir. Kayıp veri oranı arttıkça doğru sınıflandırma oranlarının düştüğü görülmektedir.

Çizelge 4.3. Göğüs kanseri verilerine ait doğru sınıflandırma oranları

Göğüskanseri	tam	%5kayıp	%10kayıp	15% kayıp	20% kayıp	25% kayıp	30% kayıp
k-ortalama	94,27	92,00	91,00	87,00	82,00	82,00	82,00
som	68,00	67,00	66,82	66,12	65,00	65,00	64,00
lvq	96,13	93,00	92,00	91,98	89,67	83,38	81,26

Şekil 4.2'de göğüs kanseri verilerine ait doğru sınıflandırma oranları oranları görülmektedir.



Şekil 4.2. Göğüs kanseri verilerine ait doğru sınıflandırma oranları.

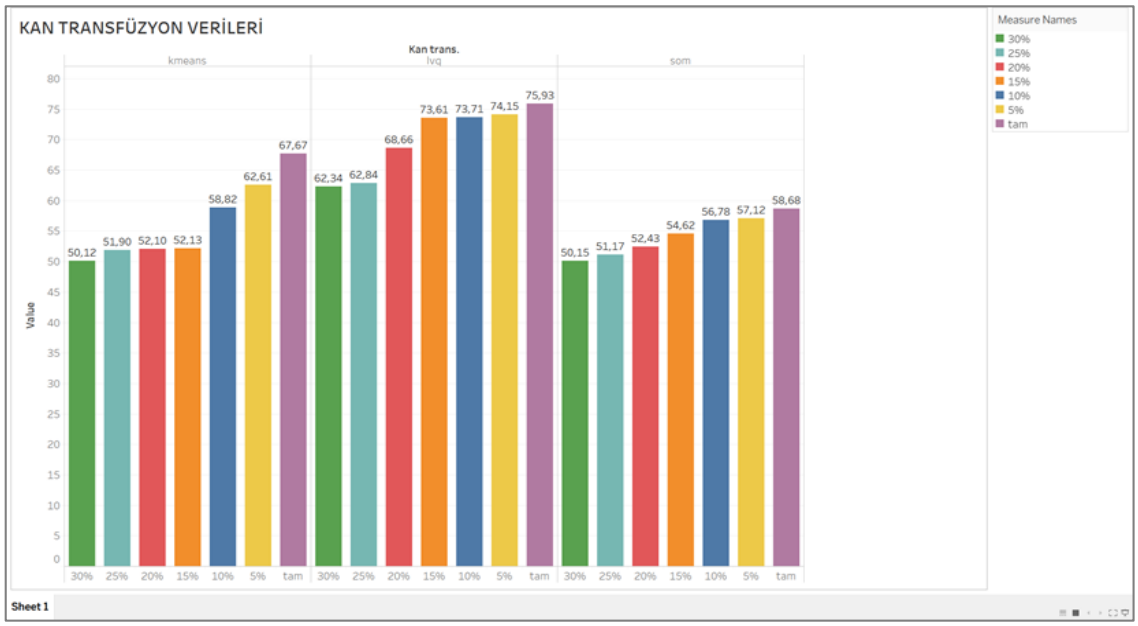
c) Kan Transfüzyon veri seti için değerlendirme

Kan transfüzyon verileri, kan bankasına başvurmuş vericilere ait iki kümeli, nümerik ve nominal veriler içeren bir veri setidir. Çizelge 4.4'te kan transfüzyon verilerine ait doğru sınıflandırma oranları görülmektedir. Kayıp veri oranı arttıkça doğru sınıflandırma oranlarının düştüğü görülmektedir.

Çizelge 4.4. Kan transfüzyon verilerine ait doğru sınıflandırma oranları

Kan trans.	tam	5%	10%	15%	20%	25%	30%
kortalama	67,67	62,61	58,82	52,13	52,1	51,9	50,12
som	58,68	57,12	56,78	54,62	52,43	51,17	50,15
lvq	75,93	74,15	73,71	73,61	68,66	62,84	62,34

Şekil 4.3'te kan transfüzyon verilerine ait doğru sınıflandırma oranları görülmektedir.



Şekil 4.3. Kan transfüzyon verilerine ait doğru sınıflandırma oranları.

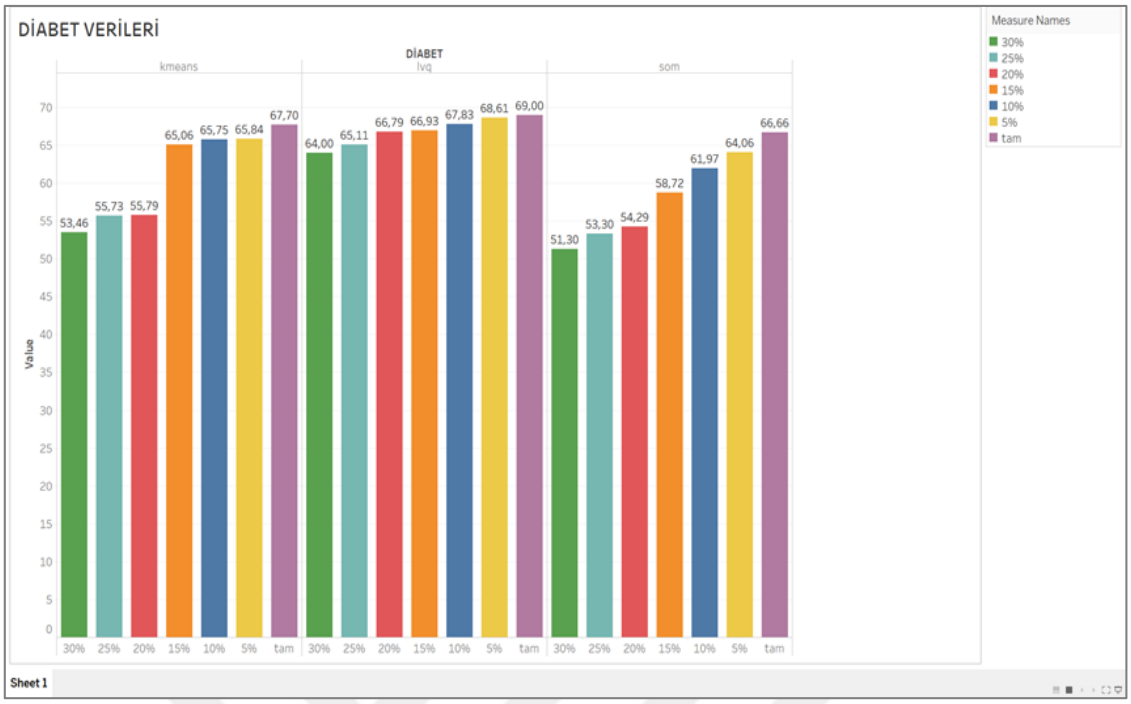
c) Diabet Veri seti için değerlendirme

Diabet, diabet hastalarına ait iki kümeli, nümerik ve nominal veriler içeren bir veri setidir. Çizelge 4.5'te diabet verilerine ait doğru sınıflandırma oranları görülmektedir. Kayıp veri oranı arttıkça doğru sınıflandırma oranlarının düştüğü görülmektedir.

Çizelge 4.5. Diabet verilerine ait doğru sınıflandırma oranları

DİABET	tam	%5kayıp	%10kayıp	15% kayıp	20% kayıp	25% kayıp	30% kayıp
k-ortalama	67,70	65,84	65,75	65,06	55,73	55,73	53,46
som	66,66	64,06	61,97	58,72	54,29	53,30	51,3
lvq	69,00	68,61	67,83	66,93	66,79	65,11	64

Şekil 4.4'te Diabet verilerine ait doğru sınıflandırma oranları görülmektedir.



Şekil 4.4. Diabet verilerine ait doğru sınıflandırma oranları.

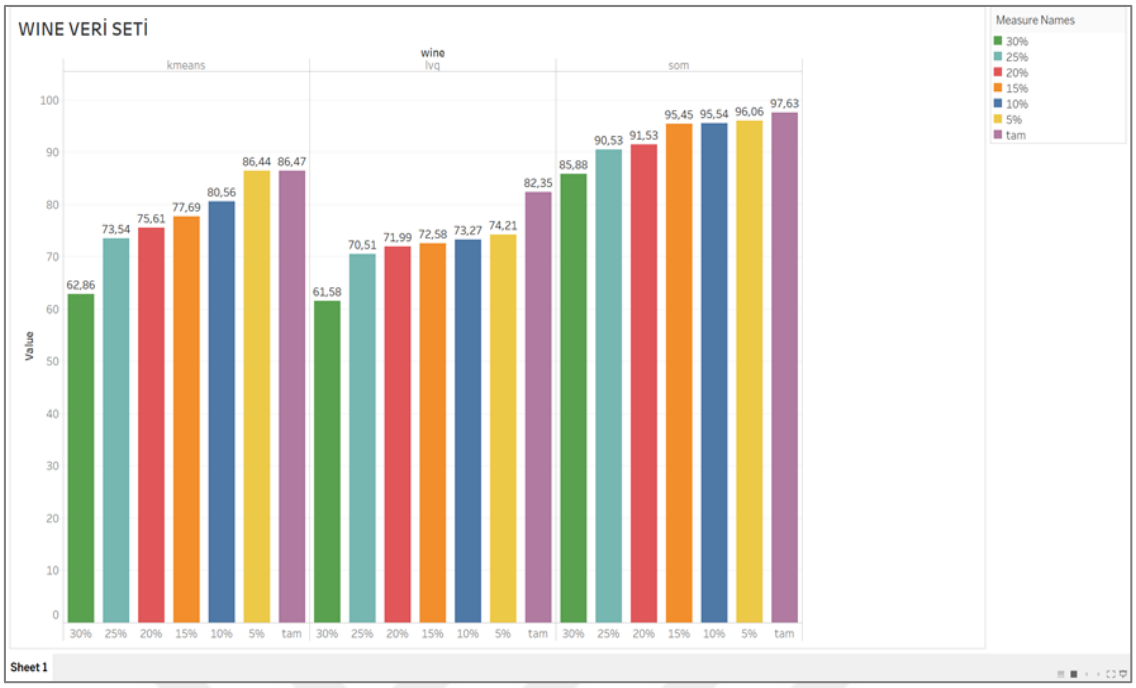
d) Şarap Veri seti için değerlendirme

Şarap verileri, şarap içeriğine ait iki kümeli, nümerik veriler içeren üç kümeli bir veri setidir. Çizelge 4.6'da şarap verilerine ait doğru sınıflandırma oranları görülmektedir. Kayıp veri oranı arttıkça doğru sınıflandırma oranlarının düştüğü görülmektedir.

Çizelge 4.6. Şarap verilerine ait doğru sınıflandırma oranları

şarap	tam	%5kayıp	%10kayıp	15% kayıp	20% kayıp	25% kayıp	30%
k-ortalama	86,47	86,44	80,56	77,69	75,61	73,54	62,86
som	97,63	96,06	95,54	95,45	91,53	90,53	85,88
lvq	82,35	74,21	73,27	72,58	71,99	70,51	61,58

Şekil 4.5'te şarap verilerine ait doğru sınıflandırma oranları görülmektedir.



Şekil 4.5. Şarap verilerine ait doğru sınıflandırma oranları.

4.1.2.2. Orta veri setleri

Orta veri setleri Abalone, kredi kartı, ionosphere, segment ve sensor veri setleridir.

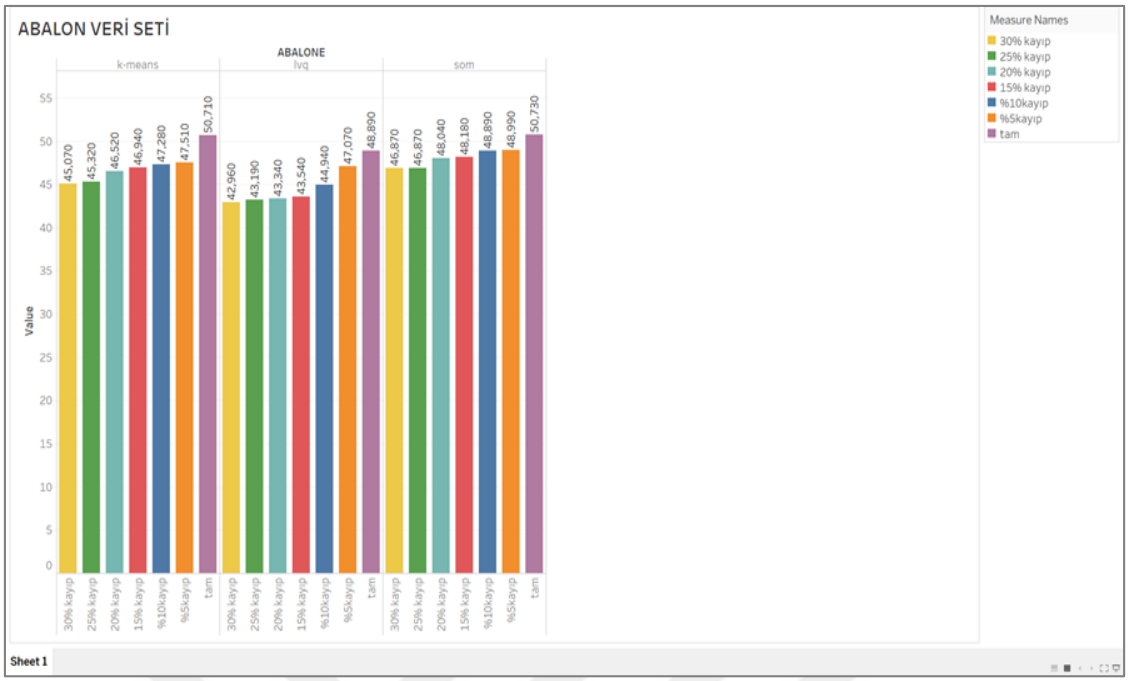
a) Abalone veri setine ait değerlendirme

Abalone verileri, abalone canlısına ait, canlının yaşını tahmin etmek amacıyla oluşturulmuş üç kümeli, nümerik veriler içeren bir veri setidir. Çizelge 4.7'de abalone verilerine ait doğru sınıflandırma oranları görülmektedir. Kayıp veri oranı arttıkça doğru sınıflandırma oranlarının düştüğü görülmektedir.

Çizelge 4.7. Abalone verilerine ait doğru sınıflandırma oranları

ABALONE	tam	%5kayıp	%10kayıp	15% kayıp	20% kayıp	25% kayıp	30% kayıp
k-ortalama	50,71	47,51	47,28	46,64	46,52	45,32	45,07
som	50,73	48,99	48,89	48,18	48,04	46,87	46,87
lvq	48,89	47,07	44,94	43,54	43,34	43,19	42,96

Şekil 4.6'da abalone verilerine ait doğru sınıflandırma oranları görülmektedir.



Şekil 4.6. Abalone verilerine ait doğru sınıflandırma oranları.

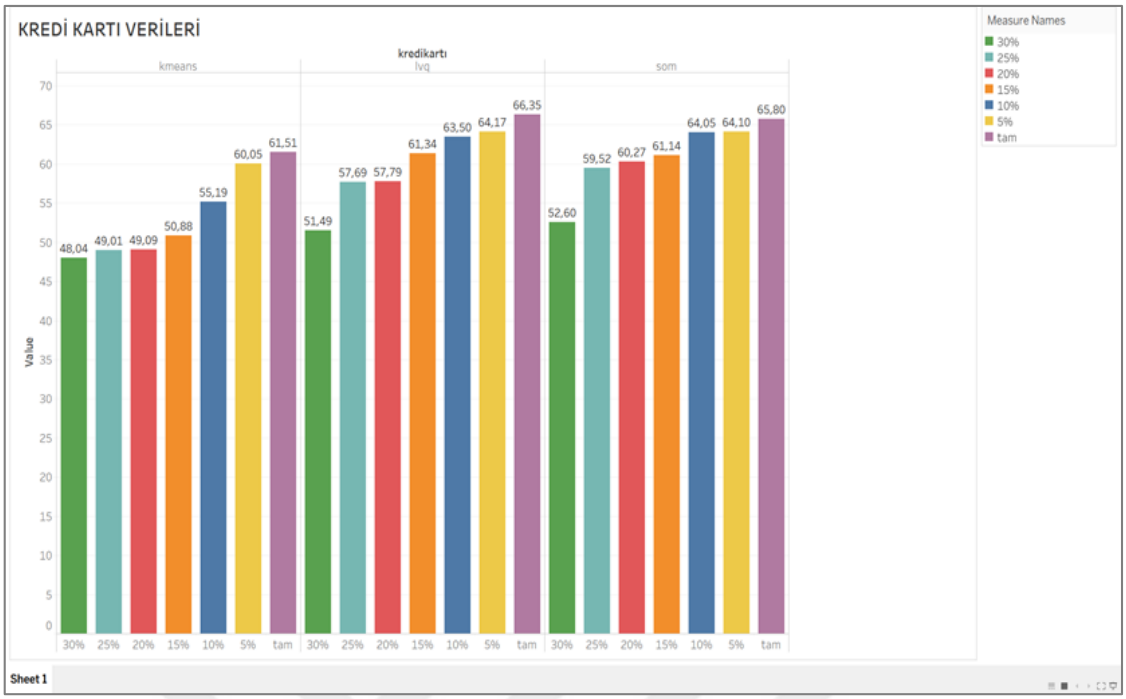
b) Kredi Kartı veri setine ait değerlendirme

Kredi kartı verileri, kredi kartı kullanıcılarının borç durumunu tahmin etmek amacıyla oluşturulmuş, iki kümeli, nümerik ve nominal veriler içeren bir veri setidir. Çizelge 4.8'de kredi kartı verilerine ait doğru sınıflandırma oranları görülmektedir. Kayıp veri oranı arttıkça doğru sınıflandırma oranlarının düştüğü görülmektedir.

Çizelge 4.8. Kredi kartı verilerine ait doğru sınıflandırma oranları

KREDİ KARTI	tam	%5kayıp	%10kayıp	15% kayıp	20% kayıp	25% kayıp	30% kayıp
k-ortalama	61,51	60,05	55,19	50,88	49,09	49,01	48,04
som	65,8	64,1	64,05	61,14	60,27	59,52	52,6
lvq	66,35	64,17	63,50	61,34	57,79	57,69	51,49

Şekil 4.7'de kredi kartı verilerine ait doğru sınıflandırma oranları görülmektedir.



Şekil 4.7. Kredi kartı verilerine ait doğru sınıflandırma oranları.

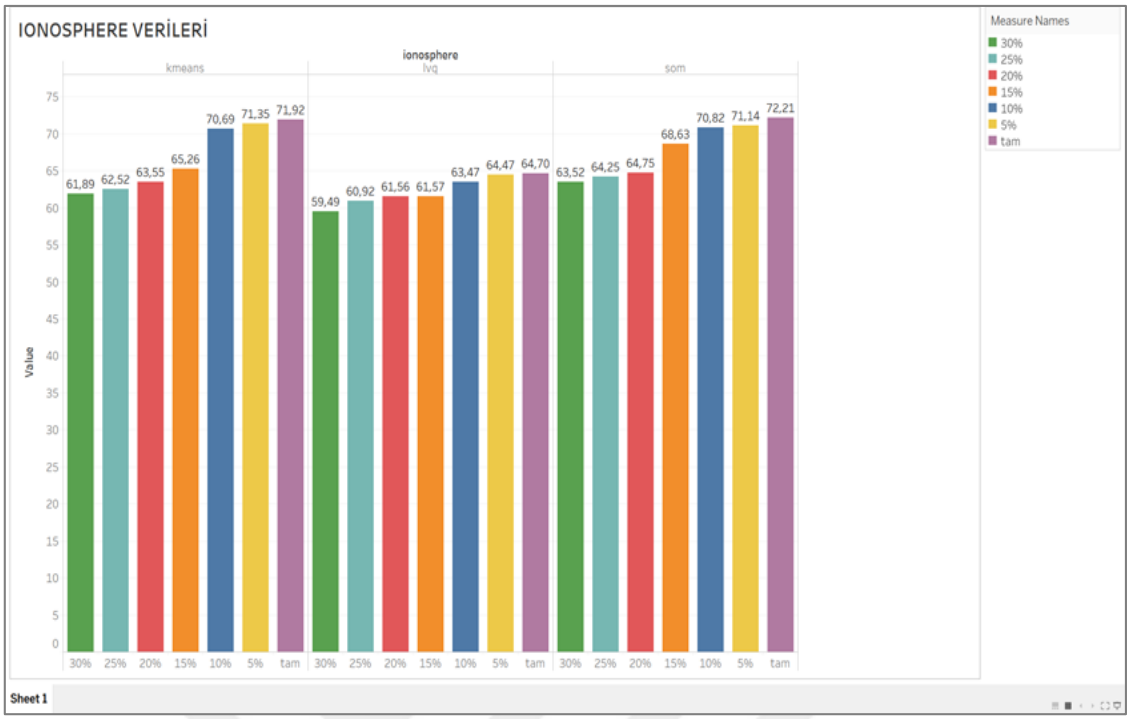
c) Ionosphere veri setine ait değerlendirme

Ionosphere verileri, iyonosfere it iki kümeli, nümerik ve nominal veriler içeren bir veri setidir. Çizelge 4.9'da iyonosfer verilerine ait doğru sınıflandırma oranları görülmektedir. Kayıp veri oranı arttıkça doğru sınıflandırma oranlarının düştüğü görülmektedir.

Çizelge 4.9. Ionosphere verilerine ait doğru sınıflandırma oranları

IONOSPHERE	tam	5%	10%	15%	20%	25%	30%
k-ortalama	71,92	71,35	70,69	65,26	63,55	62,52	61,89
som	72,21	71,14	70,82	68,63	64,75	64,25	63,52
lvq	64,7	64,47	63,47	61,57	61,56	60,92	59,59

Şekil 4.8'de Ionosphere verilerine ait doğru sınıflandırma oranları görülmektedir.



Şekil 4.8. Ionosphere verilerine ait doğru sınıflandırma oranları.

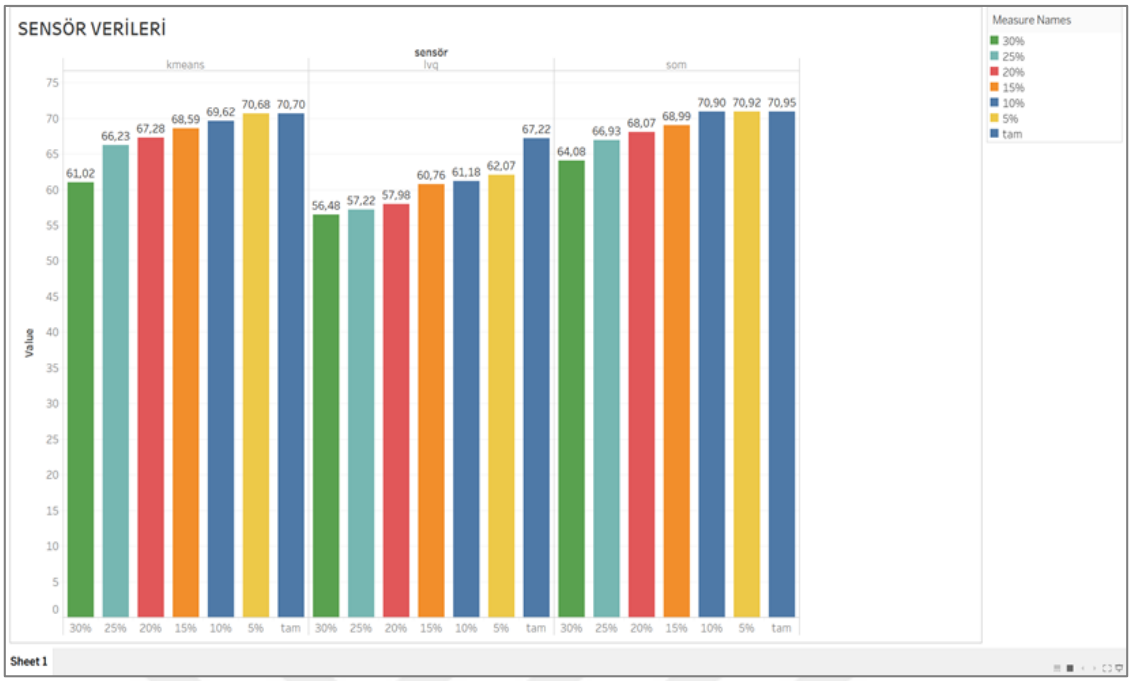
d) Sensör veri setine ait değerlendirme

Sensör verileri, fiziksel ortamdan bir tür girdiyi algılayan ve yanıt veren bir cihazın çıktılarıdır. Nümerik veriler içeren bir veri setidir. Çizelge 4.10'da sensör verilerine ait doğru sınıflandırma oranları görülmektedir. Kayıp veri oranı arttıkça doğru sınıflandırma oranlarının düştüğü görülmektedir.

Çizelge 4.10. Sensör verilerine ait doğru sınıflandırma oranları

SENSÖR	tam	%5kayıp	%10kayıp	15% kayıp	20% kayıp	25% kayıp	30% kayıp
k-ortalama	70,7	70,68	69,62	68,59	67,28	66,23	61,02
som	70,95	70,92	70,9	68,99	68,07	66,93	64,08
lvq	67,22	62,07	61,18	60,76	57,98	57,22	56,48

Şekil 4.9'da sensör verilerine ait doğru sınıflandırma oranları görülmektedir.



Şekil 4.9. Sensör verilerine ait doğru sınıflandırma oranları.

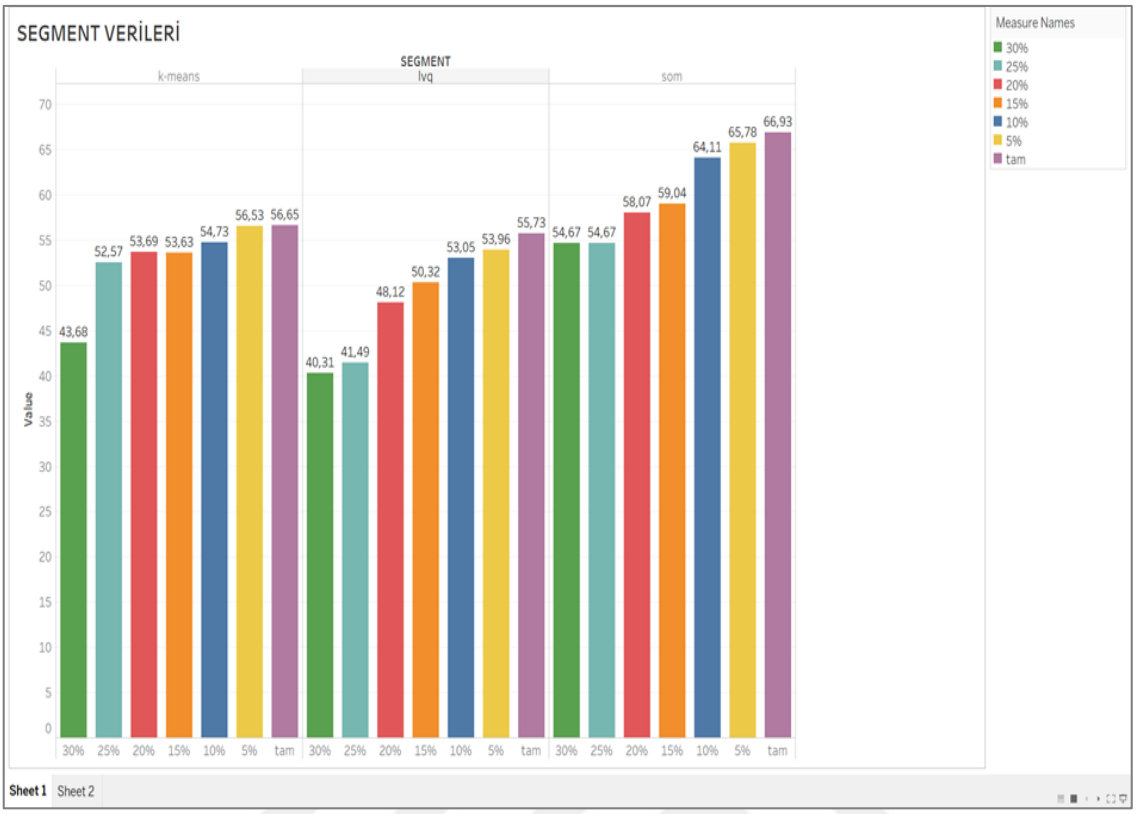
e) Segment veri setine ait değerlendirme

Segment verileri, resim bölümlendirme verileridir. Yedi kümeli, nümerik ve veriler içeren bir veri setidir. Çizelge 4.11'de segment verilerine ait doğru sınıflandırma oranları görülmektedir. Kayıp veri oranı arttıkça doğru sınıflandırma oranlarının düştüğü görülmektedir.

Çizelge 4.11. Segment verilerine ait doğru sınıflandırma oranları

SEGMENT	tam	5%	10%	15%	20%	25%	30%
k-ortalama	56,65	56,53	54,73	53,63	53,69	52,57	43,68
som	66,93	65,78	64,11	59,04	58,07	54,67	54,67
lvq	55,73	53,96	53,05	50,32	48,12	41,49	40,31

Şekil 4.10'da segment verilerine ait doğru sınıflandırma oranları görülmektedir.



Şekil 4.10. Segment verilerine ait doğru sınıflandırma oranları.

4.1.2.3. Büyük veri seti

Büyük veri seti gaz veri setidir.

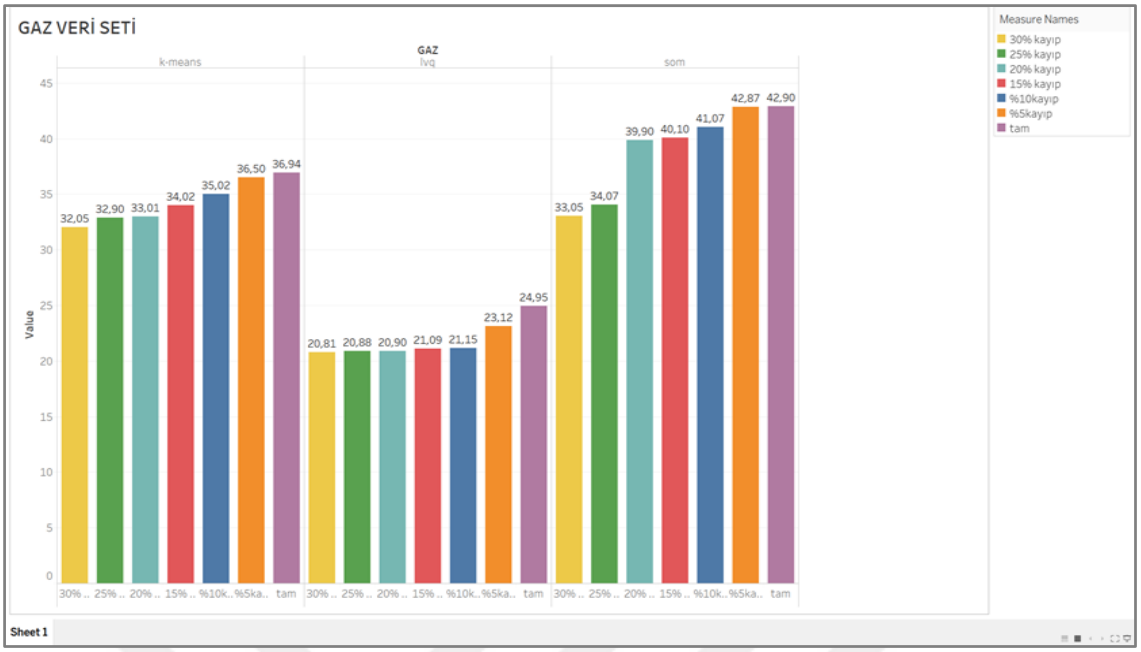
a) Gaz veri setine ait değerlendirme

Gaz verileri, özel olarak tasarlanmış sensörlerin çalışma, sıcaklık kontrolü ve sinyalleri içeren, altı kümeli, nümerik veriler içeren bir veri setidir. Çizelge 4.12'de gaz verilerine ait doğru sınıflandırma oranları görülmektedir. Kayıp veri oranı arttıkça doğru sınıflandırma oranlarının düştüğü görülmektedir.

Çizelge 4.12. Gaz veri setine ait doğru sınıflandırma oranları

Gaz	tam	%5kayıp	%10kayıp	15% kayıp	20% kayıp	25% kayıp	30% kayıp
k-ortalama	36,94	36,5	35,02	34,02	33	32,9	32,05
som	42,9	42,87	41,07	40,1	39,9	34,07	33,05
lvq	24,95	23,12	21,15	21,09	20,9	20,88	20,81

Şekil 4.11'de Gaz verilerine ait doğru sınıflandırma oranları görülmektedir.



Şekil 4.11. Gaz veri setine ait doğru sınıflandırma oranları.

Eşleştirilmiş-t testi sonuçlarına göre veri setlerinin değerlendirilmesi

Weka Experimenter aracı kullanılarak eşleştirilmiş- t testi ile kayıp verili veri setlerinin performansları ikili karşılaştırmalara tabi tutulmuştur. Weka'nın program çıktısında (*) ifadesi önemsizlik, (v) ifadesi de önemlilik anlatır.

İris veri setine ait eşleştirilmiş-t testi sonuçları;

Eşleştirilmiş-t testi sonuçları			
İris	k-ortalama 81.67	LVQ 80.87	SOM **88.67**

Kümeleme sonuçlarına göre 150 nesne, 4 değişken ve 3 kümeye sahip iris verilerinde kullanılan üç yöntemin doğru sınıflandırma oranları eşleştirilmiş -t testi ile değerlendirilmiş, üç yöntem arasında 0,05 anlamlılık düzeyinde anlamlı bir fark bulunamamıştır. Eşleştirilmiş-t testi sonuçları Ek-1 de sunulmuştur. Doğru sınıflandırma oranlarına göre SOM, k-ortalamalar ve LVQ sıralaması en iyi yöntem sıralamasıdır.

Göğüs kanseri veri setine ait eşleştirilmiş-t testi sonuçları;

Eşleştirilmiş-t testi sonuçları			
Göpüs Kanseri	k-ortalama 76.03	LVQ **86.15**	SOM 66.01

Kümeleme sonuçlarına göre 699 nesne, 10 değişken ve 2 kümeye sahip göğüs kanseri verilerinde kullanılan üç yöntemin doğru sınıflandırma oranları eşleştirilmiş -t testi ile değerlendirilmiş, üç yöntem arasında 0,05 anlamlılık düzeyinde anlamlı bir fark bulunamamıştır. Doğru sınıflandırma oranlarına göre LVQ, k-ortalamalar ve SOM sıralaması en iyi yöntem sıralamasıdır.

Kan transfüzyon veri setine ait eşleştirilmiş-t testi sonuçları;

Eşleştirilmiş-t testi sonuçları

	k-ortalama	LVQ	SOM
Kan Tr.	61.21	**65.17**	53.86*

Kümeleme sonuçlarına göre 748 nesne, 5 değişken ve 2 kümeye sahip kan transfüzyon verilerinde kullanılan üç yöntemin doğru sınıflandırma oranları eşleştirilmiş -t testi ile değerlendirilmiş, üç yöntem arasında 0,05 anlamlılık düzeyinde, k-ortalamalar'in SOM'a göre anlamlı bir farkla daha iyi kümeleme yaptığı görülmüştür. Doğru sınıflandırma oranlarına göre LVQ, k-ortalamalar ve SOM sıralaması en iyi yöntem sıralamasıdır.

Diabet veri setine ait eşleştirilmiş-t testi sonuçları;

Eşleştirilmiş-t testi sonuçları

	k-ortalama	LVQ	SOM
Diabet	62.67	**66.78**	57.40*

Kümeleme sonuçlarına göre 768 nesne, 8 değişken ve 2 kümeye sahip diabet verilerinde kullanılan üç yöntemin doğru sınıflandırma oranları eşleştirilmiş -t testi ile değerlendirilmiş, üç yöntem arasında 0,05 anlamlılık düzeyinde, k-ortalamalar yönteminin SOM yöntemine göre anlamlı bir farkla daha iyi kümeleme yaptığı görülmüştür. Doğru sınıflandırma oranlarına göre LVQ, k-ortalamalar ve SOM sıralaması en iyi yöntem sıralamasıdır.

Şarap veri setine ait eşleştirilmiş-t testi sonuçları;

Eşleştirilmiş-t testi sonuçları

	k-ortalama	LVQ	SOM
Şarap	77.60	73.29	**94.15 v**

Kümeleme sonuçlarına göre 178 nesne, 15 değişken ve 3 kümeye sahip şarap verilerinde kullanılan üç yöntemin doğru sınıflandırma oranları eşleştirilmiş -t testi ile değerlendirilmiş, üç yöntem arasında 0,05 anlamlılık düzeyinde, SOM'un k-

ortalamalara göre anlamlı bir farkla daha iyi kümeleme yaptığı görülmüştür. Doğru sınıflandırma oranlarına göre SOM, k-ortalamalar ve LVQ sıralaması en iyi yöntem sıralamasıdır.

Abalone veri setine ait eşleştirilmiş-t testi sonuçları;

Eşleştirilmiş-t testi sonuçları

Ablone	k-ortalama	LVQ	SOM
	47.04	46.34	**48.07**

Kümeleme sonuçlarına göre 4177 nesne, 9 değişken ve 3 kümeye sahip Abalone verilerinde kullanılan üç yöntemin doğru sınıflandırma oranları eşleştirilmiş -t testi ile değerlendirilmiş, üç yöntem arasında 0,05 anlamlılık düzeyinde, önemli bir fark bulunamamıştır. Doğru sınıflandırma oranlarına göre SOM, k-ortalamalar ve LVQ sıralaması en iyi yöntem sıralamasıdır.

Kredi kartı veri setine ait eşleştirilmiş-t testi sonuçları;

Eşleştirilmiş-t testi sonuçları

Kredi Kartı	k-ortalama	LVQ	SOM
	56.60	**58.60**	51.50

Kümeleme sonuçlarına göre 1000 nesne, 21 değişken ve 2 kümeye sahip Kredi kartı verilerinde kullanılan üç yöntemin doğru sınıflandırma oranları eşleştirilmiş -t testi ile değerlendirilmiş, üç yöntem arasında 0,05 anlamlılık düzeyinde, önemli bir fark bulunamamıştır. Doğru sınıflandırma oranlarına göre LVQ- k-ortalamalar ve SOM sıralaması en iyi yöntem sıralamasıdır.

Ionosphere veri setine ait eşleştirilmiş-t testi sonuçları;

Eşleştirilmiş-t testi sonuçları

Ionosphere	k-ortalama	LVQ	SOM
	65.33	62.88*	**70.31**

Kümeleme sonuçlarına göre 351 nesne, 31 değişken ve 2 kümeye sahip Ionosphere verilerinde kullanılan üç yöntemin doğru sınıflandırma oranları eşleştirilmiş -t testi ile değerlendirilmiş, üç yöntem arasında 0,05 anlamlılık düzeyinde, k-ortalamalar'in LVQ ya göre daha iyi kümeleme yaptığı gözlemlenmiştir. Doğru sınıflandırma oranlarına göre SOM, k-ortalamalar ve LVQ sıralaması en iyi yöntem sıralamasıdır.

Sensör veri setine ait eşleştirilmiş-t testi sonuçları;

Eşleştirilmiş-t testi sonuçları

	k-ortalama	LVQ	SOM
Sensör	68.26	60.52*	**68.72**

Kümeleme sonuçlarına göre 2212 nesne, 12 değişken ve 3 kümeye sahip Sensör verilerinde kullanılan üç yöntemin doğru sınıflandırma oranları eşleştirilmiş -t testi ile değerlendirilmiş, üç yöntem arasında 0,05 anlamlılık düzeyinde, k-ortalamalar'ın LVQ' ya göre daha iyi kümeleme yaptığı görülmüştür. Doğru sınıflandırma oranlarına göre SOM, k-ortalamalar ve LVQ sıralaması en iyi yöntem sıralamasıdır.

Segment veri setine ait eşleştirilmiş-t testi sonuçları;

Eşleştirilmiş-t testi sonuçları

	k-ortalama	LVQ	SOM
Segment	56.53	15.73*	**66.93 v**

Kümeleme sonuçlarına göre 1500 nesne, 20 değişken ve 7 kümeye sahip Segment verilerinde kullanılan üç yöntemin doğru sınıflandırma oranları eşleştirilmiş -t testi ile değerlendirilmiş, üç yöntem arasında 0,05 anlamlılık düzeyinde, k-ortalamalar'a göre Som'un daha iyi, LVQ'nun daha kötü kümeleme yaptığı belirlenmiştir. Doğru sınıflandırma oranlarına göre SOM, k-ortalamalar ve LVQ sıralaması en iyi yöntem sıralamasıdır.

Gaz veri setine ait sonuçlar;

Kümeleme sonuçlarına göre 13980 nesne, 128 değişken ve 6 kümeye sahip Gaz verilerinde kullanılan üç yöntemin doğru sınıflandırma oranlarına göre SOM, k-ortalamalar ve LVQ sıralaması en iyi yöntem sıralamasıdır.

Genel olarak sonuçlar incelendiğinde, sürekli verileri kümelemede literatürde de en çok kullanılan gerçek veri setleri kullanılarak yapılan analizlerde k-ortalamalar, SOM ve LVQ kümeleme yöntemlerinin kümeleme performansları tam, yüzde 5, yüzde 10, yüzde 15, yüzde 20, yüzde 25 ve yüzde 30 tamamen rastgele kayıp veri oranlarında değerlendirildi. Veri setleri küçük, orta ve büyük olarak sınıflandırıldı. Genel olarak kayıp veri oranı arttıkça doğru sınıflandırma oranının düştüğü görülmektedir. Veri setlerinin içeriğine bağlı olarak en iyi kümeleme sıralaması değişmektedir. İncelenen veri setlerinden nominal ve nümerik veriler içeren ve iki küme olan verilerin LVQ ile en iyi kümelendiği görülmektedir. LVQ yapısı gereği Kohonen tabakasından sonra

destekleyici öğrenme gereği verileri 1 veya 0 olmaya zorlar. Bu da incelenen veri setlerinin sonuçlarını doğrular. Nümerik veriler içeren veri setlerinde ise SOM' un en iyi kümeleme performansı gösterdiği görülmektedir. Büyük veri seti incelendiğinde kayıp veri oranının artmasının doğru kümeleme oranlarını çok fazla etkilemediğini görülmektedir.

5. TARTIŞMA VE SONUÇ

Literatürde de en çok kullanılan gerçek veri setleri kullanılarak yapılan analizlerde tam ve kayıp veri içeren veri setlerinde, k-ortalamalar, SOM ve LVQ kümeleme yöntemlerinin kümeleme performansları tam, yüzde 5, yüzde 10, yüzde 15, yüzde 20, yüzde 25 ve yüzde 30 tamamen rastgele kayıp veri oranlarında değerlendirildi. Veri setleri küçük, orta ve büyük olarak sınıflandırıldı.

Bu tezde, verilerin tam ve tamamen rastgele kayıp olduğu durumlarda kümeleme performansları test edilmiştir. Çalışmada kullanılan veri setleri çok değişkenli, nümerik veya nominal veriler olarak seçilmiştir. Bu nedenle, bu sonuçların kapsamı, verileri tamamen rastgele kayıp olan nümerik ve nominal verilerle sınırlıdır.

Kümeleme, veri setinde yer alan benzer nesnelerin aynı gruplarda yer alacak şekilde yerleştirilmesidir. Kümeleme yöntemlerinde birçok seçenek vardır. Kayıp verileri işleme yöntemleri de büyük çeşitlilik gösterir ve bu seçimler genellikle kayıp veri mekanizmasına göre yapılır. Bu çalışmada kümeleme yöntemlerinden bazılarının kayıp veri varlığında nasıl kümeleme yaptığı araştırıldı. Bölümlemeli kümeleme yöntemlerinden k-ortalamalar ve YSA tabanlı kümeleme yöntemlerinden SOM ve LVQ ile ilgilenildi. k-ortalamalar yöntemi kümeleme problemlerini çözmede en fazla bilinen, denetlenmeyen öğrenme sınıfında olan bir yöntemdir. k-ortalamalar yöntemi, bir dizi nesneyi, özniteliklerine ve özelliklerine bağlı olarak, k küme sayısının kullanıcı tarafından tanımlanan bir sabit olduğu k adet kümeye ayırmayı amaçlar. Gerçekleştirilen çalışma sonucunda kullanılan tüm veri setlerinde k-ortalamalar yönteminin doğru sınıflandırma oranlarında çok fazla değişme olmadığı gözlemlenmektedir. Buna göre k-ortalamalar yönteminin kayıp veriyi tolere edebildiği söylenebilir.

Yapay sinir ağı tabanlı kümeleme yöntemleri; YSA terminolojisinde kümeleme yöntemleri danışmansız öğrenme kategorisindedir. Eğitim için etiketlenmiş veya sınıflandırılmış veri setine ihtiyaç duyulmaz. Veri setindeki benzer desenleri inceleyerek öğrenilen benzerliklerine göre veri setini olabilecek en iyi şekilde kümelemeye çalışırlar. Yapılan çalışmada bu ağlardan SOM ve LVQ kullanılmıştır. Danışmansız öğrenen SOM ağlarda nöronlar arasında bir yarışma söz konusudur. SOM, tipik olarak

iki katmandan oluşur. Bu katmanlar, girdi katmanı ve iki boyutlu Kohonen çıktı katmanıdır. Girdi katmanı, Kohonen katmanındaki tüm nöronlara bağlantılıdır. Yarışmacı öğrenmeyi kullanan Kohonen ağında kazanan nöron 1 diğerleri ise 0 değerini alır. Bu stratejiye kazanan hepsini alırstratejisi denir. Kazanan nöronun belirlenmesinin ardından bu nöronun ve komşularının ağırlıkları değiştirilir. Yani kazanan nöronun çevresindeki elemanlar kazanan nöronla aynı cevabı vermesi için desteklenir. Eğitim esnasında öğrenme katsayısı sürekli olarak düşürülür ve komşuluk alanı sürekli daraltılır. SOM bu çalışmada kullanılan iki küçük, dört orta ve bir büyük veri setinde tüm doğru sınıflandırma oranlarında en iyi kümeleme sonuçlarını vermiştir. Bu veri setleri nümerik verilerden oluşmaktadır. SOM, k-ortalamalar yönteminden farklı olarak bir YSA yöntemidir ve kümeleri ayırırken küme merkezlerini değil, tüm verileri kullanır. Bu durumun kümeleme başarısını arttırdığı gözlemlenmiştir.

LVQ ağlarının eğitimindeki amaç her iterasyonda girdi vektörüne en yakın referans vektörünü bulmak, yani, küme merkezlerini(Cluster Center) ayarlamak ve eğitim setindeki tüm girdi vektörlerinin niceleme hatalarını minimize etmektir. Referans vektörleri Kohonen katmanındaki nöronları girdi katmanındaki nöronlara bağlayan ağırlık değerleridir. Öğrenme esnasında sadece referans vektörlerinin ağırlık değerleri değiştirilmektedir. Bu işlem Kohonen öğrenme kuralı kullanılarak yapılmaktadır (Cozart, 1996). Kohonen katmanı ile çıktı katmanı arasındaki ağırlıklar eğitim sırasında değiştirilmemektedir. Kohonen katmanındaki nöronların çıktıları, bu nöronları çıktı katmanına bağlayan ağırlık değerleri ile çarpılarak ağın çıktısı hesaplanmaktadır. Bu Kohonen katmanında yarışmayı kazanan nöron elemana bağlı olan çıktı elemanın değerinin 1, diğerlerinin değerinin 0 olması anlamına gelmektedir. Ağın çıktıları belirlendikten sonra çıktının doğru sınıflandırılıp sınıflandırılmadığı sorgulanır. Bu sorunun cevabına göre Kohonen katmanındaki yarışmayı kazanan nöron elemanı girdi katmanına bağlayan ağırlıklar değiştirilmektedir. Bu nedenle LVQ ağları Destekleyici Öğrenme sınıfındadır. Tezde kullanılan üç küçük, bir orta veri seti hem nümerik hem de nominal veriler içermektedir. Bu veri setleri kümeleme performansı açısından değerlendirildiğinde LVQ' nun daha iyi kümeleme yaptığı gözlemlenmiştir. Bu veri setlerinde çıktı değerinin sadece iki kümeden oluşmasu LVQ yönteminin daha iyi sonuçlar üretmesini açıklamaktadır.

Genel olarak kullanılan veri atama yöntemlerinin aksine, kayıp verili veri setlerinde de çalışma yapılabilirliğini savunan çalışmalar 2013 yılından sonra gelişmeye başlamıştır. Orczyk ve Porwik, (2013)' te yaptıkları bir çalışmada kayıp verileri - özellikle tıbbi verileri - doldurmanın tehlikelerini göstermektir.

Juhola ve Laurikkala, (2013)' te beş veri setinde kayıp değerlerin gerçek pozitif oranlar ve sınıflandırma doğruluğu üzerindeki etkisini tespit etmek için yaptıkları çalışmada KNN, Diskriminant Analizi ve Naif Bayes yöntemlerini kullandıkları çalışmada, iki sınıflı veri kümelerinde, % 20-30'a kadar kayıp değerlere rağmen, kayıp değerin olmadığı kadar iyi sonuçların üretilebileceğini göstermişlerdir. Gerçekleştirilen tez çalışmasında elde edilen doğru sınıflandırma oranları değerlendirildiğinde, kayıp veri oranları arttıkça doğru sınıflandırma oranlarının düştüğü fakat kayda değer farklar olmadığı gözlenmiştir.

Zhu ve Shi (2018), kayıp veriler için yeni bir destek vektör makinesi algoritması önerdikleri çalışmalarında veri atama yöntemleri yerine kayıp değerler doldurulmasından kaynaklanan hatayı azaltmak için gözlenen verilerin tam olarak kullanılmasını önermişlerdir. Yöntem doğrulamak için de doğruluk, F skoru, Kappa istatistikleri kullanmışlardır. Gerçekleştirilen tez çalışmasında da, farklı kümeleme yöntemleri ile kayıp verileri doldurma yerine, farklı kayıp veri oranları ile yöntemlerin çalışma performansları doğru sınıflandırma oranları kullanılarak test edilmiştir.

Analiz sonuçlarına göre kullanılan on bir farklı veri setinde genel olarak kayıp veri oranı arttıkça doğru sınıflandırma oranının düştüğü görülmektedir. Veri setlerinin içeriğine bağlı olarak en iyi kümeleme sıralaması değişmektedir. İncelenen veri setlerinden nominal veriler içeren ve çıktının iki kümeden oluştuğu verilerin LVQ ile en iyi kümelendiği görülmektedir. LVQ yapısı gereği Kohonen tabakasından sonra destekleyici öğrenme ile verileri 1 veya 0 olmaya zorlar. Bu da incelenen veri setlerinin sonuçlarını doğrular. Nümerik veriler içeren veri setlerinde ise SOM yönteminin en iyi kümeleme performansı gösterdiği görülmektedir. Büyük veri seti incelendiğinde ise kayıp veri oranının artmasının doğru kümeleme oranlarını çok fazla etkilemediğini görülmektedir.

Gelecekte konu ile ilgili gerçekleştirilebilecek bazı çalışmalar aşağıda maddelenmiştir:

- Kayıp verinin rasgeleliği ihlal edilerek çalışma tekrarlanabilir.

- Farklı kümeleme yöntemleri (Canopy, Cobweb, EM, Farthest First, FilteredClusterer, Hierarchical Clusterer, Make Density Based Clusterer) ile farklı karşılaştırma kombinasyonları oluşturulabilir.
- Kümeleme yöntemlerinin kullanabildiği farklı uzaklık fonksiyonları denenebilir.
- Kayıp veri oranları değiştirilerek analizler tekrarlanabilir.
- Kategorik veriler kullanılarak, farklı veri setlerinde, farklı oranlarda kayıp veri performansları test edilebilir.

KAYNAKLAR

Akpınar H., 2014. ***Data***. Papatya Yayıncılık. İstanbul.

Aladağ Ç.H., 2009. ***Yapay Sinir Ağlarının Mimari Seçimi İçin Tabu Arama Algoritması*** (doktora tezi). Hacettepe Üniversitesi Fen Bilimleri Enstitüsü. Ankara.

Allison, Paul D.,2009. ***Missing Data***. The SAGE Handbook of Quantitative Methods in Psychology, edited by Roger E. Millsap and Alberto Maydeu-Olivares. Thousand Oaks, CA: Sage Publications Inc.

Alkarkhi A.F.M ve Alqaraghuli W.A.A.,2019. ***Cluster Analysis***. Easy Statistics for Food Science with R.,177-86.

Alpar, R.,2011. ***Uygulamalı Çok Değişkenli İstatistiksel Yöntemlere Giriş1***. Nobel Kitabevi, Ankara.

Alpdoğan Y.,2007. ***Kendinden Düzenlenen Haritalar ile Doküman Sınıflandırma*** (Yüksek Lisans Tezi). Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

Anonim, 2018. ***https://www.saedsayad.com/naive_bayesian.htm***. Erişim Tarihi: 13.09.2018

Anonim, 2018. ***https://tr.sciencewal.com/10226-understanding-neural-networks-from-neuron-to-rnn-cnn-and-deep-learning-cd88e90e0a90-42***. Erişim Tarihi: 15.07.2018

Anonim, 2018. ***http://veribilimci.org/yapay-sinir-aglari-ysa-nedir-bolum-2***. Erişim Tarihi: 15.09.2018

Anonim, 2018. ***https://tr.sciencewal.com/10226-understanding-neural-networks-from-neuron-to-rnn-cnn-and-deep-learning-cd88e90e0a90-43.*** Erişim Tarihi: 16.09.2018

Arous, N. ve Ellouze, N., 2010. On the search of organization measures for a Kohonen map case study. ***Speech signal recognition***. ***International Journal of Digital Content Technology and its Applications***, **4**(3): 75-84.

Ayadi, T., Hamdani, T. M. ve Alimi, A. M., 2012. MIGSOM: Multilevel interior growing self-organizing maps for high dimensional data clustering. ***Neural Processing Letters***, **36**(3): 235-256.

Baker S.G. ve Laird N.M.,1988. Regression analysis for categorical variables with outcome subject to non-ignorable non-response. ***Journal of the American Statistical Association***, **83**: 62-69.

Baker S.G., Rosenberger W.F. ve Der Simonian R.,1992. Closed-form estimates for missing counts in two-way contingency tables. ***Statistics in Medicine,*** **11**: 643-657.

Bal, C., 2003. ***Çok Gruplu Veri Setlerinde Eksik Gözlem Sorununun Çözümlenmesi ve Sağlık Alanında bir Uygulama*** (Doktora Tezi). Osmangazi Üniversitesi, Fen Bilimleri Enstitüsü, Eskişehir.

Batista, G. E. ve Monard, M. C.,2003. An analysis of four missing data treatment methods for supervised learning. ***Applied Artificial Intelligence***, **17** (5-6): 519–533.

Baygül A.,2007**.** ***Eksik veri analizinde kullanılan etkili yöntemlerin değerlendirilmesi*** (Yüksek Lisans Tezi). İstanbul Üniversitesi, Sağlık Bilimleri Enstitüsü, İstanbul.

Baş, N., 2006. ***Yapay Sinir Ağları Yaklaşımı ve bir Uygulama*** (Basılmamış Yüksek Lisans Tezi). Mimar Sinan Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Burns, R. P. ve Burns, R.,2008***. Business Research Methods and Statistics Using SPSS***. SAGE, ISBN: 978-1-4462-0476-4.

Brugger, D., Bogdan, M. ve Rosenstiel, W., 2008. Automatic cluster detection in Kohonen's SOM. ***IEEE Transactions on Neural Networks***, **19**(3): 442-459.

Bolshakova, N. ve Azuaje, F., 2003. Cluster Validation Techniques for Genome Expression Data. ***Signal Processing***, **83**(4): 825-833.

Brecheisen, S., Kriegel, H. ve Pfeifle, M., 2003. Efficient density-based clustering of complex objects. ***4th IEEE International Conference on Data Mining***, United Kingdom, 43–50.

Bridge, D.,2013. Classification: k Nearest Neighbours. ***http:// www.cs.ucc.ie/dgb/courses/ tai/notes/handout4.pdf.*** Erişim Tarihi: 26.04.2018.

Caruana R.,Elhawary M., Nguyen N.,Smith C., 2006. Meta clustering. ***Sixth Int. Conf. on Data Mining (ICDM'06***) 107-118.

Carpenter, J., Bartlett, J. ve Kenward, M., 2015. Introduction to missing data. ***http://missingdata.lshtm.ac.uk.*** Erişim Tarihi: 16.02.2018.

Cheng, B.ve D. M. Titterington, 1994. Neural Networks: A Review from a Statistical Perspective. ***Statistical Science*** **(9):** 2-54.

Chen, S. F., Wang, S., ve Chen, Y. C., 2011. A simulation study using EFA and CFA programs based the impact of missing data on test dimensionality. ***Expert Systems with Applications*** **(39)**: 4026–4031.

Chi, S. C. ve Yang, C. C., 2006. Integration of ant colony som and k-ortalamalar for clustering analysis. ***International Conference on Knowledge-Based Intelligent Information and Engineering Systems***, 1–8.

Chi, S. C. ve Yang, C. C., 2008. A two-stage clustering method combining ant colony SOM and k-ortalama. ***Journal of Information Science and Engineering***, **24**(5): 1445-1460.

Clayton D. ve Hills M., 1993. ***Statistical Methods in Epidemiology.*** Oxford University Press, Oxford.

Cohen J. A., 1960. Coefficient of agreement for nominal scales. ***Educational and Psychological Measurement***, **(XX):**1.

Conaway M.R., 1992. The analysis of repeated categorical measurements subject to nonignorable nonresponse. ***Journal of the American Statistical Association,*** **(87):** 817-824.

Conaway M.R., 1993. Non-ignorable non-response models for time-ordered categorical variables. ***Applied Statistics,*** **(42**): 105-115.

Cottrell, M., Gaubert, P., Eloy, C., Francois, D., Hallaux, G., Lacaille, J. ve Verleysen, M., 2009. Fault prediction in aircraft engines using self-organizing maps. ***International Workshop on Self-Organizing Maps***. Springer Berlin Heidelberg, 5629, 37-44.

Cozart M. T., 1996. ***Evaluation of 'The Neural Gas' Network Vector Quantization and Approximation Components*** (Master of Science Thesis). The University of Tennessee, Knoxville.

Czepiel, S. A., 2002. Maximum likelihood estimation of logistic regression models: theory and implementation. ***http://ww.saedsayad.com/docs/mlelr.pdf***. Erişim tarihi: 27.03.2018.

Çanakçı, A., 2006. ***Yapay Sinir Ağlarının Makroekonomik Bir Model Üzerine Uygulaması*** **(**Yüksek Lisans Tezi). Gazi Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Davies D. ve Bouldin D., 1979**.** A cluster separation measure. ***IEEE PAMI***, **1**(2): 224–227.

Dawson B. ve Trap R.G., 2004. ***Basic and Clinical Biostatistics***. Lange Medical Books/McGraw-Hill, Third Edition.

Demir E., 2013. Eksik veri varlığında çoktan seçmeli testlerde madde ve test parametrelerinin tahmini: SBS örneği. ***Eğitim Bilimleri Araştırmaları Dergisi***, **3** (2): 48-68.

Dempster A.P, Laird N.M ve Rubin D.B., 1997. Maximum likelihood from incomplete data via the EM algorithm. ***JRSSB*** **39**: 1-38.

Dudoit, S., Fridlyand, J., 2002. A prediction-based resampling method for estimating the number of clusters in a dataset. ***Genome Biology***, **3**(7): 1-21.

Dumitrescu, D., Lazzerini, B. ve Jain, L., 2000. Fuzzy Sets and Their Application to Clustering and Training. ***CRC Press LLC***, **2**(5):1-13.

Dunn J., 1974. Well separated clusters and optimal fuzzy partitions. ***J. Cybern***, **4**(1): 95–104.

Durrant, G. B., 2005. Imputation methods for handling item-nonresponse in the social sciences: a methodological review**.** ***http://missingdata.lshtm.ac.uk/preprints/durrantOct05.pdf.*** Erişim Tarihi: 12.05.2018.

Donner A., 1982. There lative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values. **Am Stat, 36**: 378-81.

Enders, C. K., Bandalos, D. L., 2001. The relative performance of full information maximum likelihood estimation for missing data in structural equation models. Structural Equation Modeling. ***A Multidisciplinary Journal***. **8**(3): 430-457.

Enders, C. K.,***2010. Applied Missing Data Analysis***. New York: The Guilford Press.

Elmas, Ç., 2010. ***Yapay Zeka Uygulamaları***. Seçkin Yayıncılık, Ankara.

Ertorsun, A. D., Bağ, B., Uzar, G. ve Turanoğlu, M. A., 2009. ROC (Receiver Operating Characteristic) Eğrisi Yöntemi ile Tanı Testlerinin Performanslarının Değerlendirilmesi. ***XIII. Öğrenci Sempozyumu***, Eylül, Ankara.

Fisher, D., 1987. Knowledge acquisition via conceptual clustering. ***Machine Learning***, **(2**)2: 139–172.

Fiona, M. S., Heather, S., Hude, G., ve William G., 2006. Dealing with missing data in a multi-question depression scale: A comparison of imputation methods**.** ***BMC Medical Research Methodology,*** **6**:57.

Friedman L.M., Furberg C.D. ve De Mets D.L.,1998. ***Fundamentals Of Clinical Trials***. Springer, New York.

Flach, P., 2004. ***Tutorial on "The Many Faces of ROC Analysis in Machine Learning".*** **Tutorial Notes. Banff, Alberta, Canada, ICML 2004**, **1**: 49.

Frank, E., Holmes, G., Kirkby, R., ve Hall, M., 2002. Racing committees for large datasets. ***In Proceedings of the International Conference on Discovery Science,*** **(1)**:153–164.

Gorunescu, F., 2011. ***Data Mining Concepts Models, Methods and Algorithms.*** Springer.

Grabmaier, J. ve Rudolph, A., 2002. Techniques of cluster algorithms in data mining. ***Data Mining and Knowledge Discovery***, **6**: 303–360.

Goodwin, L. D., 2001. Interrater agreement and reliability. ***Measurement in Psychical Education and Exercises Science***, **5**(1): 13-14.

Greenlees, W.S., Reece, J.S., ve Zieschang, K.D., 1982. Imputation of missing values when the probability of response depends on the variable being imputed. ***J.Am Statist. Assoc***., (**77**): 251-261.

Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J. ve Bossuyt, P. M. M., 2003. **The diagnostic odds ratio: a single indicator of test performance**. ***Journal of Clinical Epidemiology.***

Gold, M. S., ve Bentler, P. M., 2000. Treatment of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation and expectation-maximization. ***Structural Equation Modelling*** **7**: 319–355.

Graham, J. W., Hofer, S. M., ve Mackinnon, D. P., 1996. Maximizing the usefulness of data obtained with planned missing value patterns: an application of maximum likelihood procedures. ***Multivariate Behavioral Research***, **31**: 197–218.

Green S.,Benedetti J. ve Crowley J.,1997. ***Clinical Trials In Oncology***. Chapman and Hall, London.

Gorgonio, F. L. ve Costa, J. A. F., 2008. Combining parallel self-organizing maps and k-ortalamalar to cluster distributed data. ***11th IEEE International Conference on Computational Science and Engineering Workshops***, 53-58.

Günay S., Eğrioğlu E. ve Aladağ Ç.H., 2007. ***Tek Değişkenli Zaman Serileri Analizine Giriş***. Hacettepe Üniversitesi Yayınları. Ankara.

Güzeller C. O. ve Aksu G., 2018. ***Matlab Yapay Zeka Uygulamaları***. Maya Akademi. Ankara.

Gwet K. 2010. ***Handbook of Inter-Rater Reliability*** (2.Ed.). ***https://books.google.com.tr/books***. Erişim Tarihi: 15.06.2018.

Halkidi M., Batistakis Y., Vazirgiannis M., 2001. On clustering validation techniques**.** ***Journal of Intelligent Information Systems*** **(17)**: 107–145.

Harabasz T. ve Calinski J., 1974. A dendrite method for cluster analysis. ***Comm. İn Statistics***, **3**(1): 1–27.

Hamer R.M. ve Simpson P.M., 2009. Last Observation Carried Forward Versus Mixed Models in the Analysis of Psychiatric Clinical Trials. Am J Psychiatry Published online ***https://doi.org/10.1176/appi.aip.2009.09040458*.** Erişim tarihi: 23.05.2017.

Hamzaçebi C., 2007. Forecasting of Turkey's net electricity energy consumption on sectorial bases. ***Information Sciences***, **178**(23):4550-4559.

Han, J. ve Kamber, M.,2001. ***Data Mining Concepts and Techniques***. Morgan Kaufmann Publishers.

Han J., Kamber M., 2006. ***Data Mining Concepts and Techniques***. San Francisco: Elsevier Inc.

Han, J., Kamber, M., Pei, J., 2012. ***Data Mining Concepts and Techniques.*** Morgan Kaufman.

Harrington, P., 2012. ***Machine Learning in Action***. 1st Edition, Manning Publications Shelter Island, NY, ISBN: 978-1-61729-018-3.

Hastie, T., Tibshirani, R., Friedman, J., 2008. ***The Elements of Statistical Learning: Data Mining, Inference, and Prediction***. 2nd Edition, Springer, ISBN: 0-387-84857-6.

Hebb D.O., 1949. ***The Organization of Behavior A Neuropsychological Theory.*** Mc Gill University.

Hopfield, J. J., 1990. The Effectiveness of Analogue Neural Network Hardware. ***Network: Computation in Neural Systems***, 1: 27-40.

Hoppner, F., Klawonn, F., Kruse, R. ve Runkler, T., 1999. Fuzzy Cluster Analysis:Methods for Classification. ***Data Analysis and Image Recognition***. John Wiley & Sons.

Hubertand L., Arabie P.1985. Comparing partitions. ***Journal of Classification***, **2**(1): 193–218.

Heyting A., Tolboom J.T.B.M. ve Essers J.G.A.,1992. Classification. ***Statistics in Medicine***. **11**: 2043-2061.

Hippel, P. T. V., 2004. Biases in SPSS 12.0 Missing Value Analysis. ***The American Statisticion***, **58**: 2.

Işık, M. ve Çamurcu, A.Y., 2007. K-ortalama, K-medoids ve Bulanık c-ortalama Algoritmalarının Uygulamalı Olarak Performanslarının Tespiti. ***İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi***, **6**(11): 31-45.

Juhola, M., ve Laurikkala, J., 2013. Missing values: how many can they be to preserve classification reliability?. ***Artificial Intelligence Review***, **40**(3): 231– 245.

Karabulut, E. ve Alpar, R., 2011. ***Lojistik Regresyon, Uygulamalı Çok Değişkenli İstatistiksel Yöntemler***. Detay Yayıncılık Ankara, ISBN: 978-605-5437-42-8.

Khanmohammadi S, Adibeig N, Shanehbandy S., 2017. An improved overlapping k-means clustering method for medical applications. ***Expert Systems With Applications,*** **67**: 12-18.

Kohonen, T., 1993. Things you haven't heard about theSelf-Organizing Map. ***In Neural Networks. IEEE International Conference,*** 1147–1156.

Kohonen, T., 2001. ***Self-organizing Maps***, Springer Series in Information Sciences. Springer Berlin.

Kohonen, T., 2012. ***Self-organization and Associative Memory***. Springer.

Kromrey, J. D., ve Hines, C. V., 1994. Nonrandomly missing data in multiple regression: An empirical comparison of common missing-data treatments. **Educational ve Psychological Measurement, 54**: 573–593.

Kröse, B. ve Smagt, P.V.D., 1996. ***An Introduction to Neural Networks.*** The University of Amsterdam.

Kim J.O ve Curry J., 1977. ***The Treatment of Missing Data in Multivariate Analysis***. Social Methods. University of Iowa.

Larose, D. T., 2005. Discovering knowledge in data: an introductionto data mining. Statistics. ***https://doi.org/10.1016/j.cll.2007.10.008.*** Erişim Tarihi: 18.04.2017.

Little, R.J.A.ve Rubin, D.B., 1987. ***Statistical analysis with missing data***. John Wiley & Sons, Inc., USA.

Little R.J. A.,1988. A Test of Missing Completely at Random for Multivariate Data with Missing Values. ***Journal of the American Statistical Association***, **83**(404): 1198-1202.

Little R.J.A. ve Rubin D.B., 2002. ***Statistical Analysis with Missing Data***. John Wiley ve Sons, New York.

Langley, P. 1996. ***Elements of machine learning.*** Morgan Kaufmann, ISBN: 1-55860-301-8.

Mallinckrodt B., Gantt D.L. ve Coble H.M.,1995. Attachment patterns in the psychotherapy relationship: Development of the client attachment to therapist scale. ***Journal of Counselling Psychology***, **56**: 549-563.

MacQueen, J.,1967. Some methods for classification and analysis of multivariate observations***. 5th Berkeley Symp. Mathemtaical Statistics&Probability,*** **1**: 281–297.

Mohri, M., Rostamizadeh, A., Talwalkar, A., 2012. ***Foundations of Machine Learning***. The MIT Press, ISBN: 0-262-01825-X.

Mooi E, Sarstedt M., 2011. ***Cluster Analysis: A Concise Guide to Market Research.*** Berlin Heidelberg: Springer, 237-84.

Navarro, J. B. ve Losilla, J. M., 2000. ***Analysis of incomplete data with artificial neural networks:*** A simulation study. Psicothema 12: 503–510.

Neter, J., Wasserman W. ve Kutner M.H., 1989. ***Applied Linear Regression Models.*** Boston: IRWIN.

Nemalhabib A.ve Shiri N., 2006. CLUC: a natural clustering algorithm for categorical datasets based on cohesion. ***https://dl.acm.org/doi/10.1145/ 1141277.1141422***. Erişim Tarihi: 12.07.2018

Nizam, H.,ve Akın, S. S., 2014. Sosyal Medyada Makine Öğrenmesi ile Duygu Analizinde Dengeli ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması***. XIX. Türkiye'de İnternet Konferansı***.

Orczyk, T. ve Porwik, P., 2013. Influence of missing data imputation method on the classification accuracy of the medical data**.** ***https://yadda.icm.edu.pl/baztech/ element.*** Erişim Tarihi: 03.04.2018.

Özdamar, K.,2002. ***Paket programlar ile istatistiksel veri analizi.*** Kaan Yayınları, Eskişehir.

Öztemel, E.,2003. ***Yapay Sinir Ağları***. Papatya yayınevi, İstanbul.

Piantadosi, S., 1997. ***Clinical Trials: A Methodologic Perspective***. John Wiley. ISBN: 978-1-118-95920-**6** August 2017 928 Pages Sons.,New York.

Rahman, M. M. ve Davis, D. N., 2013. Machine Learning-Based Missing Value Imputation Method for Clinical Datasets**.** ***IAENG Transactions on Engineering Technologies***, In: Yang, G.-C.

Rubin, D. B., 1976. Inference ve Missing Data. ***Biometrika,*** **63**(3): 581.

Rousseeuw P.,1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J**.** ***Comput. Appl.Math***., **20**(1):53–65.

Sarle, W. S.,1994. Neural Networks and Statistical Models. ***Proceedings of the Nineteeth Annual SAS Users Group International Conference***, Cary, NC: SAS Institute.

Siddiqui O. ve Ali M.W., 1998. A comparison of the random-effects pattern mixture model with last-observation-carried-forward (LOCF) analysis in longitudinal clinical trials with dropouts. ***Journal of Biopharmaceutical statistics***, **8**: 545-563.

Schafer, J. L., 1997. ***Analysis of incomplete multivariate data***. Chapman & Hall, USA.

Sijtsma, K. ve Van der Ark, L.,2003. Investigation and treatment of missing item scores in test and question naire data. ***Multivariate Behavioral Research***, **38**(4): 505-528.

Stasny E. A., 1986. Estimating gross flows using panel data with nonresponce: an example from the Canadian Labour Force Survey. ***Journal of the American Statistical Association***, **81**: 42-47.

Simonoff, J. S., 1988. Regression diagnostics to detect non-random missingness in linear regression. ***Technometrics, 30***: 205–214.

Sinharay S, Stern HS, Russell D., 2001. The use of multiple imputation for the analysis of missing data. ***Psychol Methods,* 20:** 35-40.

Sharma S., 1996. ***Applied multivariate techniques***. New York, NY, USA: John Wiley&Sons, Inc.

Somasundaram, R. S. ve Nedunchezhian, R., 2011. Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values. ***International Journal of Computer Applications***, **21**(10): 14-20.

Steyerberg, E. W., (2008).***Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating***. Springer Science & Business Media.

Şeker S.E. ve Erdoğan D.,2017. KNİME ile Uçtan Uca Veri Bilimi.

sadievrenseker.com › wp-content › uploads › veribilimi_knime. Erişim Tarihi: 12.06.2018

Tasdemir, K., Milenov, P. ve Tapsall, B., 2011. Topology-based hierarchical clustering of self-organizing maps. ***IEEE Transactions on Neural Networks,* 22**(3): 474-485.

Theodoridis, S. ve Koutroumbas, K.,2006. ***Pattern Recognition.*** 3rd Ed.,London, Academic Press.

Topuz, D. ve Çakır, M.,2003. Lojistik Regresyon Analiz Tekniğinin Eğitim Bilimleri Araştırmalarında Uygulanabilirliği İle İlgili Bir Araştırma. ***Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi,* 3** (2).

Ultsch A. ve Siemon H.P.,1990. Kohonen's self organizing feature maps for exploratory data analysis***. InProc. INNC'90, Int. Neural Network Conf***., 305-308, Dordrecht, Netherlands. Kluwer.

Xie L., Beni G., 1991. A Validity Measure For Fuzzy Clustering. ***IEEE Trans. On Pattern Analysis And Machine Int***., **13**(4): 841-846.

Xia, F. , 2013. K nearest neighbor. ***http://courses.washington.edu/ling572/winter2013/slides/class4_kNN.pdf.*** Erişim Tarihi: 26.04.2018.

Vatansever M.,2008. ***Görsel Veri Madenciliği Tekniklerinin Kümeleme Analizlerinde Kullanımı ve Uygulaması*** (Yüksek Lisans Tezi)Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Van Hulle M. M.,2012. ***Self-organizing maps***. Springer, Berlin.

Wang, W., Yang, J. ve Muntz, R., 1997. STING: A statistical information grid approach to spatial data mining. ***Int. Conf. on Very Large Data Bases (VLDB97)***, 186–195. Athens, Greece.

White, H., 1989. Learning in Artificial Neural Networks: A Statistical Perspective. ***Neural Computation***, **1**(4): 425-464.

Wisniewski S.R., Leon A.C., Otto M.W. ve Trivedi M.H., 2006. ***Prevention of missing data in clinical research studies***. Biol Psychiatry.

Wong, M. L. D., Jack, L. B. ve Nandi, A. K., 2006. Modified self-organising map for automated novelty detection applied to vibration signal monitoring***. Mechanical Systems and Signal Processing***, **20**(3): 593-610.

Wothke, W., 1998. ***Longitudinal ve multi-group modelling with missing data***. In T. D. Little, K. U. Schnabel ve J. Baumert (eds). Modeling Longitudinal ve Multiple Group Data: Practical Issues, Applied Approaches ve Specific Examples. Mahwah, NJ: Lawrence Erlbaum Associates.

Witten, I. H., Frank, E. ve Hall, M. A., 2011**.** ***Data Mining: Practical Machine Learning Tools and Techniques***. Morgan Kaufmann, ISBN: 0-12-374856-9.

Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao ve Junjie Wu., 2010. Understanding of Internal Cluster Validation Measures. ***IEEE International Conference on Data Mining.***

Yen, G. G. ve Wu, Z., 2008. Ranked centroid projection: A data visualization approach with self-organizing maps. I***EEE Transactions on Neural Networks,*** **19**(2): 245-259.

Yurtoğlu, H.,2005. ***Yapay Sinir Ağları Metodolojisi ile Öngörü Modellemesi: Bazı Mikro ekonomik Değişkenler için Türkiye Örneği.*** Ekonomik Modeller ve Stratejik Araştırmalar Genel Müdürlüğü Uzmanlık Tezi. *Ankara.*

Zhang, X., 2014. Machine Learning: Model Selection***. https://www.lri.fr/~xlzhang/KAUST/CS229_slides/c14_model_selection.pdf.*** Erişim Tarihi: 02.12.2018.

Zhu ve Shi, 2018. A Novel Support Vector Machine Algorithm for Missing Data. ***https://dl.acm.org/doi/10.1145/3194206.3194214.*** Erişim Tarihi: 05.07.2019

Zontul, M., Kaynar, O. ve Bircan, H., 2004. SOM Tipimde Yapay Sinir Ağlarını Kullanarak Türkiye' nin İthalat Yaptığı Ülkelerin Kümelenmesi Üzerine Bir Çalışma. ***Cumhuriyet Üniversitesi, İktisadi ve İdari Bilimler Dergisi***, 5(2):47-68.

ÖZ GEÇMİŞ

1975 yılında Bayburt'ta doğmuştur. İlk, orta ve lise eğitimini Van'da tamamlamıştır. 1992 yılında Van Yüzüncü Yıl Üniversitesi Fen- Edebiyat Fakültesi Matematik Bölümü'nü kazanarak eğitim hayatına devam etmiştir. Van Yüzüncü Yıl Üniversitesi Fen Bilimleri Enstitüsü Biyometri ve Genetik Anabilim Dalı'nda 2007 yılında yüksek lisansını tamamlamıştır. 2014 yılında doktora eğitimine başlamıştır. Van Yüzüncü Yıl Üniversitesi Gevaş Meslek Yüksek Okulu Bilgisayar Programlama Bölümü öğretim görevliliğine devam etmekte olup evli ve 2 çocuk annesidir.

T.C.
VAN YÜZÜNCÜ YIL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
LİSANSÜSTÜ TEZ ORİJİNALLİK RAPORU

Tarih: 27/01/2020

Tez Başlığı / Konusu:

Kayıp Veri İçeren Veri Setlerinde Kümeleme Uygulamaları

Yukarıda başlığı/konusu belirlenen tez çalışmamın Kapak sayfası, Giriş, Ana bölümler ve Sonuç bölümlerinden oluşan toplam 77 sayfalık kısmına ilişkin, 27/01/2020 tarihinde şahsım/tez danışmanım tarafından Turnitin intihal tespit programından aşağıda belirtilen filtreleme uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı % 4 (yüzde dört) tür.

Uygulanan filtreler aşağıda verilmiştir:
- Kabul ve onay sayfası hariç,
- Teşekkür hariç,
- İçindekiler hariç,
- Simge ve kısaltmalar hariç,
- Gereç ve yöntemler hariç,
- Kaynakça hariç,
- Alıntılar hariç,
- Tezden çıkan yayınlar hariç,
- 7 kelimeden daha az örtüşme içeren metin kısımları hariç (Limit inatch size to 7 words)

Van Yüzüncü Yıl Üniversitesi Lisansüstü Tez Orijinallik Raporu Alınması ve Kullanılmasına İlişkin Yönergeyi inceledim ve bu yönergede belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini bilgilerinize arz ederim.

Serpil SEVİMLİ DENİZ
27/01/2020

Adı Soyadı	: Serpil SEVİMLİ DENİZ
Öğrenci No	: 159102074
Anabilim Dalı	: İstatistik
Programı	: Doktora
Statüsü	: Y. Lisans ☐ Doktora ■

DANIŞMAN ONAYI
UYGUNDUR

Prof. Dr. H. Eray ÇELİK
(Unvan, Ad Soyad, İmza)

ENSTİTÜ ONAYI
UYGUNDUR

(Unvan, Ad Soyad, İmza)