**MARMARA ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**

# OECD SAĞLIK VERİLERİNİN VERİ MADENCİLİĞİ YÖNTEMLERİ İLE ANALİZİ

KÜBRA ÇOŞAR

**YÜKSEK LİSANS TEZİ**
Bilgisayar Mühendisliği
Anabilim Dalı
Bilgisayar Mühendisliği Programı

**DANIŞMAN**
Dr. Öğr. Üyesi Buket DOĞAN

**EŞ-DANIŞMAN**
Prof. Dr. Ali BULDU

İSTANBUL, 2020

**MARMARA ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**

# OECD SAĞLIK VERİLERİNİN VERİ MADENCİLİĞİ YÖNTEMLERİ İLE ANALİZİ

KÜBRA ÇOŞAR

(523617001)

**YÜKSEK LİSANS TEZİ**
Bilgisayar Mühendisliği
Anabilim Dalı
Bilgisayar Mühendisliği Programı

**DANIŞMAN**
Dr. Öğr. Üyesi Buket DOĞAN

**EŞ-DANIŞMAN**
Prof. Dr. Ali BULDU

İSTANBUL, 2020

# TEŞEKKÜR

Veri madenciliği alanında yüksek lisans eğitimimi tamamlamama imkân sağlayan Marmara Üniversitesi'ne, bu tezi hazırlamam için benden bilgisini ve anlayışını hiçbir zaman esirgemeyen bilgi ve deneyimleri ile yol gösteren danışman hocalarım Dr. Buket DOĞAN'a ve Prof. Dr. Ali BULDU'ya, öğrenim hayatım süresince her türlü konuda yardımlarını esirgemeyen arkadaşlarıma, yüksek lisans eğitimim sırasında zaman zaman ihmal etmek zorunda kalmama rağmen maddi ve manevi olarak beni sürekli destekleyen ve her zaman yanımda olan, motive edip bana güç veren kardeşim Hüdanur ÇOŞAR'a ve değerli aileme yürekten teşekkür sunarım.

**Haziran, 2020** **Kübra ÇOŞAR**

## İÇİNDEKİLER

# ÖZET


## OECD SAĞLIK VERİLERİNİN VERİ MADENCİLİĞİ YÖNTEMLERİ İLE ANALİZİ

Günümüzde sağlık hizmetleri, ekonomik ve sosyal kalkınmanın en temel belirleyicilerinden biri olarak kabul edilmektedir. Sağlık hizmetlerine ait sağlık verileri ülkelerin sağlık alanındaki gelişiminin tespit edilmesinde yol gösterici olmaktadır. Birleşmiş Milletler tarafından geliştirilen bir kavram olan İnsani Gelişim Endeksi, sağlık, eğitim ve gelir olmak üzere üç temel alan üzerinden toplumların insani gelişim düzeylerini belirlemek için kullanılmaktadır. İnsani Gelişim Endeksi değerinin sağlık boyutu üzerinde hangi faktörlerin etkili olduğunun belirlenmesi için OECD ülkeleri sağlık verilerinden yararlanılabilir. Literatürden yararlanılarak İnsani Gelişim Endeksi'nin sağlık boyutu üzerinde etkili olduğu belirlenen dokuz faktör; medikal teknoloji, hastane yatak sayısı, hastane sayısı, doğumda yaşam beklentisi, bebek ölüm oranı, anne ölüm oranı, kişi başına GSYH, sağlık harcaması ve kişi başına düşen ilaç satışı olarak belirlenmiştir. Bu tez çalışmasında Ekonomik Kalkınma ve İş Birliği Örgütü ülkelerindeki sağlık hizmeti göstergelerinin İnsani Gelişim Endeksi üzerindeki etkisini belirlemek için veri madenciliği metotları kullanılarak çıkarımlar yapmak amaçlanmıştır. 2011-2016 yıllarındaki bu dokuz faktöre ait değerler Ekonomik Kalkınma ve İş Birliği Örgütü sitesinde yayınlanan verilerden derlenerek bir veri seti hazırlanmıştır. Eğri uydurma ve ortalama değer atama yöntemleri ile eksik veriler tamamlanmıştır. Bilgi kazancı ve korelasyon tabanlı özellik seçimi yöntemleri kullanılarak endeks üzerinde en etkili göstergenin bebek ölüm oranı olduğu belirlenmiştir. Anne ölüm oranı ve kişi başına düşen ilaç harcaması da insani gelişime güçlü etkisi olan özellikler olarak tespit edildi. Hiyerarşik kümeleme, k-means ve k-medoids yardımıyla veri setindeki veriler kümelendi ve sonuçlar karşılaştırıldı. Naive Bayes, Yapay Sinir Ağları ve Lojistik Regresyon yöntemleri kullanılarak sınıflandırma işlemi gerçekleştirildi. Sağlık verilerini sınıflandırmada en başarılı yöntem yapay sinir ağları yöntemi oldu. Son altı yılın İnsani Gelişim Endeksi değerleri incelendiğinde doğumda yaşam beklentisinde en fazla gelişim gösteren ülkenin Türkiye, İnsani Gelişim Endeksi'ni en fazla arttıran ülkelerin ise Türkiye ve İrlanda olduğu belirlendi.



**Anahtar Kelimeler:** İnsani Gelişim Endeksi, Özellik Seçimi, Kümeleme, Sınıflandırma

# ABSTRACT

## ANALYSIS OF OECD HEALTH DATA WITH DATA MINING METHODS

Nowadays, health services are considered as one of the main determinants of economic and social development. Data of health services are guiding for the development of countries in the field of health. The Human Development Index, a concept developed by the United Nations, is used to determine the human development levels of societies in three main areas: health, education and income. Health data of OECD countries can be used to determine which factors affect the health dimension of Human Development Index. Nine factors determined to be effective on the Human Development Index based on the literature; Gross Domestic Product per capita, infant mortality, maternal mortality, hospital beds density, health expenditures, hospitals number, life expectancy, medical technology and pharmaceutical sales. It is aimed to make inferences using data mining methods to determine the impact of healthcare indicators in Organization for Economic Cooperation and Development countries on Human Development Index. A data set was prepared by compiling the values of these nine factors between 2011 and 2016 years from the data that published on the Organization for Economic Cooperation and Development website. The missing data was completed by curve fitting and average value assignment methods. The infant mortality rate was determined as the most effective indicator on the index by using info gain and correlation-based feature selection methods. Maternal mortality rate and pharmaceutical expenditure were also identified as features that had a strong impact on human development. With the help of hierarchical clustering, k-means and k-medoids, the data in the dataset was clustered and the results were compared. Classification was carried out using Naive Bayes, Artificial Neural Networks and Logistic Regression methods. The most successful method in classifying OECD health data was artificial neural networks. When the life expectancy at birth values of the past six years is examined Turkey has shown great development in this area.



**Keywords:** Human Development Index, Feature Extraction, Clustering, Classification

## SEMBOLLER

$\boldsymbol{\mu}$ **:** Ortalama

$\boldsymbol{\sigma}$ **:** Standart Sapma

$\mathbf{p}_{\boldsymbol{i}}$ **:** Olasılık Değerleri

**z** **:** Z-skor

**cov** **:** Kovaryans

**r** **:** Korelasyon Katsayısı

**n** **:** Örnek sayısı

$\boldsymbol{d_{ij}}$ **:** i. ve j. noktanın aralarındaki uzaklık

$\boldsymbol{x_{ij}}$ **:** i. noktanın j. değişkeninin değerini

$\forall$ **:** Hepsi sembolü

$\in$ **:** Kümenin elemanı sembolü

$\sum$ **:** Toplam Sembolü

**C** **:** Küme

**S(C)** **:** Ortalama Silhouette Genişliği

$\boldsymbol{M_k}$ **:** Küme merkezi

$\mathbf{e_k}$ **:** Karesel hata değeri

$\boldsymbol{E_k}$ **:** Toplam karesel hata değeri

**log** **:** Logaritmik

$\boldsymbol{w_i}$ **:** Ağırlık

$\boldsymbol{\hat{y}_i}$ **:** Tahmini çıkış değeri

$\boldsymbol{\beta_x}$ **:** Değişken katsayısı

$\boldsymbol{\pi}$ **:** pi sayısı

## KISALTMALAR

**ABD** : Amerika Birleşik Devletleri

**ANFIS** : Adaptive-Network Based Fuzzy Inference Systems

**AÖO** : Anne Ölüm Oranı

**BÖO** : Bebek Ölüm Oranı

**CLARA** : Clustering Large Applications

**CLARANS** : Clustering Large Applications based upon RANdomized Search

**CO2** : Karbondioksit

**CSV** : Comma Seperated Values

**ELECTRE TRI** : ELimination Et Choix Traduisant la REalite-Elimination

**EM** : Expectation-Maximization

**GSYH** : Gayri safi milli gelir

**HYS** : Hastane Yatak Sayısı

**IGE** : İnsani Gelişim Endeksi

**KDD** : Knowledge Discovery in Databases

**MLP** : Multilayer neural networks

**O3** : Ozon

**OECD** : Organisation for Economic Co-operation and Development

**OLAP** : On Line Transactional Processing

**PCA** : Principal component analysis

**PAM** : Partitioning Around Medoids

**SO2** : Kükürtdioksit

**UNDP** : United Nations Development Programme

**YSA** : Yapay Sinir Ağı

**DP** : Doğru Pozitif

**DN** **:** Doğru Negatif

**YP** **:** Yanlış Pozitif

**YN** **:** Yanlış Negatif

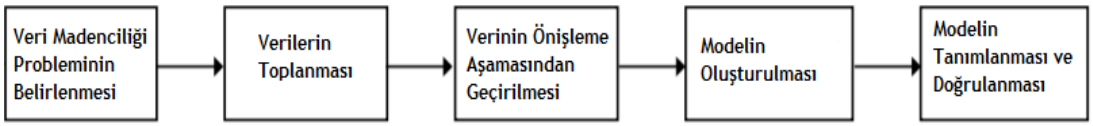
# ŞEKİL LİSTESİ

# TABLO LİSTESİ

# 1. GİRİŞ

Bilgisayarlardaki veri analizi problemini çözmesi için 1960'lı yıllarda ortaya atılmış bir kavram olan veri madenciliği, çeşitli türde yapılandırılmış ve yapılandırılmamış veri içeren büyük veri setlerindeki değerli bilgileri elde etmek ve gizli örüntüleri keşfetmek için istatistiksel teknikler, algoritmalar ve araçlar kullanan disiplinler arası bir alandır [1]. Bu kavram ilk olarak veri tarama (data dredging) ve veri yakalanma (data fishing) olarak isimlendirildi. 1990'lı yıllarda ise veri bilimciler ve mühendisler tarafından önerilen veri madenciliği ismi kullanılmaya başlandı [2]. Bilgisayar biliminin alt alanı olan veri madenciliği istatistik, veri bilimi, veri tabanı teorisi ve makine öğrenimi gibi birçok tekniği harmanlayarak sonuca ulaşır. Son yıllarda bilgisayar teknolojilerinde yaşanan ilerlemeler ile sağlık, eğitim, satış, endüstri, kamu, savunma sanayi vb. çeşitli alanlarda veri depolamaya olan yönelim önemli ölçüde artmıştır. Bunun sonucunda çok miktarda veri tabanı ve devasa büyüklüklerde veri kümeleri ortaya çıkmıştır. Veri tabanları ve bilgi teknolojisi alanındaki araştırmalar, bu değerli verilerin karar verme(decision making) işlemi için saklanması ve manipüle edilmesine yönelik bir yaklaşıma yol açmıştır. Veri madenciliği yöntemleri, birçok farklı boyut veya açıdan verilerin analiz edilmesine, kategorize edilmesine ve tanımlanan ilişkilerin özetlenmesine olanak tanır. Teknik olarak, veri madenciliği, veri tabanlarında yer alan çok sayıda alan arasında korelasyon veya örüntü bulma işlemidir [2]. Veri madenciliği tekniklerinin en yaygın olarak kullanılanları özellik çıkarımı, kümeleme, sınıflandırma, regresyon ve birliktelik analizidir. Veri önişleme ve görselleştirme veri analizi işlemlerinin temel basamaklarıdır. Veri madenciliği yöntemleri ile çıkarım yapmadan önce veri önişleme aşamasından geçer. Analiz işlemi sonrasında ise elde edilen gözlemler veri görselleştirmesi ile daha anlamlı ve okunabilir hale getirilir [3].

KDD(Knowledge Discovery in Databases) prosedürlerine göre veri madenciliği sürecinde odaklanılması gereken beş adım Şekil 1.1'de gösterilmiştir. Veri madenciliği probleminin tanımlanması, verilerinin toplanması, verilerin algılanması ve düzeltilmesi(veri önişleme), modelin oluşturulması ve model açıklaması ve doğrulama veri madenciliği sürecinde odaklanılması gereken işlemlerdir [4].




**Şekil 1.1** Veri madenciliği süreci

Veri toplama aşaması farklı kaynaklardan ve konumlardan verilerin elde edilmesi işlemidir. Veri toplama dâhili ve harici veriler üzerinden yapılır. Dâhili veriler; mevcut veri tabanları, veri ambarları ve OLAP. Harici veriler; web grafiklerinden toplanan verilerdir [5].

Sağlık hizmetlerinde farklı kaynaklardan elde edilen verilerin analiz edilerek hasta ve sağlık sistemini hizmetlerini optimize etmek, sağlık sistemlerinin ihtiyaçlarını belirlemek ve sağlıkta kaliteyi arttırmak için büyük öneme sahiptir [6]. Sağlık verilerinin analizi işleminde veri madenciliği yöntemleri ile çıkarımda bulunulabilir.

## 1.1. İnsani Gelişim Endeksi Bileşenleri

Günümüzde sağlık hizmetleri, ekonomik ve sosyal kalkınmanın en temel belirleyicilerinden biri olarak kabul edilmektedir. Sağlık hizmetleri, yapısı gereği hizmeti alan bireyden ve çevresinden başlayarak tüm topluma yayılan bir etkiye sahiptir. İnsani Gelişim Endeksi (IGE), ülkelerin insani gelişim düzeylerinin ölçülmesi için kullanılan Birleşmiş Miletler tarafından ortaya çıkarılmış bir kavramdır. IGE sağlık, eğitim ve gelir olmak üzere üç temel alan üzerinden toplumların insani gelişim düzeylerini belirlemektedir. 2010 yılında Birleşmiş Milletler tarafında IGE'yi hesaplamak için kullanılan göstergelerde ve hesaplama yönteminde değişiklik yapılmıştır. Daha önceleri de kullanılan üç boyut (gelir, eğitim ve sağlık) aynı kalmıştır ancak gelir ve eğitimi ölçmek için kullanılan ölçütler değiştirilmiştir. Geliri ölçmek için kullanılan kişi başına GSYH(Gayri safi milli gelir) yerine satın alma gücü paritesine dönüştürülmüş kişi başı gayrisafi milli gelir, eğitimi için ise ortalama okullaşma yılı ve beklenen okullaşma yılı kullanılmaya başlanmıştır [7].

Tablo 1.1'de Birleşmiş Milletler Kalkınma Programı (UNDP)'nın son olarak 2016 yılında endeks değerleri için belirlemiş olduğu maksimum ve minimum değerler sunulmuştur [7].

**Tablo 1.1** İnsani Gelişim Endeksi Minimum ve Maksimum Değerleri

| IGE Bileşeni | Göstergeler | Minimum-Maksimum Değerler |
|---|---|---|
| Eğitim | Beklenen Okullaşma Yılı | 0-18 Yıl |
| | Ortalama Okullaşma Yılı | 0-15 Yıl |
| Sağlık | Doğumda Yaşam Beklentisi | 20-85 Yıl |
| Gelir | Satın Alma Gücü Paritesi Kişi Başı GSMG | 100-75 000 $ |

Beklenen okullaşma yılı ve ortalama okullaşma yılı endeksinin minimum değerinin 0 olmasının nedeni toplumların eğitim olmadan da var olabilmesidir. Doğumda yaşam beklentisinin minimum değerinin 20 yaş olarak belirlenmesinin nedeni tarihsel kayıtlara dayanmaktadır. Savaşlar, soykırımlar, kıtlık vb. nedenler doğumda yaşam beklentisini oldukça aşağı çeker [8, 9]. 20 yaşın altındaki değerlerde de doğumda yaşam beklentisi görülmüştür ancak geçici durumlar olduğundan UNDP tarafından minimum değer 20 yıl olarak belirlenmiştir.

İnsani Gelişim Endeksini hesaplamak için öncelikle sağlık, eğitim ve gelir göstergelerinin ayrı ayrı hesaplanması gerekir. Her bir göstergenin değeri Denklem 1.1'de belirtilen Min-Max Normalizasyon yöntemi ile 0-1 aralığında bir değere dönüştürülür. Denklem 1.1'de yer alan minimum değer ve maksimum değer Tablo 1.1'de belirtilmiştir.

$$D = \frac{\text{Gerçek Değer - Minimum Değer}}{\text{Maksimum Değer - Minimum Değer}} \tag{1.1}$$

Sırasıyla sağlık, eğitim ve gelir göstergelerine ait $D_{sağlık}$ , $D_{eğitim}$ ve $D_{gelir}$ değerleri Denklem 1.1'e göre hesaplanır. Hesaplanan tüm göstergeler 0 ile 1 aralığında bir değer alabilir. Denklem 1.2'de her üç endeksin geometrik ortalaması alınmakta ve çıkan sonuç İnsani Gelişim Endeksi (İGE) değerini vermektedir.

$$\text{İGE} = \sqrt[3]{\text{Dsağlık} * \text{Değitim} * \text{Dgelir}} \tag{1.2}$$

İGE değerine göre ülkeler dört sınıfa ayrılmaktadır [10] :

- 0,800 puan ve üstü = Çok yüksek insani gelişim
- 0,700- 0,799 puan = Yüksek insani gelişim
- 0,550- 0,699 puan = Orta insani gelişim
- 0,550 puan ve altı = Düşük insani gelişim

Bu çalışmada OECD(Organisation for Economic Co-operation and Development) ülkelerindeki sağlık hizmeti harcamaları, kaynakları ve kalite göstergelerinin İnsani Gelişim Endeksi(İGE) üzerindeki etkisini belirlemek için veri madenciliği metotları kullanılarak çıkarımlar yapmak hedeflenmektedir. IGE değerinin üzerinde etkili olan belirleyiciler tespit edilerek ülkelerin sağlık hizmeti planlamalarında bu alanlara yönelmesine yol gösterilmesi amaçlanmaktadır.

## 1.2. İlgili Çalışmalar

### 1.2.1. IGE sağlık endeksine etki eden belirleyiciler

Ülkelerin genel sağlık düzeyi iyileştirildiğinde ve sağlık yatırımları arttığında insani gelişim açısından da ilerleme kaydedildiği belirlenmiştir [11]. İGE hesaplanırken sağlık bileşeni olarak doğumda yaşam beklentisi indeksi kullanılmaktadır. Doğumda yaşam beklentisi, bireyin doğduktan sonra hayatta kaldığı ortalama süre olarak ifade edilir [12]. Toplumun genel sağlık düzeyi ortalama yaşam süresi üzerinde etkiye sahiptir. Sağlık alanında yaşanan gelişmeler, başta gelişmekte olan ülkeler olmak üzere tüm ülkelerde doğumda yaşam beklentisinin artmasına neden olmuştur [11, 12]. Literatürde doğumda yaşam beklentisi üzerinde etkili faktörlerle ilgili yapılan çalışmalardan bazılarına aşağıda yer verilmiştir:

Potter [13] hükümetlerin başlıca hedeflerinden birinin, ölüm oranını en düşük seviyeye indirerek yaşam beklentisini arttırmak olduğunu belirtmiştir. Yüksek yaşam standartlarına sahip ülkelerin vatandaşları ortalama olarak daha uzun yaşama ve daha düşük ölüm oranına sahiptir.

Bilas ve arkadaşları [12] panel veri analizi yaklaşımını kullanarak, 2001-2011 yılları arasında yıllık bazda GSYH büyüme oranı, nüfus artış hızı, eğitim düzeyi, kişi başına düşen GSYH ve yaşam beklentisi değişkenlerini dikkate alarak Avrupa Birliği ülkelerindeki doğumda yaşam beklentisinin belirleyicilerini araştırmıştır. Kişi başına düşen GSYH ve eğitim düzeyindeki değişim sonucunda yaşam beklentisinde değişimin ortaya çıktığı tespit edilmiştir.

Hassan ve ark. [14] 108 gelişmekte olan ülkede, 2006-2010 yılları arasında yaşam beklentisi ve sağlık harcamaları, gayri safi yurt içi hasıla, eğitim endeksi, sağlık tesislerinde iyileştirme ve su kaynaklarında iyileştirme arasındaki ilişkiyi panel veri

analizi kullanılarak incelenmiştir. Analiz sonucunda, yaşam beklentisi ve ele alınan tüm açıklayıcı değişkenler arasında pozitif yönde bir ilişki olduğu görülmüştür.

Doğumda yaşam beklentisinin sosyal ve ekonomik gelişim ile ilişkili olduğunu öne süren Delavari ve arkadaşları [15] çalışmalarında kişi başına düşen GSYH, 10.000 kişi başına düşen doktor sayısı, gıda bulunabilirliği, okur-yazarlık oranı ve toplam doğurganlık belirleyicilerinin İran'da yaşam beklentisini etkileyen temel faktörler olarak belirledi.

Bu alanda yapılmış bir diğer çalışma Lin ve ark. [16]'nın yaptığı az gelişmiş ülkelerde 1970 - 2004 yılları arasında doğumda yaşam beklentisi üzerinde etkili olan sosyal ve politik belirleyicileri tespit ettiği çalışmadır. Demokratik siyasi yönetimin kısa vadede yaşam beklentisini arttırmaya yönelik küçük bir etkisi olmasına rağmen uzun vadede etkili bir belirleyici olduğu tespit edilmiştir.

Balan ve Jaba [17], 1970-2008 dönemleri arasında Romanya'da yaşam beklentisini belirleyen faktörler analiz edilmiştir. Sonuç olarak, yaşam beklentisi ve ücretler, hastanedeki yatak sayısı, doktor sayısı, kütüphanelere abone olan okuyucu sayısı arasında pozitif bir ilişki mevcutken; okuryazar olmayan nüfus ve nüfus artışı arasında negatif yönde ilişki olduğu görülmüştür.

Bayın [18] OECD'ye üye olan 34 ülkenin 2013 yılı doğumda yaşam süresi ve 65 yaşta yaşam sürelerine etki eden yaşam sürelerine etki ettiği düşünülen algılanan sağlık statüsü, hastane yatak sayısı, kişi başı milli gelir, kişi başı sağlık harcaması, kişi başı ilaç tüketim harcaması, anne ölüm hızı, bebek ölüm hızı, doktor ziyareti, hastane yatış gün sayısı ve kentsel nüfus oranı değişkenlerinin kadın ve erkeklerin beklenen yaşam sürelerine olan etkisi olanlar regresyon analizleri ile incelenmiştir. Araştırma sonucunda, hem kadınlarda hem de erkeklerde doğuşta beklenen yaşam süresine en çok etki eden değişkenin bebek ölüm hızı olduğu sonucuna ulaşılmıştır.

Doğu Akdeniz Bölgesi'nde doğumda yaşam beklentisi üzerindeki belirleyicileri araştıran Gillian ve Skrepnek [19] kentsel nüfus yüzdesi, güvenli içme suyuna erişim yüzdesi, hekim yoğunluğu (1000 kişi başına düşen yoğunluk), yetersiz beslenenlerin yüzdesi ve difteri-tetanos, boğmaca, menenjit ve çocuk felcine karşı ortalama aşılama oranlarını belirleyiciler olarak kullanmışlardır. Çalışmada GSYH, aşılama ortalamaları ve kentleşmenin yaşam beklentisinin anlamlı pozitif belirleyicileri olduğu tespit edilmiştir.

Amerika ve İngiltere'de uzun yılları boyunca yapılan çalışmalara göre gelir ve sağlık arasında karmaşık bir ilişki vardır. 1950'lerden beri ölümlerde önemli bir azalış olmuştur. Ancak bu azalış, Folland ve arkadaşlarının [20] yaptığı incelemelere göre artan gelirlerden ziyade sağlık alanındaki teknolojik ilerlemelerden dolayı gerçekleşmiştir.

Mathers ve arkadaşları [21], son yıllarda gelişmiş ülkelerde, beklenen yaşam süresindeki artışı vurgulamışlar ve yaşam beklentisinin kronik hastalıklara bağlı ölümlerden etkilendiğini belirtmişlerdir.

İnsani gelişim endeksi üzerine yapılan çalışmaların çoğunda genel kamu harcamaları ve sağlık harcamaları ele alınmıştır. Doğrudan sağlık hizmetleri bileşenlerinin kullanıldığı çalışmalar oldukça sınırlıdır.

Yapılan literatür incelemesi sonucunda IGE hesaplamasında sağlık endeksinin hesaplanmasında kullanılan doğumda yaşam beklentisi üzerinde etkili olan değişkenler Tablo 1.2'de gösterilmiştir.

**Tablo 1.2** Literatürde yer alan doğumda yaşam beklentisi üzerinde etkili belirleyiciler

| BELİRLEYİCİ | VERİ KAYNAĞI | REFERANS |
|---|---|---|
| Doğumda Yaşam Beklentisi | OECD | [12] [14] [15] [16] [17] [18] |
| Kişi Başına Düşen GSYH | OECD | [12] [15] |
| Sağlık Harcaması (GSYH'nin % kaçı sağlığa ayrılmış) | OECD | [12] [14] [18] |
| Bebek Ölüm Oranı | OECD | [18] |
| 10 000 kişi başına düşen doktor sayısı | OECD | [15] [17] |
| 10 000 kişi başına düşen hastane sayısı | OECD | [17] [18] |
| Hastane yatak sayısı | OECD | [17] [18] |
| Kişi başı ilaç tüketimi harcaması | OECD | [18] |
| Anne ölüm oranı | OECD | [18] |
| Aşılamalar (immunisation) | OECD | [19] |
| Medikal Teknoloji (Bilgisayarlı tomografi cihazı, manyetik rezonans görüntüleme üniteleri…) | OECD | [20] |
| Hastalıklılık(Morbidity) | OECD | [21] |

Toplamda 12 etkili özellik tespit edildi. Bu değişkenlerden 10 bin kişi başına düşen doktor sayısı, aşılamalar ve hastalıklılık için 2011-2016 yılları arasındaki OECD ülkelerine ait verilerde %27'den fazla eksik olduğu için çalışma kapsamı dışında bırakıldı. Doğumda yaşam beklentisi, kişi başına düşen GSYH, sağlık harcaması, anne ve bebek ölüm oranı, hastane sayısı, hastane yatak sayısı, kişi başına ilaç tüketim harcaması ve medikal cihaz sayısı çalışmada kullanılan OECD sağlık verileridir. İGE değerleri Birleşmiş Milletler Gelişim Programı kapsamında sunulduğu için çalışma kapsamındaki OECD ülkelerinin 2011-2016 yıllarındaki İGE değerleri Birleşmiş Milletler'in yayınladığı değerlerden alındı.

### 1.2.2. Kümeleme analizi ile ilgili çalışmalar

Ülkelerin sağlık politikalarının belirlenmesinde sağlık göstergeleri önemli bir role sahiptir. Sağlık göstergeleri sadece sağlık alanında değil aynı zamanda ülkelerin gelişmişlik düzeylerinin belirlenmesinde de büyük paya sahiptir [22]. Her yapılan sağlık harcaması da göstergeler üzerinde olumlu etkiye sahip değildir. Önemli olan doğru alanda doğru yatırım ve işin gerçekleştirilmesidir [23].

Günümüzde dünya çapında yaşanan Covid-19 salgını ile ülkelerin sağlık alanında yaşadığı gelişimler bir sınavdan geçmiş oldu. Sağlık alanındaki atılımlar, yenilikler ve sağlık ekibi organizasyonu bu salgında başrol oynadı. Koordineli bir küresel müdahaleye ihtiyaç duyuldu [24]. Covid-19 ile beraber birçok salgın ve hastalık toplumu tehdit etmektedir. Ülkelerin halklarının sağlığı için yeni adımlar atması, eksik kaldığı alanlarda ilerlemesi ve mevcut sağlık sisteminin geliştirilmesi büyük bir zorunluluk haline gelmektedir.

OECD ülkeleri sağlık alanında OECD dışındaki ülkelere göre büyük bir alt yapıya sahiptir. OECD ülkeleri sağlık alanın gelişiminde farklı yöntemlere başvurmaktadır. Sağlık alanında insani gelişim için ülkeler farklı politikalar izlemektedir. Avrupa Birliği üyesi ülkeler Avrupa Birliği Sağlık Politikası çerçevesinden sağlık alanındaki şartları belirlemişlerdir. Birliğe üye olmak isteyen bir ülke bu sağlık politikalarını karşılıyor olması gerekmektedir [25]. OECD ülkelerinden biri olan Türkiye, Avrupa Birliği aday ülkesi konumunda olması sebebiyle Avrupa Birliği sağlık politikalarını karşılamaya çaba sarf etmektedir. İskandinav ülkeleri tarafından uygulanan İskandinav modeli sağlık politikaları halkın sağlık alanında eşit durumda olması ve sağlık seviyesinin

yükseltilmesini hedeflemektedir. Amerikan sağlık modelinde sağlık politikası sağlık harcamasında devlet desteğinin az olduğu bir politikadır [26]. Sağlık alanında birbirine en benzer politikalar izleyen ülkeleri tespit etmek için kümeleme yöntemi kullanılabilir.

Proksch ve arkadaşları [27], OECD ülkelerinin sağlık hizmetlerindeki inovasyonların çıktılarına göre kümelenmesinin incelendiği çalışmasında çok göstergeli bir yaklaşım kullanmışlardır. Karesel Öklid uzaklığı ve Ward's metodunun kullanıldığı hiyerarşik kümeleme ile 34 ülkenin verileri incelenerek kümeleme analizi yapılmıştır.

OECD ülkelerinin ekonomik özgürlüklerine göre kümelendiği çalışmada Gülden ve Karakış [28], Ward metodu, centroid, medyan, en uzak komşu, en yakın komşu, gruplar arası bağlantılı ve grup içi bağlantılı hiyerarşik kümeleme ve k-means kümeleme yöntemini kullanmıştır. Değişkenler kareli Öklid uzaklıklarına göre analiz edilmiş ve yöntemlerin sonuçları arasında korelasyon analizi yapılmıştır.

Kangallı ve arkadaşları [29] OECD ülkelerini 2011 yılındaki ekonomik özgürlük ve gelişmişlik düzeyleri açısından incelemişlerdir. Kümeleme analizi yöntemlerinden olan k-ortalamalar ve Ward yöntemine göre yapılan analiz çalışmasında ülkeler 3 kümeye ayrılmıştır.

Mylevaganam [30], PCA(temel bileşen analizi) ve K-means kümeleme yöntemleri ile R programlama dilini kullanarak IGE'ye ait doğumda yaşam beklentisi, gelir indeksi ve eğitim indeksine ait verileri kategorize ederek, Birleşmiş Milletlere üye ülkeler arasında sıralama yapmışlardır.

Son yıllarda yeni verilerin ortaya çıkmasıyla, bu yeni verinin analizini farklı bakış açılarıyla değerlendirmek insan sağlığına ve hayatına olumlu etki edecektir. Literatürde yol gösterici olması açısından bu çalışmada OECD sağlık verileri üzerinde hiyerarşik kümeleme, k-means ve k-medoids yöntemleri ile kümeleme yapıldı. Bu üç yöntemin uygulanması sonucunda elde edilen kümeleme sonuçları ve performansı karşılaştırıldı.

### 1.2.3. Sınıflandırma ile ilgili çalışmalar

Sınıflandırma pek çok farklı model ve algoritmayı içerisinde barındıran bir veri madenciliği yöntemidir. Bu kadar çeşitli yöntemi içermesi nedeniyle kullanılacak veri için en uygun yöntemin seçilmesi büyük önem taşımaktadır [55]. Algoritma sayısının ve versiyonlarının fazla olması, her bir algoritmanın farklı bir çalışma prensibine sahip

olması, kullanılabilecek veri tipinin algoritmadan algoritmaya değişmesi ve eldeki verilerin önişleme aşamasından geçirilip geçirilmemesi algoritmanın çıktısına doğrudan etki edebilmektedir. Veri için uygun sonucun belirlenmesi için birden fazla algoritmanın sonuçlarının karşılaştırılması ve en etkilisinin seçilmesi gerekebilir. Sınıflandırma yöntemlerine ve OECD sağlık verileri ve insani gelişim üzerinde etkili olan sağlık belirleyicileri üzerinde yapılan sınıflandırma çalışmalarına ait yayınlar incelendi.

Yakut ve arkadaşları [31], bebek ölüm oranı, gayri safi milli hasıla, lise kayıt oranı, büyüme, doğrudan yabancı yatırım, enerji tüketimi, enerji üretimi, enflasyon, ihracat, internet kullanıcı sayısı, işsizlik, ithalat, mobil telefon abone sayısı ve sağlık harcamaları göstergesini kullanarak 81 ülkenin İnsani Gelişim Endeksi değeri üzerinde sıralı lojistik regresyon analizi ve yapay sinir ağlarını kullanarak yaptıkları sınıflandırma karşılaştırmasında çok katmanlı yapay sinir ağları sınıflandırmasının daha başarılı olduğunu belirlemiştir.

RL do Carvalhal Monteiro ve arkadaşları ELECTRE TRI (ELimination Et Choix Traduisant la REalite-Elimination) yöntemini kullanarak İnsani Gelişim Endeksi sonuçlarının sınıflandırması için yeni bir yaklaşım önermiştir [32].

Burmaoğlu ve arkadaşları, 120 ülkeye ait toplam nüfus, kırsal nüfus, kadınların parlamentodaki koltuk yüzdesi, özel ve kamu sağlık harcamaması, doğumda yaşam beklentisi, ilkokula kayıt oranı, sabit telefon sayısı, cep telefonu sayısı, internet kullanıcı sayısı, GSYH, ithalat, ihracat, elektrik tüketimi ve cezaevi nüfusu göstergeleri üzerinde Diskriminant analizi yöntemi ile sınıflandırma yapmıştır. Çıkış değişkeninin iki değerden oluşması nedeniyle tek bir diskriminant fonksiyonunun kullanıldığı çalışmada 4 değişkenli olarak kurulan modelde; kişisel sağlık harcamaları ve 1000 kişi başına düşen cep telefonları olumlu, kamu ve özel sağlık harcamaları olumsuz etkiye sahip olduğu belirlendi [33].

Yi ve arkadaşları [38], sağlık göstergelerinin kirletici emisyonlara($CO_2$, $O_3$, $SO_2$ vb.) matematiksel bağımlılığını belirlemek için klasik bir regresyon analizi yönteminin kullanılmasını önerdikleri çalışmalarında, morbidite göstergeleri ve morbiditeyi belirleyen faktörleri arasındaki ilişkileri araştırdığı gösterilmiştir. Korelasyon bazlı regresyon analizi sırasında morbiditenin farklı korelasyon ve regresyon modellerinin oluşturulduğu da gösterilmiştir. Atıfta bulunulan çalışmanın yazarları, bağımlı değişkenin

tahmini değerlerini belirlemek için bir iletişim formu ve bir model tipi seçimini gösteren bir regresyon analizini tanımlamıştır.

Nüfus sağlığı göstergelerinin dış faktörlerin neden olduğu hastalıklara bağımlılığını belirlemek için yapay sinir ağlarının (YSA) kullanımının önerildiği çalışmada Alam ve arkadaşları [39], YSA'ların doğrusal modellere, genelleştirilmiş doğrusal modellere ve doğrusal olmayan modellere dayanan çeşitli bağımlılıkların simülasyonunu mümkün kıldığını göstermektedir. YSA'nın güvenilir istatistiki tahminlerin elde edilmesinin altında yatan girdi ve çıktı verileri arasındaki gizli bağımlılıkları genelleme ve vurgulama yeteneği olduğu belirtilmiştir.

Permai, Tanty ve Rahayu [40] yaptıkları çalışmada coğrafi ağırlıklandırma tabanlı çoklu lojistik regresyon analizini kullanarak İGE'yi tahmin eden bir lojistik model geliştirdiler.

Sarıtaş ve Yaşar [45], meme kanseri şüphesiyle kliniğe başvuran hastaların verileri üzerinde hastalık tanısı tahmin etme ve Naive Bayes ve yapay sinir ağlarının sınıflandırma performansını karşılaştırdıkları çalışmalarında yapay sinir ağıları %86,95 ve Naive Bayes sınıflandırıcı ise %83,54 doğrulukla sınıflandırma yapmıştır. Her iki algoritmanın da meme kanseri teşhisinde erken tanı için kullanılabileceği belirtilmiştir.

Rençber ve Mete [46] IGE değerini sınıflandırdıkları çalışmalarında YSA ve ANFIS yöntemlerinden yararlanarak bu iki sınıflandırma yönteminin performansını karşılaştırmışlardır. Yapılan sınıflandırma sonucunda ülkelerin İGE değerlerini YSA %87,5 ve ANFIS %91,36 oranında doğru sınıflandırdığı ve ANFIS'in daha başarılı sonuçlar verdiği belirtilmiştir.

Literatürde YSA, Naive Bayes ve lojistik regresyon için birçok sınıflandırma ve tahmin araştırması yapılmıştır. Bu üç yönteme ait incelenen çalışmalardan Naive Bayes'e ait olanlar Tablo 1.3'te verilmiştir.

**Tablo 1.3** Literatürde sınıflandırma ve tahmin üzerine yapılan Naive Bayes çalışmaları

| Naive Bayes ile Sınıflandırma ve Tahmin Çalışmaları |
|---|
| Kharya, Agrawal ve Soni (2014) [46] |
| Yang ve Webb (2002) [47] |
| Ng ve Jordan(2002) [48] |
| Halloran(2009) [49] |

Literatürde YSA kullanılarak yapılan çalışmalar incelendi. Yapılan literatür taramasında incelenen çalışmalar Tablo 1.4'de sunulmuştur.

**Tablo 1.4** Literatürde yer alan sınıflandırma ve tahmin üzerine yapılan YSA çalışmaları

| YSA ile Sınıflandırma ve Tahmin Çalışmaları |
|---|
| Fish ve arkadaşları (1995) [35] |
| Schumacher, Robner and Vach (1996) [36] |
| Zhang, Hu, Patuwa ve Indro (1999) [37] |
| Yakut ve arkadaşları (1999) [31] |
| Ottenbacher ve ark. (2001) [41] |
| Dreiseitl and Machado (2002) [42] |
| Liew ve arkadaşları (2007) [43] |
| Radhimeenakshi (2016) [44] |
| Sarıtaş ve Yaşar (2019) [45] |

Lojistik regresyon üzerine yapılmış çalışmalardan Tablo 1.5'de yer alana çalışmalar incelenmiştir. Burada belirtilen çalışmalar genellikle lojistik regresyonun çalışma prensibi, Naive Bayes ve yapay sinir ağları ile karşılaştırılmasını içermektedir.

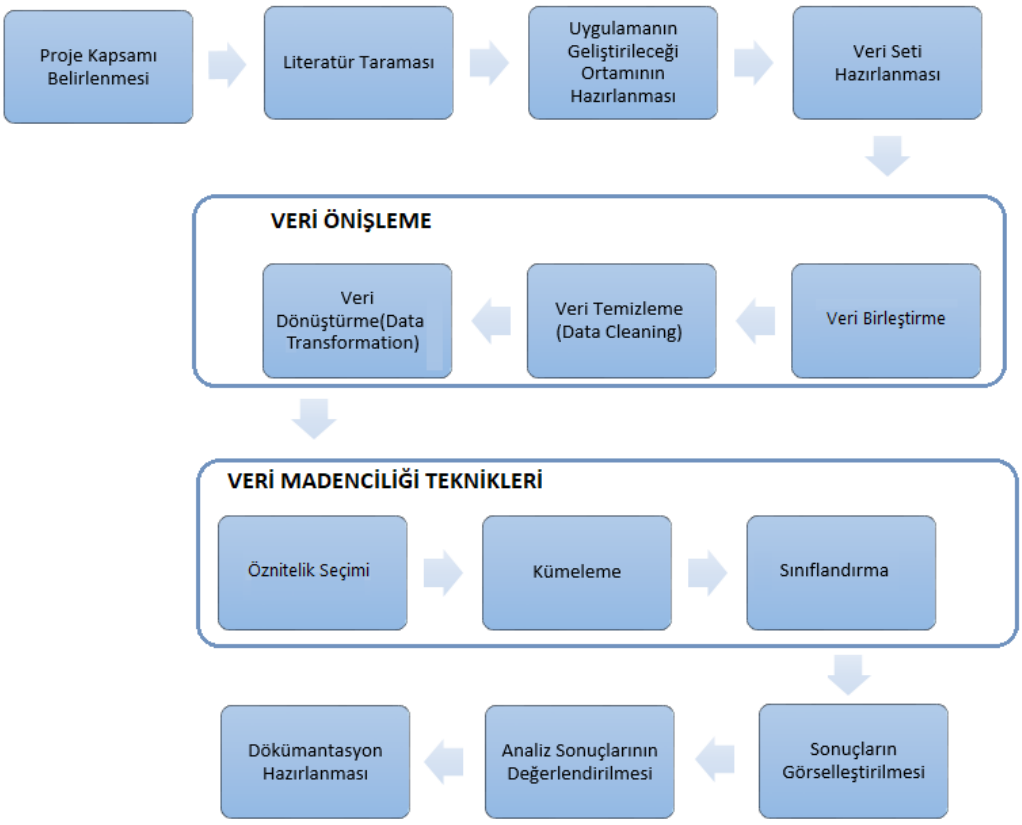
**Tablo 1.5** Literatürde sınıflandırma ve tahmin üzerine yapılan Lojistik Regresyon çalışmaları

| Lojistik Regresyon ile Sınıflandırma ve Tahmin Çalışmaları |
|---|
| Wright (1995) [50] |
| Lin ve arkadaşları (2011) [51] |
| Tu (1996) [52] |
| Xue ve Titterington (2008)[53] |
| Caruana ve Niculescu-Mizil. (2006) [54] |

Literatürdeki çalışmalar incelendiğinde sağlık verileri üzerinde YSA, Naive Bayes ve Lojistik Regresyon sınıflandırma yöntemlerinin birlikte kullanılarak karşılaştırma yapılmasının faydalı olacağı düşüncesi ile yapılacak çalışmada bu üç yöntemin kullanılması kararlaştırılmıştır.

## 1.3. Amaç

OECD ülkeleri sağlık verilerinin kullanılacağı bu çalışmada veriler üzerinde veri önişleme çalışması ve paralel koordinat sistemi ile verilerin yorumlaması işlemi yapıldıktan sonra İnsani Gelişim kavramının temel bileşenlerinden biri olan sağlık göstergesi üzerinde etkili olan özniteliklerin belirlenmesi, belirlenen özniteliklere göre kümeleme ve sınıflandırma çalışmaları yapılması, sonuçların görselleştirilmesi ve analiz edilmesi amaçlanmaktadır. Kullanılan yöntemlerden elde edilen sonuçlar hem kendi aralarında hem de literatürde yer alan sonuçlar ile karşılaştırılacaktır. Veri seti belli bir zaman serisini kapsadığı için bu zaman aralığında veriler üzerinde meydana gelen değişimlerin de çalışmada kapsamında sunulması hedeflenmektedir. Tez çalışması kapsamında yapılması hedeflenen işlemler Şekil 1.2'de verilmiştir.




**Şekil 1.2** Tez kapsamında yapılması amaçlanan işlerin akışı

Uluslararası bir ekonomi örgütü olan OECD Nisan 2020'de Kolombiya'nın da aralarına katılması ile beraber 37 üyeden oluşmaktadır. OECD üye ülkeleri Tablo 1.6'da

gösterilmiştir. Bu 37 üyeden 20 tanesi kurucu üyeler arasında yer almaktadır. Örgüte sonradan 17 üye daha dahil olmuştur.

**Tablo 1.6** OECD üye ülkeleri

| OECD Kurucu Üyeleri | | | | OECD'ye Sonradan Dahil Olan Ülkeler | | | |
|---|---|---|---|---|---|---|---|
| 1 | Almanya | 11 | İspanya | 21 | Avusturalya | 31 | Meksika |
| 2 | ABD | 12 | İsveç | 22 | Çekya | 32 | Polonya |
| 3 | Avusturya | 13 | İsviçre | 23 | Estonya | 33 | Slovakya |
| 4 | Belçika | 14 | İzlanda | 24 | Finlandiya | 34 | Slovenya |
| 5 | Danimarka | 15 | Kanada | 25 | İsrail | 35 | Şili |
| 6 | Fransa | 16 | Lüksemburg | 26 | Güney Kore | 36 | Yeni Zelanda |
| 7 | Hollanda | 17 | Norveç | 27 | Japonya | 37 | Kolombiya* |
| 8 | İngiltere | 18 | Portekiz | 28 | Litvanya | | |
| 9 | İrlanda | 19 | Türkiye | 29 | Letonya | | |
| 10 | İtalya | 20 | Yunanistan | 30 | Macaristan | | |

Nisan 1948'de merkezi Fransa'da bulunan bu örgüt demokrasi, insan hakları, finansal istikrar, özgürlük, yaşam standartlarının iyileştirilmesi, işsizlik, ekonomik ve sosyal gelişme, sağlık hizmetleri, ticareti yardım ve sosyal destekler başta olmak üzere birçok alanda iş birliği sağlamayı hedeflemektedir. Bu alanlarda ülkelerin her yıl elde ettiği, gerçekteştirdiği değerler OECD'nin belirlediği standartlar çerçevesinde halka açık bir veri tabanı üzerinden paylaşılmaktadır. Bu çalışmanın konusunda OECD ülkelerinin seçilmesinin nedeni üye ülkelerin sağlık verilerinin OECD tarafından belirlenen standartlar çerçevesinde toplanması ve açık kaynak olarak tek bir çatı altında sunulmasıdır.

OECD veri tabanından elde edilen değerler başta yüzde, sayısal ve kategorik veriler olmak üzere birçok türde olabilir. Eksik veriler ve tahmini veriler OECD veri tabanında sıkça karşılaşılan bir durumdur. Bu durumlardan kaçınmak için veri setini seçerken veri önişlemeden geçirmek gerekir. Çalışma kapsmında kullanılacak olan OECD sağlık veri setine veriler toplandıktan sonra verideki eksiklikleri gidermek ve veri madenciliği

yöntemlerini uygulamaya hazır hale getirmek için çalışmanın ilk aşaması olan veri önişleme işlemi gerçekleştirilecektir.

Önişleme aşamasından son veri setini tanımak için paralel koordinat sisteminden yararlanılacaktır. Paralel koordinat sistemi verinin hangi gruplara ayrıldığını ve değişkenlerin hangi aralığında yoğunlaşmaların olduğunu görmek açısından yararlı bir veri analiz ve görselleştirme yöntemidir.

Veri madenciliği yöntemleri ile OECD sağlık verilerinin analiz edileceği bu çalışmada veriler üzerinde bir veri madenciliği yöntemi olan özellik çıkarımı işlemleri uygulanacaktır. OECD ülkelerindeki sağlık hizmeti harcamaları, kaynakları ve kalite göstergelerinin İGE üzerindeki etkisini belirlemek için öznitelik seçimi yöntemi kullanılarak çıkarımlar yapmak hedeflenmektedir. Özellikle OECD ülkelerinde medikal cihazların sayısındaki artışın ve anne-bebe ölümlerinin oranının insani gelişime etkisi konusunda literatürdeki eksiklikler göz önüne alınarak, yapılan çalışma ile bu eksikliğin giderilmesine katkıda bulunma ve bundan sonra sağlık verileri ile İnsani Gelişim Endeksi ilişkisi üzerine yapılan çalışmalara yol gösterici olma amacı güdülmüştür.

Ülkelerin uluslararası ölçekte gelişmişliğini ölçen İnsani Gelişim Endeksi temelde sağlık, eğitim ve gelir olmak üzere üç temel unsura dayansa da incelenen literatür çalışmaları gösterdi ki bu üç unsur dışında insani gelişim üzerinde birçok belirleyici bulunmaktadır. Bu belirleyiciler ve eldeki sınıf değerleri ile bir sınıflandırma yapısı kurup yeni gelen bir değerin hangi sınıfa ait olacağını önceden tahmin etmekte yardımcı olacaktır. Bu sınıflandırma işlemi için bu çalışma kapsamında IGE değeri üzerinde etkili olan sağlık göstergeleri yardımı ile YSA, Naive Bayes ve lojistik regresyon yöntemlerinin sınıflandırma performansları karşılaştırılacaktır.

Son olarak da veri setinde yer alan özniteliklere göre en iyi ilk on ülkenin belirlendiği analiz çalışması yapılacaktır. Kullanılacak veri seti 2011-2016 arası verileri kapsadığı için 31 ülkenin bu altı yıl içerisinde yaşadığı değişim grafikler yardımıyla görselleştirilerek sunulacaktır.

## 2. MATERYAL VE YÖNTEM

Analiz aşamasında kullanılacak olan 2011 - 2016 yılları arasındaki verilerin yer aldığı OECD sağlık veri seti, OECD ülkelerine ait verilerin bulunduğu https://stats.oecd.org/ adresinden ve İnsani Gelişim Endeksi değerleri http://hdr.undp.org/en/data adresinden 15.06.2019 tarihinde alınmıştır.

Veri setinde yer alan ülke isimleri nominal veri türündedir. Kişi başına GSYH, sağlık harcaması, yatak sayısı, hastane sayısı, bebek ölüm oranı, doğumda yaşam beklentisi, anne ölüm oranı, medikal teknolojik cihaz sayısı, kişi başına düşen ilaç harcaması ve IGE değeri nominal türde değerler içermektedir. IGE sınıfı değişkeni ise ordinal veri türündedir.

Tablo 1.2'de belirtilen özelliklerden 10.000 kişi başına düşen doktor sayısı, aşılamalar ve hastalıklılık özellikleri için OECD sağlık verilerinde yeterli kayıt bulunamamıştır. Bu nedenle bu özellikler çalışma kapsamından çıkartıldı ve Tablo 2.1'deki özniteliklerin kullanılmasına karar verildi.

Tablo 1.6'da gösterilen 37 OECD üye ülkesinden 31 tanesi çalışma kapsamına alınmıştır. Kolombiya OECD ülkeleri arasına 2020 yılında katıldığı için çalışma kapsamına alınmamıştır.

**Tablo 2.1** OECD sağlık veri seti içerisinde yer alan öznitelikler

| Değişkenler |
|---|
| Kişi başına GSYH |
| Sağlık Harcamasının Kişi başına GSYH'ye Oranı |
| 1000 Kişi Başına Düşen Yatak Sayısı |
| Hastane Sayısı |
| Bebek Ölüm Oranı |
| Doğumda Yaşam Beklentisi |
| Anne Ölüm Oranı |
| Medikal Teknoloji |
| Kişi Başına Düşen İlaç Harcaması (US $) |
| IGE Değeri |
| IGE sınıfı |

Verilerin web sitesinden alındığı tarihte Tablo 2.1'deki çalışma kapsamına alınan 11 gösterge için Belçika, Yunanistan, Portekiz, Norveç ve İsveç'e ait verilerde %20 oranında eksik veri bulunmaktadır. Bu oranlardaki eksik veriler çalışmanın sonucunu büyük oranda etkileyebileceği için bu ülkeler çalışma kapsamına alınmamıştır. Bu ülkeler haricinde Danimarka, ABD, Polonya, İsrail ve Litvanya ait verilerde tüm veri seti içerisinde toplamda %12 oranında eksik veri bulunmaktadır. Buradaki eksik veriler eğri uydurma yöntemi ve ortalama değer atama yöntemi ile tamamlandı.

## 2.1. Veri Önişleme

Veri önişleme işlemleri, kullanılan veri setine göre farklılık göstermektedir. Veri önişlemede temel amaç gereksiz, gürültülü ve eksik verilerin araştırmanın sonucunu etkilemeyecek şekilde manipüle edilmesidir [55].

OECD sağlık veri seti içerisinde yer alan diğer özelliklerdeki eksik veriler eğri uydurma ve ortalama değer atama yöntemleri kullanılarak tamamlanmıştır. Eksik değer bir zaman serisinde belirli bir yerde eksikse bu değer eğri uydurma yöntemi kullanılarak tamamlandı. Bu yöntem ile tamamlanamayan özelliklere ortalama değeri atanarak veri seti tamamlanmıştır.

### 2.1.1. Veri temizleme

#### 2.1.1.1. Eğri uydurma ile eksik veri tamamlama

Kullanılan veriler arasında doğrusal bir ilişki bulunmadığı durumlarda eksik verilerin tamamlanması için uygun bir eğri seçilerek en küçük kareler yöntemine göre hesaplama yapılır. Eğri uydurma için kullanılan en popüler metot olan en küçük kareler yöntemi; bağımlı değişkenin değerleri ile tahmin edilen değerlerinin arasındaki farkın yani hatanın karelerinin toplamının en küçük olmasını amaçlayan bir yöntemdir [56]. En küçük kareler yöntemi birçok farklı tipteki fonksiyonu kullanılarak veriye uygun eğri uydurma için kullanılır. Eğri uydurma işleminde, eğrinin formatı önceden tahmin edilemiyor olabilir. Doğrusal, ikinci ve üçüncü dereceden polinom, üslü veya logaritmik denklem formatında olabilir. Bu formatlardan verilere uygun olan bir seçilir ve daha sonra $R^2$ uyumu incelenir. $R^2$ olarak kullanılacak değer gözlem sonuçları ile tahmin edilen değer arasındaki ilişkiyi gösterir [57].

Bu çalışmada kullanılan veri setinde yer alan eksik veriler polinom denklemi eğri uydurma ve ortalama değer atama yöntemleri ile tamamlanmıştır. Bir ülkenin seçilen zaman aralığındaki verilerinde bir eksik veri varsa bu değer eğri uydurma yöntemi ile tahmin edildi. R programlama dilinde yapılan işlemlerde eğri uydurma ile değer atama için 2. ve 3. dereceden polinominal denklemler kullanıldı. Hata oranı $R^2$'nin $R^2 = 1$'e en çok yaklaştığı yöntem ile tamamlandı.

#### 2.1.1.2. Ortalama değer atama ile eksik veri tamamlama

Ortalama değer atama yöntemi, bir özelliğin eksik verilerinin tamamlanmasında eksik verilerin yerine o özelliğin ortalama değerinin atanması işlemidir [58]. Denklem 2.1'de gösterildiği gibi bir özelliğe ait eksik değere sahip olanlar hariç tüm veri noktaları belirlenir. Bu veri noktalarındaki değerlerinin toplamının veri noktasına oranı o özellik için ortalama değeri verir [59].

$$\text{Ortalama} = \frac{\text{Tüm Veri Noktalarının Toplamı}}{\text{Veri Noktası Sayısı}} \tag{2.1}$$

Eğri uydurma ile değeri belirlenemeyen eksik değerlerin değerini belirlemek için ortalama değer atama yöntemi kullanıldı. Her bir özelliğin 2011-2016 yılları arasındaki eksik değerler bulunduğu yılın ortalaması atanarak tamamlandı.

### 2.1.2. Veri bütünleştirme - birleştirme

Birden fazla kaynaktan gelen farklı türlerdeki verilerin anlamlı ve değerli hale getirilmesi için ortak tek bir türe dönüştürülmesi işlemidir. Bir veri tabanında Evet/Hayır şeklinde tutulan bir alan, başka bir veri tabanında 1/0 şeklinde tutuluyor olabilir. Bu kaynaklardan alınan örnekler bütünleştirilirken tek bir ortak türe dönüştürülür. Veri birleştirme ise farklı veri tabanlarından alınan verilerden tek bir veri seti oluşturma işlemidir. Bu çalışmada OECD sağlık verileri ve Birleşmiş milletler IGE indeksi değerleri üzerinde veri birleştirme işlemi yapıldı.

### 2.1.3. Veri indirgeme

Uygulanacak veri madenciliği yönteminin sonucunun değişmeyeceği durumlarda veri seti içerisindeki örnek sayısı ve öznitelik sayısında yapılan azaltma işlemine veri indirgeme denir [55].

### 2.1.4. Veri dönüştürme

Veri dönüştürme farklı aralıklardaki verileri belli bir aralığa çekerek uygulanacak veri madenciliği modelinin ortalama ve varyans farklılıklarından etkilenmemesini sağlar. Veri dönüştürme yöntemi kullanılacak veri madenciliği modele uygun şekilde yapılmalıdır. Veri normalizasyonu, standartlaştırma ve düzeltme temel veri dönüştürme yöntemleridir [55, 59].

#### 2.1.4.1. Min-max normalizasyonu

Verileri belirli bir minimum ve maksimum ile sınırlandırarak dönüştürme işlemidir. Genellikle [0-1] aralığı veya [-1,1] aralığı tercih edilebilmektedir. Min-max normalizasyonda veri seti içerisindeki sütundaki en büyük değer maksimum ve en düşük değer minimum kabul edilir. Aralıkta kalan değerler Denklem 2.2'deki gibi hesaplanır [59].

$$X_{yeni} = \frac{x - x_{min}}{x_{min} - x_{maks}} \tag{2.2}$$

Min-max normalizasyonu ile orijinal verileri belli sayısal bir aralığa çekmek veri madenciliğinde veri gizliliği için de tercih edilen bir yöntemdir [55]. R programlama dilinde verilerin min-max normalizasyonu için rescale() fonksiyonu kullanılır. Bu çalışmada kullanılacak OECD sağlık veri seti farklı aralıklarda verilere sahip kolonlar içermektedir. Kümeleme işlemi öncesi verilere rescale fonksiyonu ile min-max normalizasyonu yapıldı.

#### 2.1.4.2. Z-skor normalizasyonu

Z-skor normalizasyonu verilerin ortalamasını ve standart sapmasını temel alan bir yöntemdir. Hem ortalama hem de standart sapma aykırı değerlere duyarlı değildir. Bundan dolayı z-skor normalizasyonu güçlü bir yöntem değildir. z-skor normalizasyonu x data noktası, $\mu$ = ortalama ve $\sigma$ = standart sapma olacak şekilde Denklem 2.3'teki gibi hesaplanır [60].

$$z = \frac{x - \mu}{\sigma} \tag{2.3}$$

Ortalama ve standart sapma değerleri sadece Gauss dağılımı için en uygun ölçek parametresidir. Giriş skoru Gauss dağıtılmamışsa çıkış skoru giriş dağılımını korumaz [60].

## 2.2. Özellik Seçimi (Feature Selection)

Özellik seçimi, içerisinde fazla sayıda özelik bulunduran veri kümelerinde çıkış değişkenine en fazla etkisi olan özelliklerin belirlenmesi kullanılan yöntemdir. X adet özellik arasında en iyi Y tanesinin çeşitli özellik belirleme algoritmaları kullanılarak çıkarılır [61]. Özellik sayısını azaltarak gürültülü verileri veri setinden çıkararak veri setinin boyunu küçültürüz ve faydalı bilgiye bir adım daha yaklaşırız. Böylelikle veri setinden daha sade, anlaşılır ve kaliteli bilgiler elde edilebilir. Veri setindeki bu değişim kullanılacak modelde başarı oranını arttırır [62].

### 2.2.1. Korelasyon tabanlı özellik seçimi

İki değişkenin arasında ilişkinin bulunup bulunmadığını tespit etmenin en basit yöntemi bu iki değişkenin birbirlerine göre değişimini gösteren kovaryans değerini hesaplamaktır. Denklem 2.4'teki eşitlik iki değişkenin birlikte değişimini bulabilmek için her bir değerin ortalamadan farkının çarpımı kovaryans değerini vermektedir [63].

$$\text{Kovaryans } = cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(N-1)} \tag{2.4}$$

Kovaryans hesaplarken kullanılan iki değişken farklı birimlerde gösteriliyorsa bu değerleri ortak bir birime dönüştürmemiz gerekir. Bunun için kovaryans formülünü standart sapma değerleri ile bölünerek Denklem 2.5'te gösterildiği gibi korelasyon hesaplanır [64].

$$\text{Korelasyon Katsayısı } = r = \frac{cov_{xy}}{s_x s_y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y} \tag{2.5}$$

Korelasyonun büyüklüğü iki değişken arasındaki ilişkinin gücünü gösterirken işareti değişkenlerin aynı yönde pozitif (+) ya da zıt yönlerde negatif (-) artışı ve azalışı gösterir. İki değişken arasında hesaplanan korelasyon (r) değeri aşağıdaki gibi yorumlanır;

- 0-0.20 çok zayıf ilişki

- 0.20-0.39 zayıf ilişki
- 0.40-0.59 orta düzeyde ilişki
- 0.60-0.79 yüksek düzeyde ilişki
- 0.80-1.0 çok yüksek düzeyde ilişki

Korelasyon tabanlı özellik seçimi, korelasyon katsayısını kullanarak veri setinde yer alan özelliklerin çıkış değişkeni ile olan korelasyon katsayısını göz önünde bulundurarak özellikler arasında bir sıralama yapar. Sıralama sonucunda çıkış değişkeni üzerinde en etkili olan özellikler belirlenir.

### 2.2.2. Bilgi kazancı ile özellik seçimi

Entropi, bir veri seti içindeki belirsizliği, düzensizliği ve rastgeleliği ölçmek için kullanılır. Tüm olasılıkların toplamı 1' eşit olmalıdır(Denklem 2.6) [63].

$$\sum_{i=1}^{n} p_i = 1 \qquad p_1, p_2, .., p_n \tag{2.6}$$

$p_i$ olasılık değerlerini temsil eder. Yukarıdaki denklemden yola çıkarak bir sistemin entropisi Denklem 2.7'deki gibi hesaplanır.

$$I(P) = \sum_{i=1}^{n} p_i * \log(p_i) \tag{2.7}$$

Bilgi Kazancı(Information Gain), entropi ile doğru orantılıdır. Bilgi kazancının artması için entropinin azalıyor olması gerekir. Her bir niteliğin bilgi değeri Denklem 2.8 ile hesaplanır [64];

$$\text{Bilgi (Sınıf, Özellik)} = \sum_{i=1}^{n} \frac{T_i}{T} I(T_i) \tag{2.8}$$

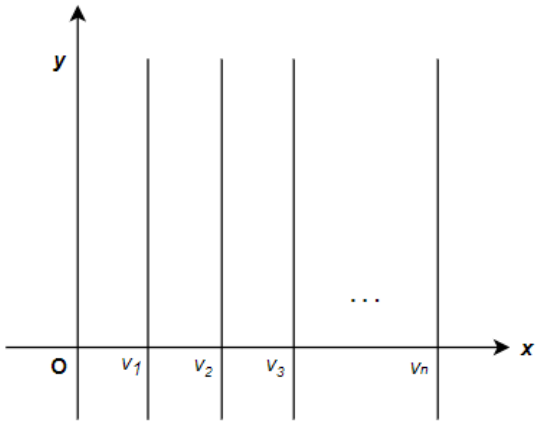
Bilgi değeri hesaplanan niteliklerin kazanç değeri aşağıdaki Denklem 2.9 ile hesaplanır;

$$\text{Kazanç (Sınıf, Özellik)} = \text{Bilgi(Sınıf)} - \text{Bilgi(Sınıf, Özellik)} \tag{2.9}$$

Bilgi kazancı ile özellik seçimi yönteminde değişkenlerin kazanç değerleri hesaplanır ve bu değerlere göre bir sıralama yapılır. Kazanç değeri en yüksek özellik sonuca en fazla etkisi olan özelliktir [65].

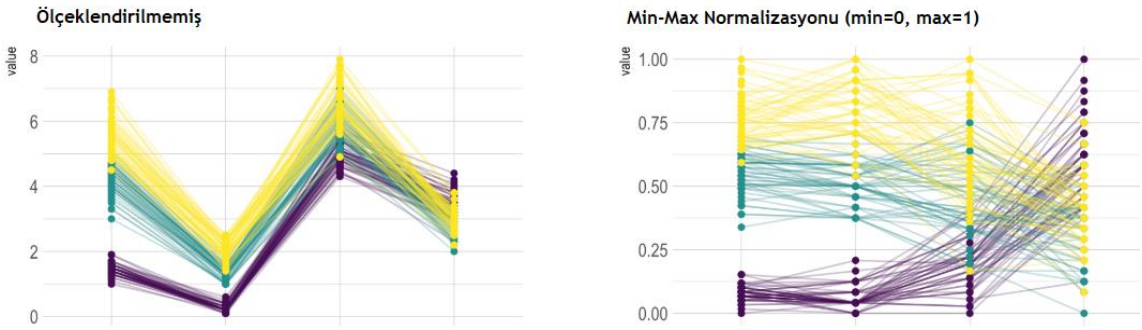
## 2.3. Paralel Koordinat Sistemi

Paralel grafik veya paralel koordinat sistemi olarak adlandırılan bu grafik bir dizi sayısal verinin gözlem değerlerine göre karşılaştırılması amacıyla kullanılır. Inselberg ve Dimsdale [66] tarafından 1990 yılında analitik ve sentetik geometriyi görselleştirmek için geliştirilmiştir. Paralel koordinat sisteminin temel gösteriminin yer aldığı Şekil 2.1'deki koordinat sisteminde her dikey çubuk bir değişkeni temsil eder ve değişkenin kendi ölçeği ile temsil edilir. Bu yöntem gösterilecek değişken sayısının az olduğu durumlarda daha iyi çalışmaktadır. Çok fazla değişken olması durumu paralel koordinat sistemini işlevsiz hale getirmektedir. Şekil 2.1'de yatay eksendeki çizgiler ise veri seti içerisinde yer alan örnekleri göstermektedir. Aynı sınıfa dahil olan örnekler aynı renk ile gösterilir. Farklı sınıflar belirtmek için farklı renkler kullanılır. Paralel koordinat sistemini en doğru şekilde oluşturabilmek için ölçeklendirme, dikey eksen yerleştirme ve vurgulama gibi yöntemler kullanılmaktadır [66].




**Şekil 2.1** Paralel Koordinat Sistemi Temel Gösterimi

Ölçeklendirme, ham verileri diğer değişkenlerle ortak bir aralığa çekerek grafiğin daha okunur hale gelmesini sağlar. Bu işlem aynı birime sahip olmayan değişkenleri karşılaştırmak için önemli bir adımdır. Şekil 2.2'de grafikte ölçeklendirilmemiş, min-max normalizasyonu ile ölçeklendirilmiş veri setine ait paralel koordinat sistemleri gösterilmiştir. Paralel koordinat sistemleri min-max normalizasyonu dışında tek değişkenli normalizasyon ve Z score normalizasyon yöntemleri ile de ölçeklendirilebilir.




**Şekil 2.2** Ölçeklendirilmenin Paralel Koordinat Sistemine Etkisi

Ölçeklendirme dışında grafiğin karmaşıklığını ve dağınıklığını azaltmanın bir diğer yolu da dikey eksende yer alan özelliklerin eksen sıralaması yapılarak optimize edilmesi işlemidir. Bu işlemin temel amacı seriler arası çapraz bağlantıları azaltarak daha okunur bir grafik elde etmektir [67].

R programlama dilinde ggparcoord kütüphanesi kullanılarak eksen sıralaması yapılabilir. Eksen sıralaması yapılan grafik eksen sıralaması yapılmamış olana oranla daha az karmaşık görünmektedir.

Paralel koordinat sistemi grafiğinin okunabilirliğini arttıran bir diğer yöntem ise belirli bir grubun vurgulanmasıdır. Bu yöntem belirli sınıflar öne çıkarılmak istendiğinde kullanılır.

## 2.4. Kümeleme

Eğitimsiz öğrenme yöntemlerinden biri olan kümeleme, benzer özellikler taşıyan verilerin birbirlerine olan uzaklıkları hesaplanarak kendilerine en yakın olan kümeye atanması işlemidir. Aynı küme içinde benzerlikler fazladır fakat kümeler arası benzerlikler azdır. Başka bir deyişle, kümeler arası benzerliğin az, küme içi benzerliğin çok olacak biçimde benzer özellikleri gösteren örnekleri aynı kümede toplayarak, özet bir bilgi elde etmektir. Birbirine en çok benzeyen örnekler bir küme içerisine toplanır ve her küme kendi içerisinde homojen bir yapıya sahiptir [68].

Kümeleme analizi işlemi ile eldeki belli miktardaki verinin sınıflandırılarak veriye anlam kazandırılması hedeflenmektedir [69]. Kümeleme analizi yöntemi; örüntü tanıma, görüntü işleme, veri analizi ve pazar araştırması gibi pek çok uygulama alanına sahiptir [70]. Veri madenciliği ve çok değişkenli istatistik yöntemlerinden biri olan kümeleme

algoritmaları içerisinde en yaygın olarak kullanılan algoritmalar hiyerarşik kümeleme, k-means ve k-medoids kümeleme yöntemleridir [71].

Kümeleme analizinde ilk adım uzaklık ölçüsünün belirlenmesi işlemidir. Sonrasında hangi kümeleme yönteminin kullanılacağı ve kullanılacak kümeleme yönteminin türü belirlenir. En başta verdiğimiz kümeleme sayısı optimum sonuç vermeyebilir. Farklı denemeler yaparak, optimum küme sayısının ne olacağı bu sonuçların karşılaştırılması ile bulunabilir [72].

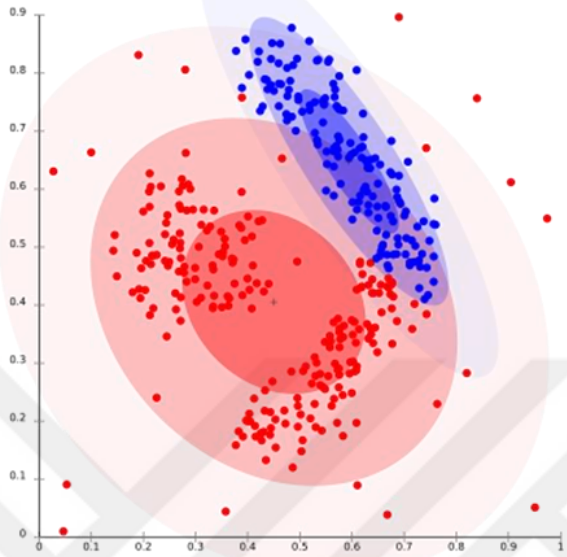
Uzaklık, iki örneğin birbirinden ne kadar farklı olduğunu gösterirken, benzerlik iki örneğin birbirine ne kadar yakın olduğunu gösterir. Büyük bir uzaklık değeri varsa karşılaştırılan iki örnek birbirine farklı, az bir uzaklık değeri varsa iki örnek birbirine çok benzerdir. Uzaklık ölçüsünün seçimi kümeleme analizinin sonucunda da bir etkiye sahip olduğu için doğru yöntemin kullanılması sonucun doğruluğuna katkı sağlayacaktır. Öklid, karesel Öklid, Mikowski, Manhatten, Chebyshev, Karl-Pearson, Hotelling T2 ve Mahalobis uzaklığı bu uzaklık ölçülerinin bazılarıdır. Öklid uzaklığı Pisagor bağlantısından yararlanarak iki nokta arasındaki uzaklığın ölçümünde kullanılan yöntemdir ve uzaklık ölçüleri içerisinde en çok tercih edilenidir [55].

### 2.4.1. Kümeleme yöntemleri

Kümeleme yöntemleri çalışma prensiplerine göre bölümlemeli, hiyerarşik, ızgara tabanlı, model tabanlı ve yoğunluk tabanlı yöntemler şeklinde incelenebilmektedir. Kümeleme algoritmaları bölümlemeli, hiyerarşik, ızgara tabanlı, model tabanlı ve yoğunluk tabanlı olarak beş gruba ayrılmaktadır [74]. Bu çalışma kapsamında bölümlemeli ve hiyerarşik algoritmalar kullanılacaktır.

**Bölümlemeli Algoritmalar:** n adet örneğin k tane bölüme ayrıldığı algoritmalara bölümlemeli algoritmalar denilmektedir. Bu tür algoritmalarda k değeri n değerine eşit veya küçük olmalıdır(k<=n). Veri seti içerisindeki her bir eleman k adet kümeden birine dâhil edilir. Nesnelerin mantıksal gruplara ayırılarak analiz edildiği bu yöntemde kullanılan veri seti küçük veya orta boyuttayken birkaç küme oluşurken, veri seti büyük boyuttaysa çok fazla sayıda küme oluşabilir. Çok fazla sayıda küme oluştuğu durumlarda bir gruplandırma ölçütü belirlenmesi gerekebilir. Gruplandırma ölçütü kullanılması kümeleme analizinin sonucunu değiştirebileceğinden büyük veri setlerinde kullanılması

uygun olmayan bir yöntemdir. k-means, k-medoids, CLARA- CLARANS ve EM(Expectation-Maximization)(Şekil 2.3) bölümlemeli algoritmalardır [55].

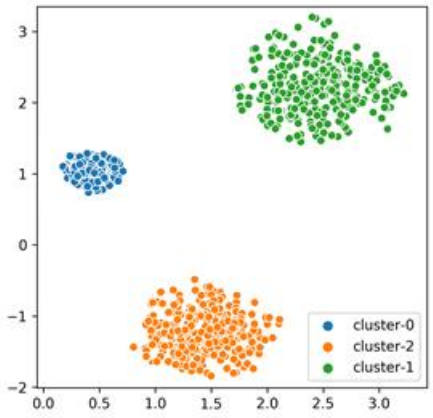



**Şekil 2.3** Expectation Maximization Kümeleme

**Hiyerarşik Algoritmalar:** Birleştirici ve ayrıştırıcı olarak iki yöntem ile yapılan hiyerarşik kümeleme eldeki veriyi hiyerarşik bir yapıya dönüştürmek için kullanılır. Birleştirici hiyerarşik kümelemede başlangıçta her örnek bir kümeyi temsil eder. Birleştirme işlemi tek bir küme elde edilene kadar devam eder. Ayrıştırıcı hiyerarşik kümelemede ise veri setindeki tüm örnekler tek bir kümede toplanır. Her örnek kendi başına bir kümeyi temsil edene kadar ayrıştırma işlemi yapılır. Her iki yöntemin sonucunda da bir hiyerarşik yapı elde edilir. Oluşan hiyerarşik yapının görselleştirilmesi için "Dendogram" kullanılır [73].

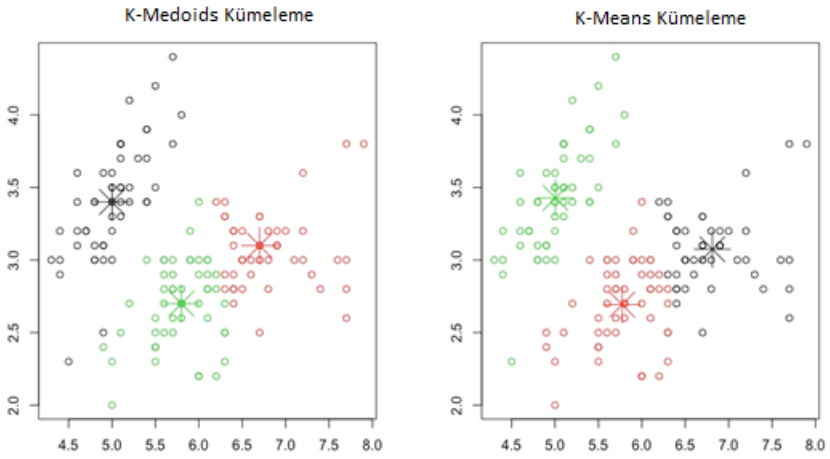
Karakteristiklerine göre kümeler, iyi ayırılmış kümeler, ortak merkezli kümeler, yoğunluk tabanlı kümeler ve bitişik kümeler olmak üzere beş türe ayrılmaktadır [74].

**İyi ayrılmış kümeler:** İyi ayrılmış kümeler, küme içindeki herhangi bir nesnenin bulunduğu küme içindeki düğümlere diğer kümelerdeki düğümlerden daha yakın olması şartını sağlayan kümelerdir. İyi ayrılmış kümeler Şekil 2.4'teki gibi birbirinden uzakta yer almaktadır. Bu kümedeki düğümler arasındaki benzerliği ölçmek için eşik değeri de kullanılabilir [75].




**Şekil 2.4** İyi ayırılmış kümeler

**Merkez tabanlı kümeler:** Küme içerisindeki nesneleri içinde bulunduğu kümenin merkezine diğer kümelerin merkezine olduğundan daha yakın olmadı durumudur [75]. Bir sentroid (kümedeki tüm noktaların ortalaması) veya bir medoid (kümedeki en temsili nokta) genellikle bir kümenin merkezini temsil edebilmektedir. Kmeans ve K-medoids yöntemlerine ait merkez tabanlı kümeleme örneği Şekil 2.5'te gösterilmiştir.

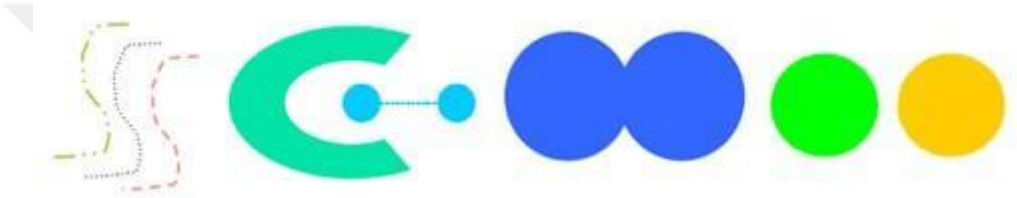



**Şekil 2.5** Merkez tabanlı kümeleme

**Yoğunluk tabanlı kümeler:** Nesnelerin yoğunluklarına göre kümelere ayırılır. Bu yöntem, kümelerin birbiri içine geçmiş, düzensiz veya gürültülü olduğu durumlarda tercih edilebilir. Şekil 2.6'da yoğunluklarına göre ayrılmış küme örnekleri gösterilmiştir.

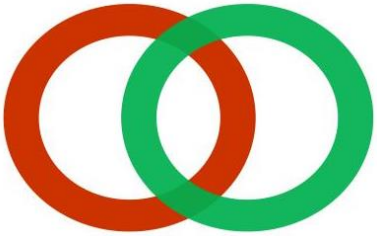
**Şekil 2.6** Yoğunluk tabanlı kümeleme

**Bitişik kümeler (en yakın komşu veya geçişli küme):** Bitişik kümeler, Şekil 2.7'de gösterildiği gibi küme içerisindeki bir noktanın aynı kümedeki bir veya daha fazla noktaya bu kümede olmayan bir noktadan daha yakın olduğu durumlarda oluşabilen kümelerdir [76].



**Şekil 2.7** Bitişik küme

**Kavramsal kümeler (shared property or conceptual clusters):** İki kümede kesişen bazı nesneler ker iki kümenin de özelliklerini taşır. Bu tür kümelerde kesişen nesneler her iki kümenin de elemanı olarak kabul edilir [76]. Şekil 2.8'de yer alan görsel kavramsal kümelerin gösterimine ait bir örnektir.



**Şekil 2.8** Kavramsal küme

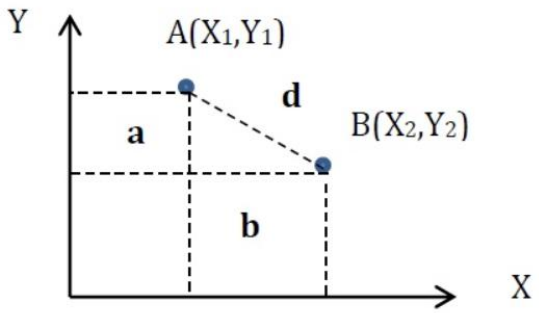
### 2.4.2. Uzaklık hesaplama yöntemleri

Uzaklık fonksiyonu seçimi araştırmada kullanılacak değişken seçimi kadar önemlidir. Seçilen uzaklık fonksiyonu kümeleme analizinin sonucuna doğrudan etki edebilmektedir [68].

Gözlemler arası benzerlikleri hesaplamada uzaklık ölçüleri kullanılır. Öklid uzaklığı, sayısal veriler için şimdiye kadar kullanılan en yaygın uzaklık hesaplama yöntemidir [77]. Yaygın olarak kullanılan diğer uzaklık ölçüleri Minkowski ve Manhattan'dır. i. gözlem ve j. gözlem arasındaki uzaklık $d_{ij}$ üç yöntem için aşağıdaki başlıklarda açıklanmaktadır. Bu üç yöntem dışında literatürde kullanılan Karesel Öklid uzaklığı, Chebyshev uzaklığı, Karl-Pearson uzaklığı, Mahalanobis uzaklığı, Hotelling T2 uzaklığı ve Canberra uzaklığı başta olmaz üzere birçok uzaklık belirleme algoritması mevcuttur. Literatürde bu çalışmada ele alınan probleme benzer problemlerin araştırıldığı çalışmalarda hiyerarşik kümeleme, k-means ve k-medoids yöntemleri için çoğunlukla Öklid uzaklık ölçüsü kullanılmıştır. Bu nedenle bu çalışmada da uzaklık ölçüsü olarak Öklid uzaklığının kullanılmasına karar verildi.

#### 2.4.2.1. Öklid uzaklığı (euclidean distance)

Veri gözlemlerinin birbirleri ile olan uzaklığını ölçmede en sık kullanılan uzaklık ölçülerinden biridir. Çok boyutlu uzayda iki birey ya da nesnenin arasındaki geometrik uzaklıktır. Öklid uzaklık ölçüsü, standartlaştırılmış veriden değil ham veriden hesaplanmaktadır [78].




**Şekil 2.9** Öklid uzaklığının koordinat sistemi ile gösterimi

Koordinat sistemindeki iki nokta arasındaki uzaklık Şekil 2.9'da gösterilmiştir. Burada A ve B noktaları arasındaki mesafe olan d Öklid uzaklık hesaplama yöntemi ile Denklem 2.10'daki kullanılarak hesaplanır. Denklem 2.10'da yer alan terimlerden;

$d_{ij}$ = i. ve j. noktanın aralarındaki uzaklığı (i=1..n , j=1..n ve k=1..p),

$x_{ik}$ = i. noktanın k. değişken değerini,

$x_{jk}$ = j. noktanın k. değişken değerini belirtir.

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \quad (2.10)$$

Denklem 2.10'daki p değişken sayısını, n ise birim sayısını göstermektedir. Öklid uzaklığı iki nokta arasındaki doğrusal uzaklığı hesaplamak için kullanılır ve tek, iki ve üç boyutlu olarak hesaplanabilir [79].

### 2.4.3. K küme değerinin belirlenmesi

#### 2.4.3.1. Silhouette küme geçerlilik indeksi

Silhouette küme geçerlilik indeksi, önceden bir bilgiye ihtiyaç duymadan kümelerin birbirine uzaklığına ve küme içindeki uzaklıklara bakarak kümeleme işlemi performansını hesaplayabilmektedir [80]. Hem küme içi hem de kümeler arası mesafeler dikkate alınarak formüle edilir. Belli bir x(i) noktası için ilk önce kümedeki tüm noktalara olan uzaklıkların ortalaması Denklem 2.11 ile hesaplanır. Hesaplanan değer a(x)'e atanır.

$$a(x) = \frac{1}{n_{C_i}-1} \sum_{\forall y \in C_i} dist(x,y) \qquad x \neq y \ , \ C = C_1, C_2, .. , C_k \quad (2.11)$$

Sonrasında $x_i$'yi içermeyen tüm kümelerdeki noktalar ile $x_i$ arasındaki ortalama mesafe b(x) Denklem 2.12 ile hesaplanır.

$$b(x) = minimum\left\{\frac{1}{n_{C_i}} \sum_{\forall y \in C_i} dist(x, y_i)\right\} \; x \in C_i \; ve \; j = 1,2,..,k \quad (2.12)$$

a(x) ve b(x) değerleri kullanılarak bir noktanın Silhouette katsayısı Denklem 2.13a ve Denklem 2.12b yardımıyla elde edilir. Veri setinde yer alan tüm Silhouette katsayılarının ortalaması S(C) Silhouette genişliği olarak adlandırılır. Bu değer kümelemenin kalitesini değerlendirmek için kullanılır [81, 82].

$$S(C_i) = \frac{1}{n_{C_i}} \sum_{\forall x \in C_i} \frac{b(x)-a(x)}{maksimum\{a(x),b(x)\}} \qquad i = 1,2,..,k \quad (2.13a)$$

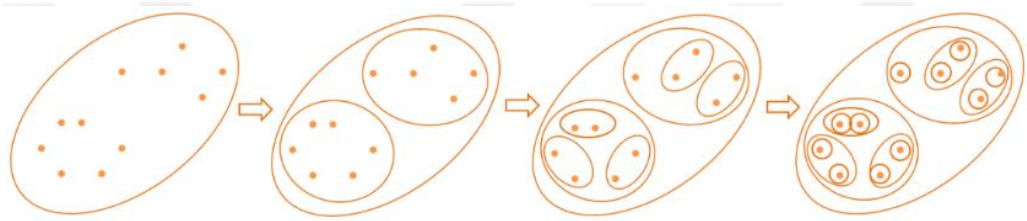
$$S(C) = \frac{1}{k}\sum_{i=1}^{k} S(C_i) \quad (2.13b)$$

Silhouette küme geçerlilik indeksi($S$) değeri ne kadar yüksekse o kadar iyi bir kümeleme performansı gerçekleşmiş demektir [82]. Ortalama Silhouette Genişliği S(C) küme geçerliliğinin tespitinde kullanılmasının haricinde kümeleme işlemlerinde optimum küme sayısının belirlenmesi için de kullanılır. Bu çalışmada hiyerarşik kümeleme, kmeans ve

k-medoids kümeleme çalışmalarında optimal küme sayısını belirlemek için Ortalama Silhouette Genişliğinden yararlanılmıştır.

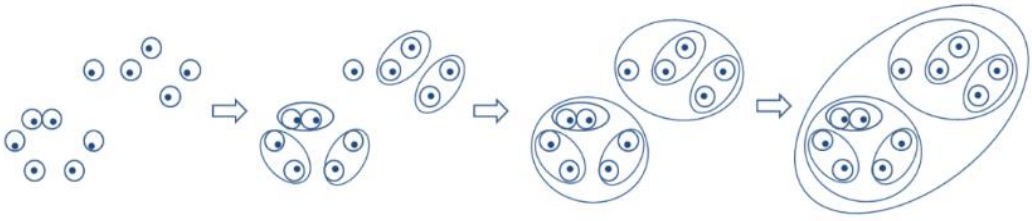
### 2.4.4. Hiyerarşik kümeleme

Hiyerarşik kümelemede ayrıştırıcı ve birleştirici yaklaşım ile alt küme oluşturulur. Ayrıştırıcı yöntemde veri setindeki tüm örnekler en başta tek bir küme ile temsil edilir. Her adımda benzerliklerine göre küçük kümelere ayrılırlar. Bu ayrışma işlemi Şekil 2.10'de gösterildiği gibi her bir örnek kendi başına bir kümeyi temsil edene kadar devam eder.



**Şekil 2.10** Ayrıştırıcı(Divisive) Hiyerarşik Kümeleme

Birleştirici (Agglomerative) hiyerarşik kümeleme ise ayrıştırıcı hiyerarşik kümelemenin tam tersidir. Bu yöntemde en başta her bir örnek bir kümeyi temsil ederken sonuçta tüm örnekler bir kümeyi oluşturur. Şekil 2.11'de birleştirici hiyerarşik kümelemenin işlem aşamaları gösterilmiştir.



**Şekil 2.11** Birleştirici (Agglomerative) Hiyerarşik Kümeleme

Birleştirici ve ayırıcı hiyerarşik teknikler, aşama sırasını birbirlerine göre ters yönde inşa etmektedirler [73].

Hiyerarşik kümeleme analizinde kümeleri birleştirebilmek amacıyla birçok bağlantı yöntemi (linkage method) bulunmaktadır. Tek bağlantı, tam bağlantı, ortalama bağlantı ve Ward yöntemi bunların başlıcalarıdır. Her bir kümeleme yönteminin aynı veri setine

uygulanması sonucu farklı kümeleme sonuçlarına ulaşılabilmesi nedeniyle en uygun ve güçlü tek bir yöntemin seçilmesi gereklidir. Bunu sağlamak için sadece tek bir yöntem keyfi olarak seçilmek yerine elde edilen sonuçların tutarlılığını pekiştirmek amacıyla tüm yöntemlerin denenmesi uygun görülmüştür.

#### 2.4.4.1. Birleştirici(agglomerative) hiyerarşik kümeleme

Hiyerarşik kümeleme yöntemleri içerisinde en yaygın şekilde kullanılan yöntem birleştirici hiyerarşik kümeleme yöntemidir [83]. Veri seti içerisindeki her bir örnek ayrı bir küme olacak şekilde ayrılır. Daha sonra kullanılan benzerlik yöntemine bağlı olarak bu kümeler birleştirilir. Bu birleştirme işlemi bütün örnekler bir kümeye ait olana kadar iteratif bir şekilde devam eder. Kullanılan bağlantı yöntemine göre kümeler bir araya getirilir. Agglomerative (Birleştirici) hiyerarşik kümelemede mesafe hesaplamak için birçok yol vardır. Tek, tam ve ortalama bağlantı yöntemi, merkez bağlantı yöntem ve Ward yöntemi bunların en önemlileridir [84].

**Tek bağlantı (single-link):** İki küme arasındaki uzaklık birbirine en yakın iki elemanın arasındaki uzaklık olarak kabul edilir. c1 ve c2 olarak adlandırılan iki kümenin en yakın iki elemanı olan x1 ve x2 arasındaki uzaklık U ile gösterilir ve Denklem 2.14 ile hesaplanır [85].

$$U(c1\text{-}c2) = \min(x1 - x2) \quad x1 \in c1 \text{ ve } x2 \in c2 \tag{2.14}$$

**Tam bağlantı (complete-link):** İki küme arasındaki uzaklık birbirine en uzak iki elemanın arasındaki uzaklık olarak kabul edilir. c1 ve c2 olarak adlandırılan iki kümenin en uzak iki elemanı olan x1 ve x2 arasındaki uzaklık U ile gösterilir ve Denklem 2.15 ile hesaplanır [86].

$$U(c1\text{-}c2) = \max(x1 - x2) \quad x1 \in c1 \text{ ve } x2 \in c2 \tag{2.15}$$

**Ortalama bağlantı (average-link):** iki küme elemanları arasındaki uzaklıklar hesaplanır ve bunların ortalaması iki küme arasındaki uzaklık(U) olarak kabul edilir ve Denklem 2.16 ile hesaplanır [87].

$$U(c1, c2) = \frac{1}{|c1|}\frac{1}{|c2|}\sum_{x1}\sum_{x2} U(x1, x2) \tag{2.16}$$

**Merkez (centroid) bağlantı:** Bir kümedeki örneklerin ortalama değeri ile diğer kümelerin ortalama değerleri ile bağlantı kurmak için kullanılır. Birinci kümenin (c1) merkezi (p elemanlı ortalama vektör) ve ikinci kümenin (c2) merkezi arasındaki uzaklık Denklem 2.17 ile hesaplanır. Bir kümede bir örnek varsa bu örnek değer merkez kabul edilir [88].

$$U(k_1, k_2) = U\left(\left(\frac{1}{|k_1|}\sum_{x\in k_1}\vec{x}\right), \left(\frac{1}{|k_2|}\sum_{x\in k_2}\vec{x}\right)\right) \tag{2.17}$$

**Ward yöntemi**: Toplam küme içi varyansı minimize ederek, küme içi kareli sapmalardan yararlanarak hata kareler toplamını hesaplar [88]. Ward yöntemi Denklem 2.18'deki gibi hesaplanır.

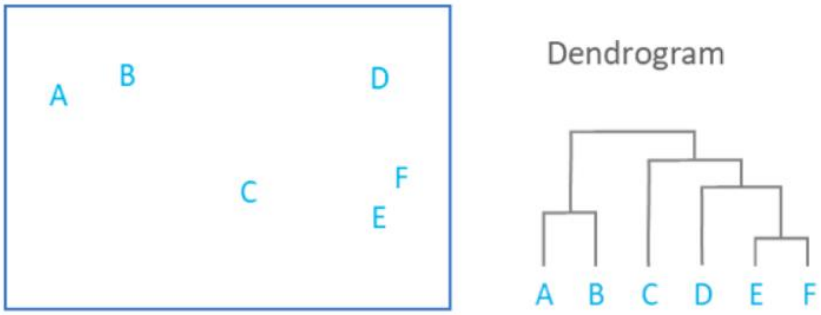
$$TU_{k_1\cup k_2} = \sum_{x\,\in\, k_1\cup k_2} U(x, \mu_{k_1\cup k_2})^2 \tag{2.18}$$

Her bir kümeleme yönteminin aynı veri setine uygulanması sonucu faklı kümeleme sonuçlarına ulaşılabilmesi nedeniyle en uygun ve güçlü tek bir yöntemin seçilmesi gereği açıktır. Bunu sağlamak için sadece tek bir yöntem keyfi olarak seçilmek yerine elde edilen sonuçların tutarlılığını pekiştirmek amacıyla tüm yöntemlerin denenmesi uygun görülmüştür.

### 2.4.4.2. Dendrogram

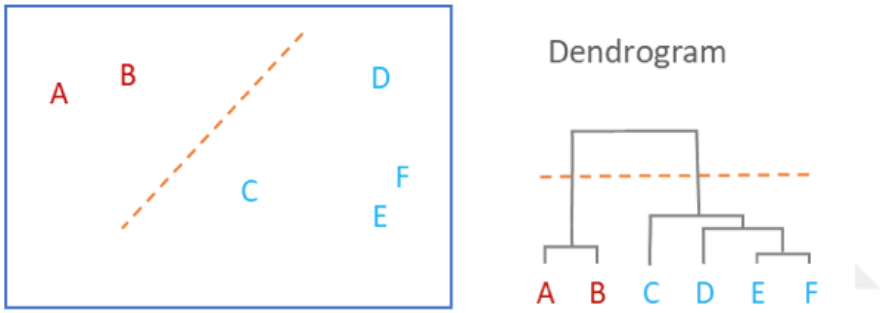
Dendrogram hiyerarşik kümeleme sonuçlarının bir ağaç grafiği üzerinde görselleştirilmesinde ve küme sayısının belirlenmesinde kullanılan bir yöntemdir. Dendrogram dikey eksende nesneleri ve kümeleri, yatay eksende kümeler arası mesafeyi ve farklılığı gösterir [106].

Bir dendrogramı yorumlayabilmek için iki nesnenin bir araya getirildiği yüksekliğe odaklanmak gerekir. Şekil 2.12'de dendrogramın yapısı verilmiştir. E-F'yi birleştiren bağlantının yüksekliğinin küçük olması birbirlerine ne kadar çok benzediklerini göstermektedir. Aynı şekilde A-B ikilisi de birbirine çok benzemektedir. Dendrogramın yüksekliği kümelerin birleştirilme sırasını göstermektedir.




**Şekil 2.12** Dendrogram yapısı

Her iki kümenin birleşimi, Şekil 2.13'deki gibi yatay bir çizginin iki yatay çizgiye bölünmesi şeklinde dendrogram üzerinde gösterilir. Çizginin altında kalan birleştirilmiş olan gözlemler kümeleri temsil etmektedir.

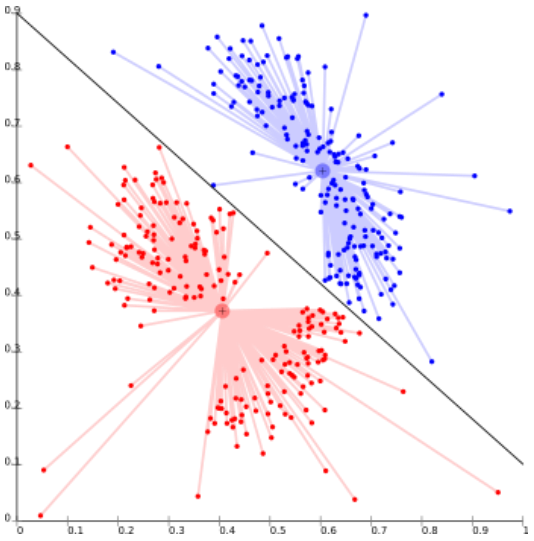



**Şekil 2.13** Dendrogramı bölümleme

Bu çalışmada dendogram yardımıyla ülkelerin sağlık değerlerinin hiyararşik kümeleme sonuçlarını dendogram ile gösterildi. Ülkeler yatay eksende(x ekseni) gösterildi. Dikey eksende(y ekseni) ise küme uzaklıkları gösterildi.

### 2.4.5. K-Means kümeleme algoritması

En çok kullanılan öğreticisiz öğrenme algoritmalarının başında 1967 yılında MacQueen tarafından geliştirilmiş olan k-means kümeleme algoritması gelir [89, 90]. K-means kümelemede Şekil 2.14'deki gibi her küme içerisinde yer alan noktaların ortalamasına karşılık gelen bir merkez ile bir diğer adıyla sentroid ile temsil edilir. K-means algoritmasının isminde yer alan k küme sayısını belirtir. Kullanılan veri seti k adet kümeye bölünür. Küme içerisindeki benzerlikler veriler arasındaki uzaklıklara göre belirlenir [77, 91]. K-means küme için varyasyonları azaltarak kümeleme yapmayı sağlar.



**Şekil 2.14** K-means Kümeleme Örneği

K-means algoritması ile kümeleme yapmanın ilk aşaması kümeleme sonucunda oluşacak k küme sayısının belirlenmesidir. k küme sayısı belirlendikten sonra rastgele k tane merkez nokta seçilir. Her veri ile rastgele belirlenen merkez noktaları arasındaki uzaklık hesaplanarak veri en yakın olduğu merkez noktasına göre bir kümeye atanır. Daha sonra her küme için yeniden bir merkez noktası seçilir ve yeni merkez noktalarına göre kümeleme işlemi yapılır. Bu durum sistem kararlı hale gelene kadar yani küme elemanları sabit kalana kadar devam eder [92].

K-Means algoritmasının işlem basamakları şu şekildedir [93]:

Adım 1: Küme sayısı k değeri belirlenir. K değerini belirlemek birçok yöntem bulunmaktadır. Elbow methodu, ortalama Silhouette genişliği ve DB indeksi k değerinin belirlenmesinde kullanılabilir.

Adım 2a: Rastgele k tane nesne seç. Seçilen k adet nesne kümelerin merkezini temsil etmektedir ve kümelerin merkezi $M_1$, $M_2$, .. , $M_k$ şeklinde gösterilir. Bu değerler Denklem 2.19'daki gibi hesaplanır.

$$M_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ik} \tag{2.19}$$

Adım 2b: Denklem 2.20a ve 2.20b ile Karesel Hata hesaplanır. k kümenin karesel hata değerleri $e_1$, $e_2$ .., $e_k$ şeklindedir. K kümesini içeren tüm kümelerin karesel hata değeri hesaplanır.

$$e_i{}^2 = \sum_{i=1}^{n_k}(x_{ik} - M_k)^2 \quad (2.20a)$$

$$E_k{}^2 = \sum_{k=1}^{k} e_k{}^2 \quad (2.20b)$$

Adım 3: Her bir örnek kendine en yakın kümeye atanır.

Adım 4: Atama işlemi tüm veriler için tamamlandığında, tekrar bu k adet küme için Denklem 2.19 kullanılarak merkezin hesaplanması işlemi gerçekleştirilir.

Adım 5: Küme merkezleri değişmeyene kadar Adım 2a, Adım 2b ve Adım 3 tekrarlanır. Küme merkezi değişmediğinde kümeleme işlemi tamamlanmış olur [94].

Bu çalışmada R programlama dilinde yukardaki adımlara göre çalışan "kmeans" fonksiyonu kullanıldı. K-means algoritmasında kullanılacak veri seti nümerik data frame, nümerik matris veya nümerik vektör olabilir. Bu fonksiyon sonuçları list formatında döndürmektedir. Elde edilen kümeleme sonuçlarının görselleştirilmesi için R'da fviz_cluster fonksiyonu kullanılması planlanmaktadır. Bu fonksiyon ikiden fazla boyut olduğu zaman PCA yaparak varyansın çoğunu tanımlayan ilk iki temel bileşene göre bir görselleştirme elde etmemizi sağlar [92].

### 2.4.6. K-Medoids kümeleme

K-means algoritması gürültülü verilere duyarlıdır yani gürültü de kümelere dahil olmaktadır. Bu olumsuzluğu gidermek için k-medoids algoritması 1990 yılında Kaufman ve Rousseeuw tarafından önerilmiştir [73]. K-medoids, Kmeans'e kıyasla aykırı değerlere daha dayanıklı bir kümeleme algoritmasıdır [95]. K-means hesaplama süresi açısından verimli bir algoritma olmasına karşın aykırı değerlere karşı duyarlıdır [96]. K-medoids algoritmasının birçok çeşidi bulunmaktadır. PAM(Partitioning Around Medoids), CLARA ve CLARANS başlıca K-medoids algoritmalarıdır [97]. Küçük ve orta büyüklükte veri setlerini medoidlere göre parçalama yönteminin sunulduğu PAM ilk ortaya çıkan k-medoids algoritmalarından bir tanesidir. CLARA ve CLARANS algoritmaları da büyük kümelerde etkili performans gösteren k-medoids algoritma çeşitleridir. K-medoids algoritması, temelde karesel hatalar toplamı yerine mutlak hata ölçütünü en aza indirmeyi hedefler. K-medoids algoritmaları istisna verinin küme merkezini kaydırması problemini gidermek için küme merkezini her yeni eklenen nesne için tekrar hesaplayarak tüm noktalara uzaklığı minimum olan noktayı küme merkezi olarak değiştirir [98]. Tüm bölümlemeli kümeleme algoritmalarında olduğu gibi k-

medoids algoritmasında da temel amaç k adet temsilci nesneyi belirleme işlemi yapılır. Veri seti için uygun uzaklık hesaplama yöntemi seçildikten sonra k adet küme merkezi rastgele seçilir.

PAM algoritmasının iki aşaması vardır. Birinci aşamada rastgele seçilen k adet nesne kullanılır. Algoritmanın ikinci aşamasında ise tespit edilen yeni medoid değerlerine göre küme merkezlerinde değişim yapılır.

K-Medoids(PAM) kümeleme algoritmasının işlem adımları aşağıdaki gibidir:

Adım 1: Veri seti yüklenir.(U=Tüm nesneler)

Adım 2: Küme sayısı k değeri belirlenir.

Adım 3: Rastgele seçilen k adet nesne temsilci nesne olarak tanımlanır. (S = seçilen nesneler)

Adım 4: Geriye kalan temsilci olmayan nesneler (O = U – S) en yakın olduğu kümeye atanır. Bu atama işleminde nesne ve küme arasındaki uzaklık, uzaklık hesaplama yöntemi ile yapılır.

Adım 5: Temsilci olmayan nesneler (O) arasından rastgele bir nesne (m) seçilir.

Adım 6: m nesnesinin toplam değişim maliyeti C hesaplanır.

Adım 7: Eğer C < 0 ise m'yi yeni küme merkezi olarak ata.

Adım 8: Değişiklik olmayana kadar 4. ve 7. Adım arasındaki işlemler tekrarlanır.

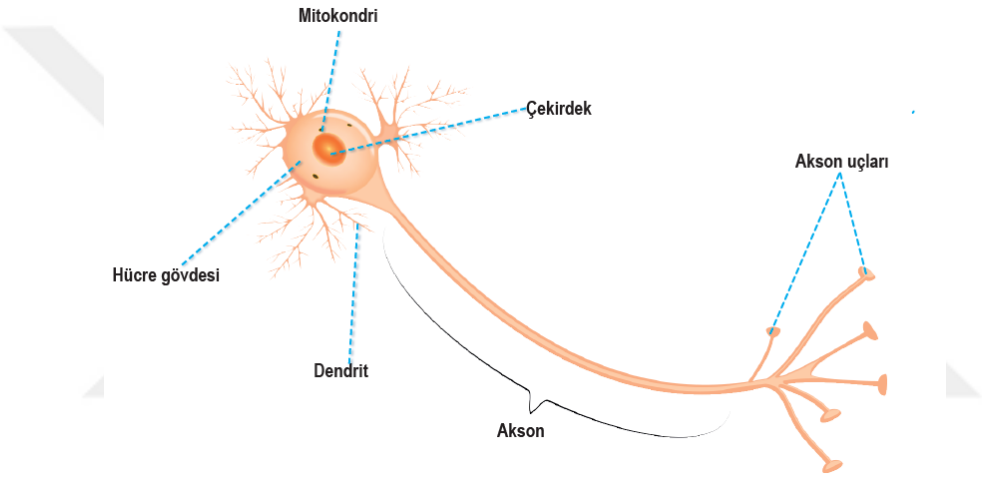
Bu çalışmada kullanılan OECD sağlık verileri veri seti büyük boyutlu bir veri seti olmadığı için PAM yöntemi tercih edilmiştir.

## 2.5. Sınıflandırma

Veri madenciliğinde sınıflandırma önemli veri sınıflarını tanımlayan ortaya çıkarmak için kullanılan bir veri analizi yöntemidir [99]. Tahmin edici yaklaşımlar içerisinde uygulama açısından zordan kolaya yapılacak sıralamada algoritmalar YSA, regresyon, kural indüksiyon ve karar ağaçları şeklinde sıralanır. Yapay sinir ağları, regresyon, karar ağaçları, Bayes yöntemleri ve genetik algoritmalar başlıca sınıflandırma yöntemleridir [100].

### 2.5.1. Yapay sinir ağları

Günümüzde yapay zeka kavramının ön plana çıkması ile birlikte yapay sinir ağları(YSA) kavramı ortaya atılmıştır. Yapay zekanın alt kollarından biri olan YSA yapay zekanın alt dallarından bir tanesidir. Yapay sinir ağları(YSA), insan beyninin çalışma şeklinden ilham alan veri madenciliği ve derin öğrenme yöntemidir. Bir insan beyni, duyular tarafından gönderilen verileri kullanarak çok miktarda bilgiyi işleyebilir. Bu işlem içinden geçen elektrik sinyalleri üzerinde çalışan ve flip-flop mantığı uygulayan nöronlar tarafından yapılır. Şekil 2.15'te bir nöronun temel yapısı gösterilmiştir. Bir nöronun 3 temel bileşeni akson uçları, dentrit ve hücre gövdesidir [101].




**Şekil 2.15** Nöronun yapısı

Dentrit, nöronlar arasında elektrik sinyallerinin iletilmesini sağlayan girdi noktalarıdır. Hücre gövdesi dentritten gelen girdileri işleyerek çıkarımlar oluşturur ve hangi eylemin yapılacağına karar verir. Akson uçları ise çıktıları bir sonraki nörona elektriksel olarak iletmekle görevlidir [102]. YSA'da yer alan yapay nöronlar beyindeki nöronlar ile aynı mantıkta çalışan ve üç katman halinde bir araya gelen sinir ağını oluşturur. YSA'nın temel elemanları [109];

- Giriş katmanı
- Çıkış katmanı
- Gizli katman
- Ağırlıklar
- Aktivasyon fonksiyonu
- Toplama fonksiyonu

Nöronun yapısında yer alan bileşenlerin YSA'daki karşılıkları Tablo 2.2'de verilmiştir.

**Tablo 2.2** İnsan beynine ait nöron yapısının bileşenlerinin YSA'daki karşılıkları

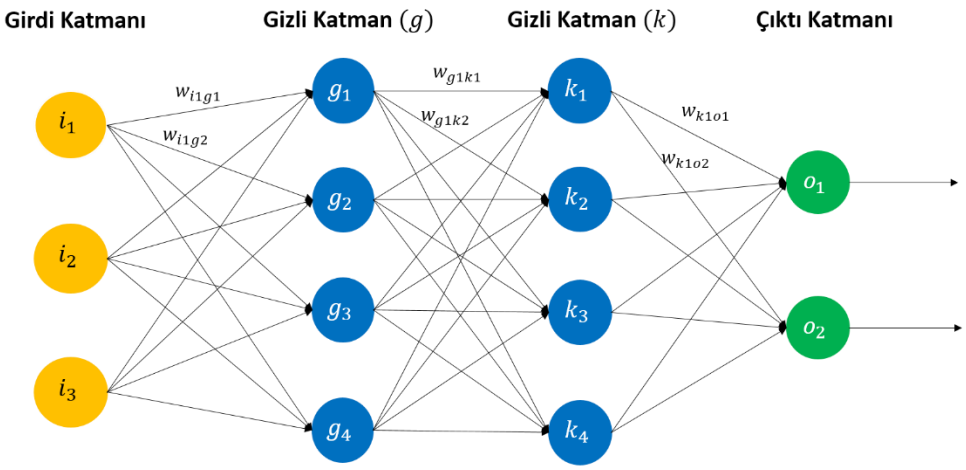
| Nöronun Yapısı | YSA'daki Görevi |
|---|---|
| Dentrit | Toplama fonksiyonu |
| Çekirdek | Aktivasyon fonksiyonu |
| Akson | Çıkış katmanı |
| Sinaps | Ağırlık |

YSA, birden fazla girişten bir çıkış oluşturma prensibine göre çalışır. YSA, tahmin, çıkarım, sınıflandırma, optimizasyon sorunlara çözüm bulma ve hızlı işlemler yapma gibi birçok amaçla kullanılabilir. Ayrıca veri ilişkilendirme, veri sıkıştırma ve veri filtreleme işlemlerinde de kullanılabilmektedir [34, 103].

Yapay sinir ağlarının güçlü yönleri;

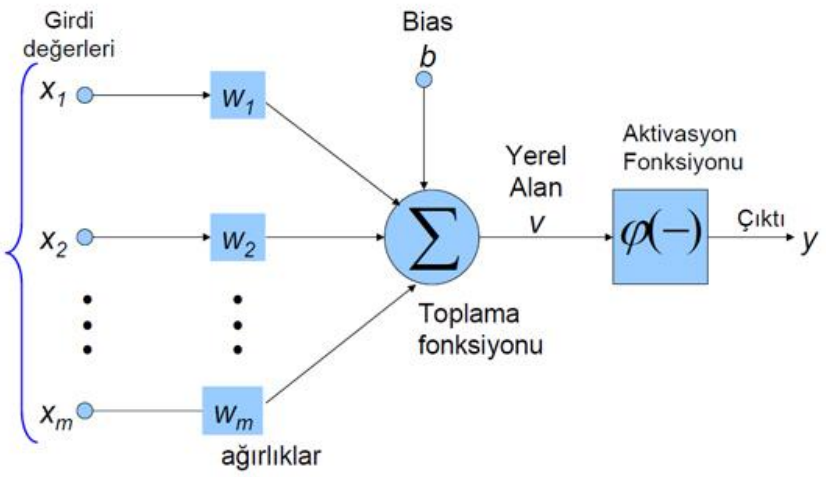
- Bir matematiksel modele ihtiyaç duymazlar.
- İlk girdilerden ve onların ilişkilerinden öğrenme işlemini gerçekleştirdikten sonra, görünmeyen veriler üzerinde görünmeyen ilişkileri de çıkarabilir, böylece modeli görünmeyen verileri genelleştirebilir ve tahmin edebilir.
- YSA'lar doğrusal olmayan ve karmaşık ilişkileri öğrenme ve modelleme yeteneğine sahiptir.
- Öğrenme işlemi için birçok öğrenme algoritmasına sahiptir.
- Eğitilmiş olan bir YSA'nın çalıştığı ortamda küçük değişiklikler meydana geldiğinde yeniden eğitilebilir. Bu da YSA'nın "uyarlanabilirlik" özelliğine sahip olduğunun göstergesidir.
- YSA'lar hatalara karşı toleranslıdır. YSA'ların eksik bilgilerle çalışabilme özelliği hata toleransına sahip olduğunu gösterir.

Yapay sinir ağlarının genel yapısı Şekil 2.16'te gösterilmiştir. Yapay sinir ağları yöntemi üç katmandan oluşur. Bunlar; giriş katmanı, gizli katman ve orta katmandır.




**Şekil 2.16** Yapay sinir ağlarının genel yapısı

Giriş katmanı gelen giriş değişkenlerini herhangi bir işlemden geçirmeden gizli katmana aktarmak ile görevlidir. Gizli katman bir veya birkaç katmandan oluşabilir. Giriş katmanından gelen değişkenlerin işlenerek çıkış katmanına aktarılmasından sorumludur [104]. Bu üç katman içerisinde yer alan temel YSA bileşenleri ve iş akışı Şekil 2.17'de gösterilmiştir.




**Şekil 2.17** YSA'nın temel bileşenleri ve birbiri ile ilişkisi

Aktivasyon fonksiyonu iki önemli amaca hizmet eder bunlar girişler arasında doğrusal olmayan ilişkiyi yakalamak ve girişi daha kullanışlı bir çıktıya dönüştürmeye yardımcı olmaktır. İncelenen literatür çalışmalarında en fazla tercih edilen yöntem olan sigmoid aktivasyon fonksiyonu 0 ile 1 arasında bir değer oluşturur. Sigmoid dışında step, tanh, softmax and RELU gibi aktivasyon fonksiyonları da mevcuttur.

Yapay sinir ağları gözetimli ve gözetimsiz olmak üzere iki şekilde eğitilebilir. Gözetimli öğrenmede ağa veriler eğitim(train) ve test seti olarak verilir. YSA'yı eğitmek için verilen veri setine eğitim veri seti deni. Eğitim veri setinde kullanılmamış test amaçlı kullanılacak verilere de test veri seti denir. Gözetimsiz öğrenmede ise veriler tek bir girdi veri seti olarak ağa verilir ve problemin çözümü ağdan beklenir [105].

YSA'da iki tip ağ kullanılmaktadır. Bunlar [107];

**İleri beslemeli ağ modeli:** bu modelde her bir katmanda bulunan hücreler sadece bir önceki katmandan aldığı değerden beslenir.

**Geri beslemeli ağ modeli:** bir katmanda yer alan hücrelerden en az bir tanesinin sonraki katmanlardan beslenmesi şeklinden çalışan ağ modelidir. Bu yöntemde en az bir hücrenin çıkışı kendisine yada diğer hücreler diğer hücrelere giriş olarak verilir. Buna geri besleme denilmektedir.

Yapay sinir ağları giriş, gizli ve çıkış olmak üzere üç katmandan oluşmaları nedeniyle MLP (Çok Katmanlı Algılayıcı – Multilayer neural networks) olarak da adlandırılır. MLP tam bağlantılı geri yayılımlı bir yapay sinir ağıdır. Geri yayılım, gradyan inişini(gradient descent) ağırlıklara göre hesaplamak için MLP tarafından kullanılan bir öğrenme algoritmasıdır. Gradyan inişi, bir maliyet fonksiyonunu en aza indiren fonksiyonun parametrelerinin değerlerini bulmak için kullanılan bir optimizasyon algoritmasıdır. MLP'de hedef çıktı ile elde edilen sistem çıktısı arasındaki fark ölçülür ve daha sonra hata fonksiyonu hesaplanır. Hatayı en düşük düzeye indirmek için bağlantı ağırlıklarının değeri yeniden hesaplanır [104, 107].

Tablo 2.2'de verilen YSA'nın çalışma prensibi göz önüne alındığında, hata oranının kabul edilebilir olmadığı durumlarda ağdaki bağlantılar için uygun ağırlıklar yeniden belirlenmelidir. Bunun için ileri beslemeli yapay sinir ağlarında birimlerin son yapılan tahmine katkısının gücü hesaplanır ve gizli birimlere giden bağlantıların ağırlıkları bu hesaplamaya göre güncellenir. Ağırlıkların yeniden güncellenmesi için gradyan indirgeme yöntemi kullanılır. Gradyan indirgeme işlemi toplu veya stokastik(rastsal) şekilde olabilir. Gradyan indirgeme yönteminin bir türü olan stokastik gradyan indirgeme(SGD) toplu gradyan indirgemeye oranla daha hızlı sonuç vermektedir [102].

Tablo 2.3'te yapılacak sınıflandırma çalışmasında kullanılması hedeflenen çok katmanlı geri beslemeli öğreticili bir yapay sinir ağı algoritmasının çalışma prensibi gösterilmiştir [111].

**Tablo 2.3** Öğreticili geri beslemeli yapay sinir ağı algoritmasının çalışma prensibi

| YSA Algoritması İşlem Adımları |
|---|
| 1.Adım: Veri setini eğitim ve test verisi olarak ikiye ayırılır. |
| 2.Adım: Başlangıç ağırlıklarını rastgele olacak şekilde seçilir. |
| 3.Adım: Eğitim veri seti giriş katmanına uygulanır. |
| 4.Adım: İşlemci elemanlar üzerinden çıkış değeri hesaplanır. |
| 5.Adım: Hata oranı kabul edilebilir düzeyde mi? Hata oranı kabul edilebilir düzeyde değilse gradyan indirgeme ile ağırlıkları yeniden düzenlenir. 3. ve 4. Adım tekrar edilir. |
| 6.Adım: Test aşamasına başlanır. |
| 7.Adım: Test veri seti giriş katmanına uygulanır. |
| 8.Adım: İşlemci elemanlar üzerinden çıkış değeri hesaplanır. |
| 9.Adım: Ağın çıkış değeri hesaplanır. |

R programlama dilinde YSA algoritması için neuralnet() fonksiyonundan yararlanılacaktır. Bu fonksiyon gradyan inişi olarak SGD içerdiği için tercih edildi.

#### 2.5.1.1. Toplama fonksiyonu

Girdi değişkenleri ağırlıklandırdıktan sonra elde edilen değerler toplama fonksiyonuna gönderilir. Hücreye gelen net girdiyi hesaplamak için kullanılır. Sinir ağlarında kullanılan birçok toplama fonksiyonu bulunmaktadır. Bunlar ağırlıklı toplama, çarpım, maksimum, minimum ve kümülatif toplam gibi yöntemlerdir. Toplama fonksiyonları içerisinde en yaygın kullanılan fonksiyonlardan biri ağırlıklı toplama yöntemidir. Girdi değerleri $x_i$ ve ağırlık değerleri $w_i$ çarpımlarının toplamı şeklinde denklem 2.21'deki gibi hesaplanmaktadır [110].

$$\text{Toplama fonksiyonu} = \text{f(x)} = \sum_{i=1}^{j} x_i w_i \tag{2.21}$$

Bir hücre kendi toplama fonksiyonunu kullanma yetisine sahiptir. Tüm hücreler aynı toplama fonksiyonunu da kullanabilir.

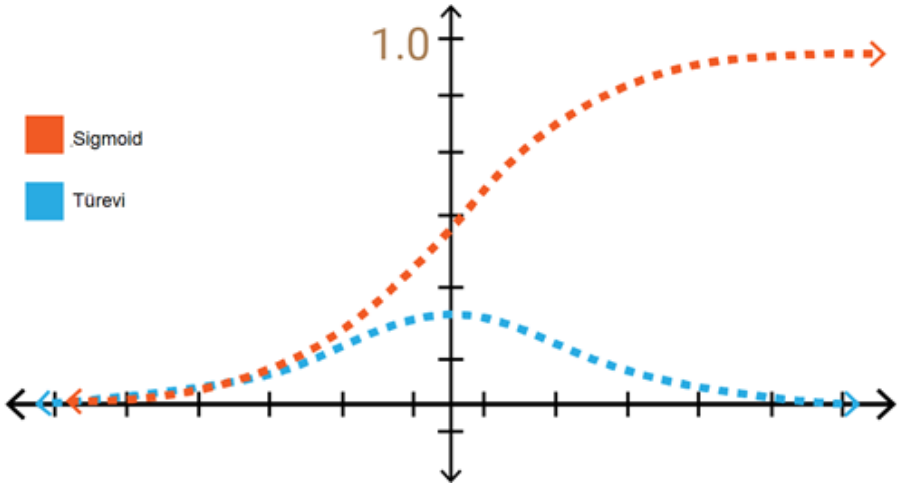
#### 2.5.1.2. Aktivasyon fonksiyonu

Bir aktivasyon fonksiyonu, bir önceki katmandan aldığı verileri beklenen çıktıya daha yakın ve anlamlı bir gösterime dönüştüren matematiksel fonksiyondur. İleri beslemeli ağların kullanılacağı çalışmalarda aktivasyon fonksiyonu seçilirken fonksiyonun lineer olmamasına ve fonksiyonun türevinin kolay alınabilir olmasına dikkat edilmelidir. Burada türevinin kolay alınabilir olmasının sebebi daha kullanışlı bir çıktı elde edebilmektir. Tablo 2.4'te YSA'larda sık kullanılan aktivasyon fonksiyonları ve hesaplanma yöntemleri verilmiştir. Burada x ve y koordinat sisteminde x ve y eksenindeki değerleri göstermektedir.

**Tablo 2.4** YSA'da kullanılan aktivasyon yöntemleri

| Aktivasyon Fonksiyonu | Hesaplanma Yöntemi |
|---|---|
| Step Fonksiyonu | $\mathrm{x} > 0$ ise $\mathrm{y} = 1$<br>$\mathrm{x} < 0$ ise $\mathrm{y} = 0$ |
| Sigmoid Fonksiyonu | $\sigma(x) = \frac{1}{1+e^{-1}}$ |
| Tanh Fonksiyonu | $tanh = 2 * \sigma(2 * x) - 1$ |
| ReLu Fonksiyonu | $f(x) = \max(0, x)$ |

Step fonksiyonun türevi öğrenme değeri temsil etmediği için aktivasyon fonksiyonlarında pek tercih edilmemektedir. Step fonksiyonun ikili bir sınıflayıcı olduğundan çıkış katmanında tercih edilebilmektedir. Doğrusal fonksiyon ise türevinde hep aynı değeri verdiği için aktivasyon fonksiyonu olarak kullanılması uygun olmayabilir. Sigmoid aktivasyon fonksiyonu Şekil 2.18'de gösterildiği gibi 0 ile 1 arasında bir değer oluşturur. x değerinde yapılan küçük bir değişiklik y değerinde büyük bir etkiye yol açar. Bu özelliği nedeniyle sınıflandırma problemleri için uygun bir yöntemdir.




**Şekil 2.18** Sigmoid fonksiyon ve türevi

Sigmoid fonksiyonunda türevin kolay alınabilir olması ve hızlı çalışması nedeniyle ileri beslemeli ağların kullanıldığı YSA çalışmalarında tercih edilmektedir. Sigmoid fonksiyonu Denklem 2.22 ile elde edilir.

$$\sigma(x) = \frac{1}{1+e^{-1}} \tag{2.22}$$

Sigmoid fonksiyonun türevi denklem 2.23 ile elde edilir;

$$\frac{d}{dx}\sigma(x) = \frac{d}{dx} * \frac{1}{1+e^{-1}} = \sigma(x) * (1 - \sigma(x)) \tag{2.23}$$

Sigmoid dışında step, tanh, softmax and RELU gibi aktivasyon fonksiyonları da mevcuttur. Bu çalışmada türevi kolay alınabilir olması ve hızlı işlem yapmaya olanak sağlamasından dolayı sigmoid aktivasyon fonksiyonu tercih edildi. Ayrıca sigmoid tabanlı bir algılayıcının öğrenme aşaması lojistik regresyondakine benzerdir. Bu da YSA ve lojistik regresyon yöntemin karşılaştırılmasında yardımcı olacaktır.

### 2.5.1.3. Hata fonksiyonu

Eğiticili yapay sinir ağlarındaki eğitim algoritmaları beklenen ve elde edilen sonucun arasındaki tutarsızlığı ölçek için hata fonksiyonunu kullanır. R programlamada neuralnetwork kütüphanesi ile iki hata fonksiyonu kullanılabilir. Bunlarda biri Toplam karesel hata yöntemidir denklem 2.24 ile hesaplanır [110].

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.24}$$

Burada n değeri test veri setinde yer alan örnek sayısına denk gelmektedir. $y_i$ tahmin edilen değeri $\hat{y}_i$ ise olması gereken değeri belirtmektedir.

### 2.5.2. Regresyon analizi

İki veya daha fazla değişken arasındaki ilişkiyi modelleyen yönteme regresyon analizi denilmektedir. Regresyon analizinin pek çok çeşidi bulunmaktadır. Bunlardan yapılan literatür incelemesinde en sık karşılaşılanları lojistik regresyon, doğrusal regresyon, çoklu doğrusal regresyon, sıralı regresyon ve çok kategorili regresyon yöntemidir.

Bu yöntemler haricinde regresyonun birçok çeşidi mevcuttur;

- Logaritmik regresyon
- Karesel regresyon
- Üstel regresyon
- Kübik regresyon
- Panel regresyon

Doğrusal regresyon ve çoklu doğrusal regresyon yöntemlerinde bağımlı değişken sayısal veri türündedir. Kategorik veri türünde bağımlı değişkene sahip veri setleri üzerinde lojistik regresyon uygulanması veri ile uyumlu olacaktır.

Eldeki probleme göre en uygun regresyon modelinin seçilmesi sınıflandırmanın başarısını arttırabilir. OECD sağlık veri setinde bağımlı değişken "Yüksek" ve "Çok yüksek" olmak üzere iki kategori değerinden oluşuyor. OECD sağlık veri seti kategorik bağımlı değişkene sahip olduğundan üzerinden sınıflandırma yapmak için en uygun modelin ikili lojistik regresyon olduğuna karar verildi.

#### 2.5.2.1. İkili lojistik regresyon

Lojistik regresyon, kategorik bir bağımlı bir yanıt değişkeni ile bir veya daha fazla girdi değişkeni ile ilişkilendiren bir istatistiksel modelleme tekniğidir. Bağımlı değişkenin kategorik olduğu durumlarda lojistik regresyon kullanılır. Lojiktik regresyonda kullanılacak bağımlı değişken sıralı, ikili veya çok kategorili olabilir. Bağımlı değişkenin iki değer aldığı ikili lojistik regresyon(binary logistic regression), bağımlı değişkenin ikiden fazla değer aldığı lojistik regresyona çok kategorili lojistik regresyon

(multinominal logistic regression) denir [113]. Bağımlı değişkenin üç veya daha fazla kategoriden oluştuğu durumlarda sıralı lojistik regresyon kullanılır. Lojistik regresyon sosyal bilimler, tıp ve imalat başta olmak üzere yaygın bir kullanım alanına sahiptir [108, 114]. Lojistik regresyonda doğrusal regresyondan farklı olarak tahmin edici ve yanıt değişkeni arasındaki ilişki doğrusal değildir.

Lojistik regresyon modelinin temeli bir olayın başarılı olma ve başarısız olma olasılığının karşılaştırıldığı odds oranına dayanmaktadır.

**Odds Oranı:** Lojistik regresyon başarısızlık olasılığı üzerinden başarı olasılığını hesaplar. Elde edilen sonuçlar odds oranı olarak gösterilir [108]. Odd oranı bir olayın meydana gelmeme olasılığı üzerinden meydana gelme olasılığını denklem 2.25a'daki gibi hesaplar. Odds başarı oranını başarsızlık oranına oranlar. Bu denklemde p başarı olasılığını ve 1-p'de başarısızlık olasılığını ifade etmektedir.

$$Odds = \frac{p}{1-p} \tag{2.25a}$$

$$Odds\ oranı = e^{\beta_i} \tag{2.25b}$$

Odds oranı ve regresyon katsayı arasındaki ilişki denklem 2.25b ile ifade edilir [115]. Lojit kavramı lojistik regresyonda kullanılan temel kavramlardan biridir. Odds oranının doğal logaritmasını ifade etmektedir.

$$\text{logit}\big(\pi(x)\big) = \log\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_x X_x \tag{2.26}$$

Lojistik regresyon Denklem 2.26'daki gibi hesaplanır. Bu hesaplamada $p$ sonuç olasılığını, $\beta_0$ durdurma terimini, $\beta_0, \ldots, \beta_x$ her değişkenle ilişkili katsayıları, $X_1, \ldots, X_x$ potansiyel tahmin edici değişkenlerin değerlerini ve x değişken sayısını gösterir [108].

Bu çalışma kapsamında yapılacak analizde bir kategorik değişken ve birden fazla bağımsız değişken arasında ilişki ikili lojistik regresyon analizi yöntemi ile incelenecektir. Modelin oluşturulması ve analizlerin yapılması için R programlama dili kullanılacak. Modeli oluşturmak için R programlama dilinde yer alan glm() lojistik regresyon fonksiyonu kullanılacaktır. Tahmin için ise predict() fonksiyonundan yararlanılacaktır.

### 2.5.3. Naive bayes sınıflandırıcı

Naive Bayes algoritmasının temeli Bayes teoremine dayanır. Thomas Bayes'in teoremine dayanan Bayes sınıflandırıcılar istatistiksel sınıflandırıcılardır ve her bir özniteliğin sonuç üzerindeki etkisini olasılıksal olarak hesaplar [55]. Veri madenciliği çalışmalarında sıkça kullanılan Naive bayes sınıflandırıcı metin madenciliği, ses madenciliği, çizge(graph) madenciliği ve web madenciliği gibi alanlarda sınıflandırma işlemleri için terch edilmektedir [117].

Bayes sınıflandırıcılarda amaç belli bir eğitim kümesine göre her sınıfa ilişkin üyeliklerin olasılıksal olarak tahmin edilmesidir. Burada kullanılan olasılık yöntemi koşullu olasılık yöntemine denk gelmektedir. Y sonuç değişkeni X öznitelik olmak üzere, X'in gerçekleştiği durumda Y'nin de gerçekleşme olasılığı P(Y|X) Denklem 2.27 ile ifade edilir [118].

$$P(X,Y) = P(Y|X)*P(X) = P(X|Y)*P(Y) \tag{2.27}$$

P(X|Y); Y olası durumunda X durumunun meydana gelme olasılığını, P(Y |A) X olası durumunda Y durumunun meydana gelme olasılığını temsil temektedir. P(X) ve P(Y) ifadeleri X ve Y olaylarının ön olasılığıdır. Denklem 2.28'deki ifadenin yeniden düzenlenmesi ile Bayes teoremi formüle edilir.

$$P(Y|X) = \frac{P(X|Y)*P(Y)}{P(X)} \tag{2.28}$$

Naive Bayes algoritması bir örnek için her durumun olasılığını hesaplar ve olasılık değerinin en yüksek olduğu duruma göre sınıflandırır. Az eğitim verisi ile başarılı sonuçlar vermesi nedeniyle küçük ve orta boyutlu veri setlerini sınıflandırmada tercih edilmektedir. Naive bayes algoritması, test kümesinde olan bir değerin eğitim kümesinde olmayan bir değere karşılık gelmesi durumunda olasılık değerini tahmin edemez yani olasılık değerini sıfır olarak verir. Sorunu düzeltilebilmesi için Laplace tahminine başvurulabilir. Sınıflandırma başarısının tespiti için karmaşıklık matrisi (confusion matrix) kullanılır.

### 2.5.4. Karmaşıklık matrisi

Yapılan sınıflandırmanın performansını ölçmek için kullanılan karmaşıklık matrisi test verisi ile tahmin değerlerini karşılaştırmamızı sağlar. Sınıflandırıcı birçok örneklemi doğru değerlendirirken, bazı örneklemleri doğruyken yanlış, bazı örneklemleri yanlışken doğru, işaretleyebilir. Bu dört durumun oluşmasına sebep verir [119]. Bu durumlar;

- Doğru Pozitif (DP veya True Positive - TP): IGE değeri Çok Yüksek ve Çok Yüksek tahmin edildi.
- Doğru Negatif (DN veya True Negative –TN): IGE değeri Yüksek ve Yüksek tahmin edildi.
- Yanlış Pozitif (YP veya False Positive – FP): IGE değeri Çok Yüksek ve Yüksek tahmin edildi.
- Yanlış Negatifler (YN veya False Negative – FN): IGE değeri Yüksek ve Çok Yüksek tahmin edildi.

Bu dört değere göre karmaşıklık matrisi Tablo 2.5'teki gibi oluşturulur.

**Tablo 2.5** Karmaşıklık matrisi değerleri

| Gerçek Değer | Tahmin = Çok Yüksek | Tahmin = Yüksek |
|---|---|---|
| Çok Yüksek | Doğru Pozitif (DP) | Yanlış Pozitif (YP) |
| Yüksek | Yanlış Negatif (YN) | Doğru Negatif (DN) |

Sınıflandırıcının yaptığı sınıflandırma değerlendirilirken kullanılan parametreler şunlardır [119];

**Doğruluk(Accuracy):** Doğru tahmin edilen DP ve DN değerlerinin tüm değerler toplamına bölünmesi ile elde edilen orandır. Denklem 2.29'da doğruluk hesabı gösterilmiştir.

$$\text{Doğruluk (Accuracy)} = \frac{DN+DP}{DN+DP+YN+YP} \tag{2.29}$$

**Hata oranı(Misclassification Rate):** Yanlış sınıflandırma oranını gösteren hata oranı sınıflayıcının ne sıklıkla yanlış tahmin ettiğinin bir ölçüdür ve Denklem 2.30'daki gibi hesaplanır.

$$\text{Hata oranı} = \frac{YN+YP}{DP+DN+YN+YP} \tag{2.30}$$

**Duyarlılık(Sensitivity, Recall veya TPR-Tue Positive Rate):** Doğru tahmin edilen doğru değer sayısı olan DP değerlerinin gerçek pozitiflerin (DP ve YN) toplamına bölümü ile elde edilen oran sınıflamanın duyarlılığıdır. Denklem 3.31'deki eşiklik yardımıyla duyarlılık hesabı yapılır.

$$\text{Duyarlılık (Recall, TPR)} = \frac{DP}{DP+YN} \tag{2.31}$$

**Özgüllük(Specifity):** Sınıflayıcının negatif değeri doğru tahmin ettiğinin bir ölçüsüdür. Denklem 2.32 özgüllük hesabında kullanılmaktadır.

$$\text{Özgüllük (Specifity)} = \frac{DN}{DN+YP} \tag{2.32}$$

**Hassasiyet(Precision):** Doğru tahmin edilen doğru değer sayısı olan DP değerlerinin tüm pozitif değerlere bölümü ile elde edilen oran sınıflamanın hassasiyetidir ve Denklem 2.33'teki gibi hesaplanır.

$$\text{Hassasiyet (Precission)} = \frac{DP}{DP+YP} \tag{2.33}$$

**F1 Skor:** F-measure olarak da bilinen hassasiyet ve duyarlılık değerlerinin harmonik ortalaması olan F1 skor uç durumların gözden kaçırılmaması için kullanılır. Eşit dağılmayan sınıf değerleri için doğruluk yerine kullanılan istatistiki bir hesaplamadır. F1 Skoru hesaplamak için denklem 2.34'teki hesaplama yöntemi kullanılır.

$$\text{F1 Skor} = \frac{2*Duyarlılık*Hassasiyet}{Duyarlılık+Hassiyet} \tag{2.34}$$

R programlama dilinde karmaşıklık matrisi hesaplamak için caret kütüphanesi içerisinde yer alan confusionMatrix() fonksiyonu kullanılabilir [120]. Bu fonksiyon sonuç olarak bir karmaşıklık matrisi ve bu matristen elde edilen doğruluk, hassasiyet, duyarlılık, özgüllük, kappa gibi ölçütleri döndürür.

# 3. BULGULAR VE TARTIŞMA

Detayları materyal ve yöntem başlığı altında verilen 2011-2016 yıllarına ait 31 OECD ülkesinin sağlık belirleyicilerine ait verilerinin yer aldığı OECD sağlık veri seti kullanılarak veri önişleme, paralel koordinat sistemi oluşturma, kümeleme ve sınıflandırma işlemleri yapıldı. Bu çalışmalarda elde edilen sonuçlar aşağıdaki başlıklarda belirtilmiştir.

## 3.1. Paralel Koordinat Sistemi Bulguları

Bu çalışmada OECD ülkelerinin 2011 – 2016 yılları arasındaki sağlık verilerinin bulunduğu Tablo 3.1'de detayları verilen on değişken ve bir çıkış değişkenine ait paralel koordinat sistemini R programlama dili ile R Studio'da oluşturuldu. R programlama dili veri madenciliği uygulamalarında yaygın bir şekilde kullanılan dildir. Diğer veri madenciliği araçları ile karşılaştırıldığından kendi kodunuzu yazmanıza sağladığı esneklik nedeniyle diğer araçlara oranla daha çok tercih edilmektedir. R programlamada bulunan hrbrthemes, GGally ve viridis kütüphaneleri yardımıyla veri setine ait paralel koordinat sistemini oluşturuldu. ggparcoord içerisinde yer alana uniminmax ölçeklendirme yöntemi ile verilere min-max normalizasyonu uygulandı. Paralel koordinat sistemini oluşturmak için R programlamada kullanılan kodlar;

```
library(GGally)
library(viridis)
ggparcoord(data,
        columns = 1:10, groupColumn = 11, order = "anyClass",
        scale="uniminmax",
        showPoints = TRUE,
        title = "Paralel Koordinat Sistemi / OECD Saglik Verileri")
+ scale_color_manual(values=c("dodgerblue4","red")) +  theme_ipsum()+
  theme( plot.title = element_text(size=10) )
```

Tablo 3.1'de paralel koordinat sisteminde yatay ekseni oluşturacak değişkenler verilmiştir. Çıkış değişkenimiz ülkenin yer aldığı IGE sınıfıdır. IGE sınıfları Birleşmiş

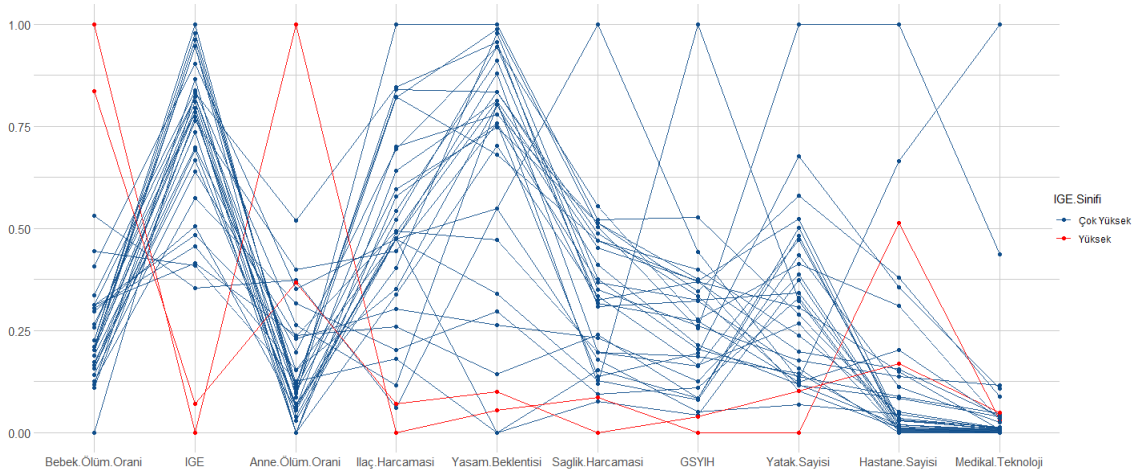
Milletler tarafından belirlenmektedir. Kullanılan veri seti içerisindeki ülkeler "Çok Yüksek" ve "Yüksek" olmak üzere iki sınıfa ayrıldı.

**Tablo 3.1** Paralel Koordinat Sisteminde Gösterilecek Değişkenler ve Sınıf Değişkeni

| Değişkenler | Değer |
|---|---|
| $X_1$ | Kişi başına GSYH |
| $X_2$ | Sağlık Harcamasının Kişi başına GSYH'ye Oranı |
| $X_3$ | 1000 Kişi Başına Düşen Yatak Sayısı |
| $X_4$ | Hastane Sayısı |
| $X_5$ | Bebek Ölüm Oranı |
| $X_6$ | Doğumda Yaşam Beklentisi |
| $X_7$ | Anne Ölüm Oranı |
| $X_8$ | Medikal Teknoloji |
| $X_9$ | Kişi Başına Düşen İlaç Harcaması (US $) |
| $X_{10}$ | IGE Değeri |
| Y | IGE sınıfı |

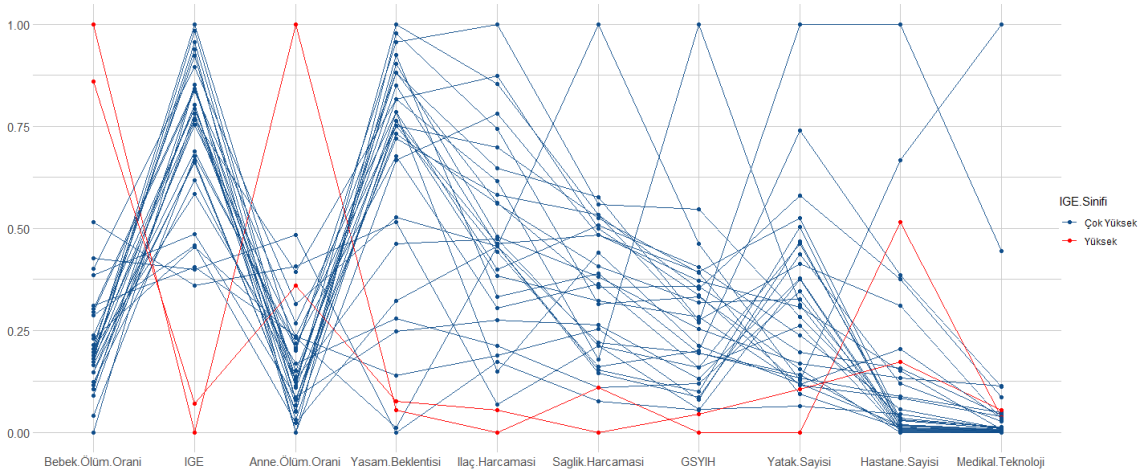
Veri seti içerisinde her bir yıl için 31 kayıt bulunmaktadır. 2011, 2012 ve 2013 yılı verilerinin her birinde veri setinin % 0,067'si "Yüksek" sınıfına %0,933'ü "Çok Yüksek" sınıfına aittir. 2015 ve 2016 yıllarında Türkiye "Çok Yüksek" sınıfına geçtiği için bu yılların her birinde veri setinin % 0,033'ü "Yüksek", 0,967'si "Çok Yüksek" olarak değişti.

Paralel Koordinat sisteminin anlaşılabilirliğini attırmak için ilk olarak veri seti üzerinde normalizasyon işlemi yapıldı. Bu çalışmada verileri çok çeşitli birimlerde olması nedeniyle Min-max normalizasyonu kullanılarak veri setlerindeki veriler 0-1 aralığına çekildi. Değişkenlerin sıralaması da grafiğin okunabilirliğini arttırmak için "AnyClass" olarak ayarlandı. Böylelikle seriler arası çapraz bağlantıları azaltılarak daha okunur bir grafik elde edildi.




**Şekil 3.1** 2011 Yılı Verilerine Ait Paralel Koordinat Sistemi

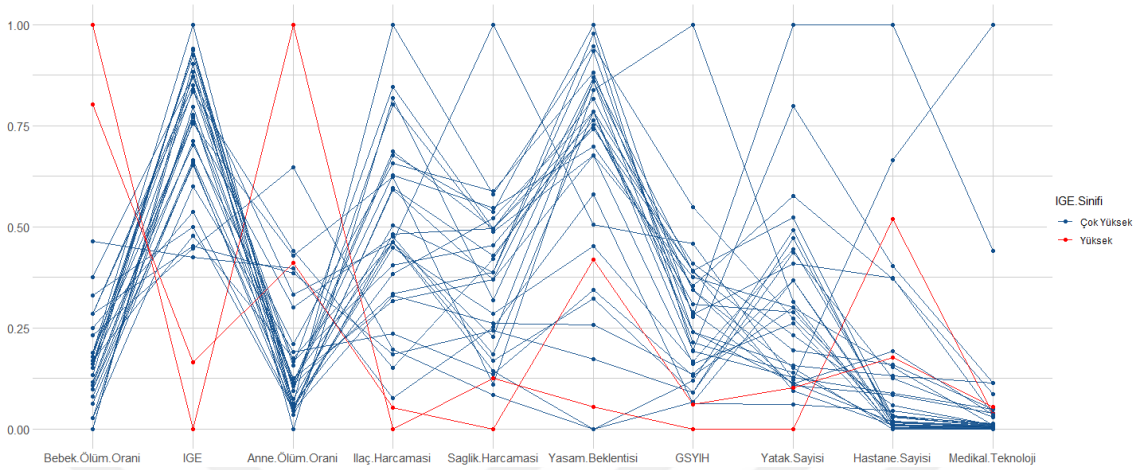
Sağlık veri seti içerisinde 2011 yılı verilerinde Türkiye ve Meksika "Yüksek" IGE sınıfındayken diğer 29 ülke "Çok Yüksek" IGE sınıfında yer aldı. 2011 yılı için Şekil 3.1'de oluşturulan paralel koordinat sisteminde grafik okunabilirliğini arttırmak için eksen sıralaması yapıldı. Bu sıralamada değişkenler sırayla bebek ölüm oranı, IGE değeri, anne ölüm oranı, doğumda yaşam beklentisi, ilaç harcaması, sağlık harcaması, GSYH, yatak sayısı, hastane sayısı ve medikal teknoloji olarak sıralandı. Bebek ölüm oranında ve IGE değerinde iki sınıf tamamen ayrıştı.




**Şekil 3.2** 2012 Yılı Verilerine Ait Paralel Koordinat Sistemi

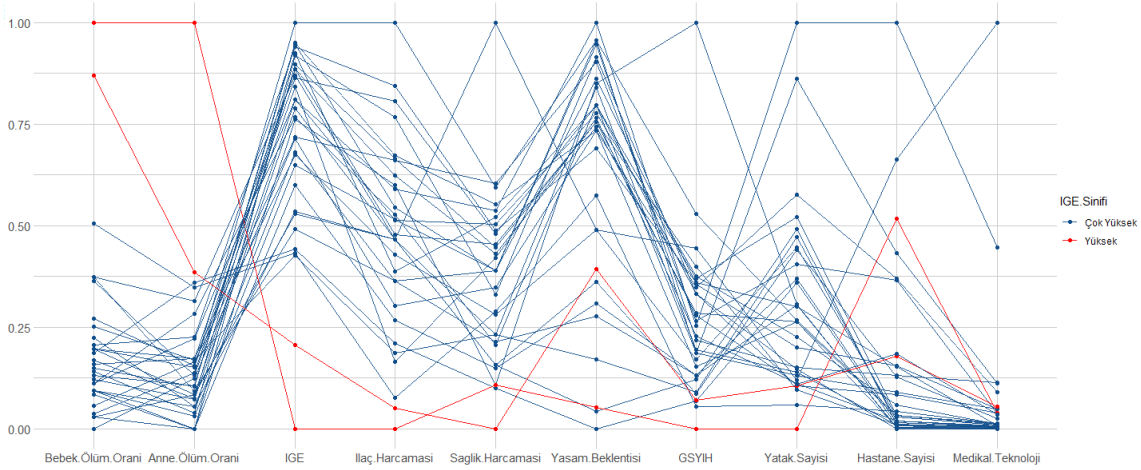
Sağlık veri seti içerisinde 2012 ve 2011 yılı verilerinde Türkiye ve Meksika "Yüksek" IGE sınıfındayken diğer 29 ülke "Çok Yüksek" IGE sınıfında yer aldı. 2012 yılı için Şekil 3.2'deki paralel koordinat sisteminde grafik okunabilirliğini arttırmak için eksen sıralaması yapıldı. 2013 yılı için Şekil 3.3'deki paralel koordinat sisteminde grafik okunabilirliğini arttırmak için eksen sıralaması yapıldı. 2011 yılına benzer şekilde 2012

ve 2013'de de IGE değeri ve bebek ölüm oranında iki sınıf birbirinden farklı aralıkta yer aldı.




**Şekil 3.3** 2013 Yılı Verilerine Ait Paralel Koordinat Sistemi

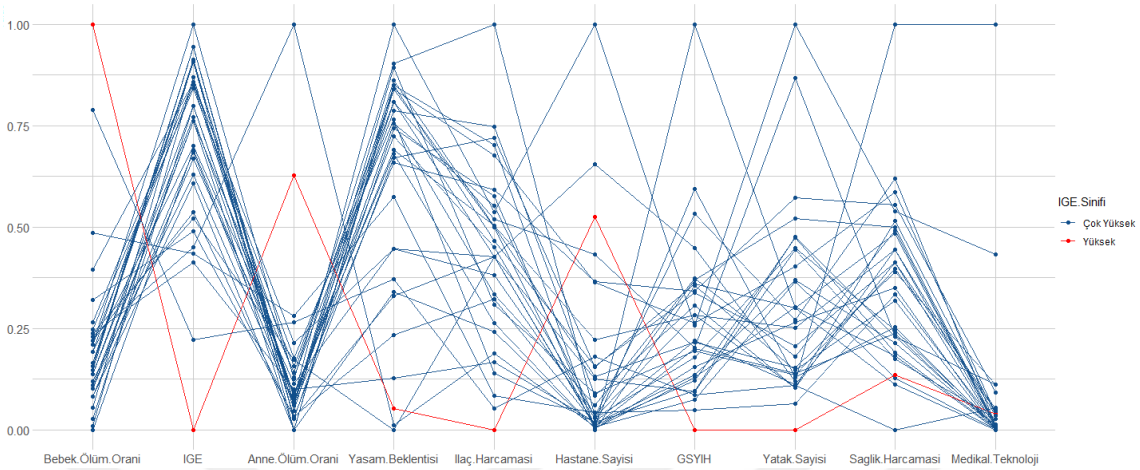
2014 yılına ait paralel koordinat sisteminin gösterildiği Şekil 3.4'te eksen sıralamasının değiştiği görüldü. Bu sıralamada değişkenler sırayla bebek ölüm oranı, anne ölüm oranı, IGE değeri, doğumda yaşam beklentisi, ilaç harcaması, sağlık harcaması, GSYH, yatak sayısı, hastane sayısı ve medikal teknoloji olarak sıralandı. Bebek ölüm oranı, IGE ve ilaç harcamasında her iki grup farklı aralıklarda değer aldı.



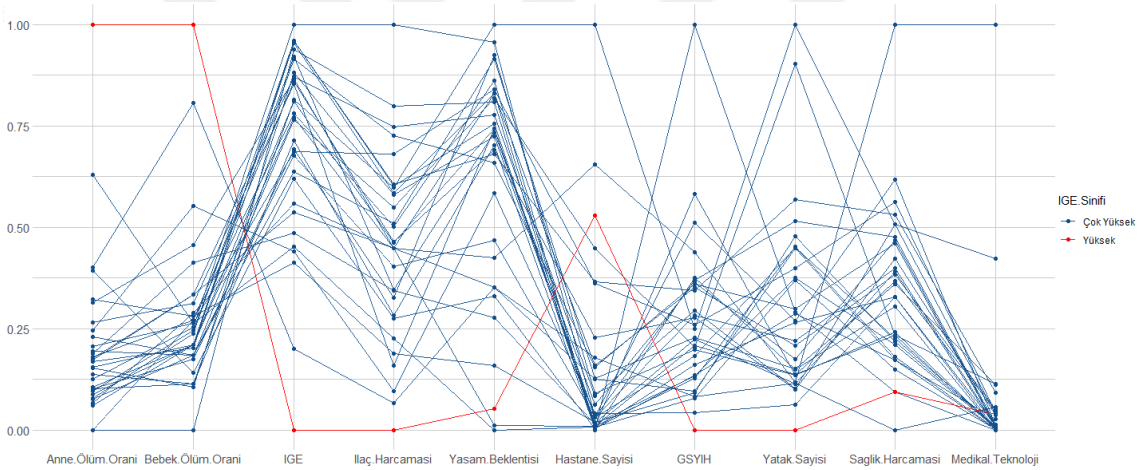

**Şekil 3.4** 2014 Yılı Verilerine Ait Paralel Koordinat Sistemi

2015 yılında Türkiye'nin "Çok yüksek" sınıfına geçmesi ile Meksika "Yüksek" sınıfında kalan tek ülke oldu. Şekil 3.5'de 2015 yılına ait koordinat sistemi gösterildi. Burada "Yüksek" sınıfında olan Meksika bebek ölüm oranı, IGE değeri, ilaç harcaması, GSYH

ve yatak sayısı bakımında en kötü değerlere sahip ülke olduğu koordinat sistemi üzerinde gösterildi.




**Şekil 3.5** 2015 Yılı Verilerine Ait Paralel Koordinat Sistemi




**Şekil 3.6** 2016 Yılı Verilerine Ait Paralel Koordinat Sistemi

2016 yılında da 2015 yılında olduğu gibi veri seti içerisinde yer alan tek "Yüksek" IGE sınıfına dahil olan ülke Meksika'dır. Şekil 3.6'daki paralel koordinat sisteminde de gösterildiği gibi anne ölüm oranı, bebek ölüm oranı, IGE değeri, ilaç harcaması, GSYH ve yatak sayısı bakımından en kötü ülke konumundadır.

## 3.2. Özellik Seçimi

### 3.2.1. Bilgi kazancı ile öznitelik seçimi bulguları

2011- 2016 yılları içeren veri setinde, her yıl için 31 tane kayıt bulunmaktadır. Sonuçlar elde edilen bilgi kazancı değerine göre sıralandı. 2011 yılına ait 31 kayda ait hesaplanan

bilgi kazancı sonrasında Tablo 3.2'deki sonuçlar elde edildi. Tabloda değişkenler için aşağıdaki kısaltmalar kullanıldı;

BÖO : Bebek Ölüm Oranı

AÖO : Anne Ölüm Oranı

HYS : Hastane Yatak Sayısı

2011 yılında bebek ölüm oranı ve GSYH özelliklerinin İnsani Gelişim Endeksine etkisinin diğer özelliklerden fazla olduğu belirlendi. Her iki özellik de 0.345 kazanç değerine sahip olduğu için her ikisinin de İnsani Gelişim Endeksine etkisi aynıdır.

Tablo 3.2'de yer alan 2012 yılına ait kişi başına düşen ilaç harcaması, bebek ölüm oranı ve gayrisafi milli hâsıla özelliklerinin insani gelişim üzerinde etkili olduğu belirlendi. Bir önceki yıla göre kişi başına düşen ilaç harcamasının IGE üzerindeki etkisinin 2012 yılında arttığı tespit edildi.

2013 yılına ait veriler incelendiğinde 2012 yılına ait verilerle birebir aynı sonuçlar elde edildi. 2013 yılında da kişi başına düşen ilaç harcamasının, bebek ölüm oranı ve gayrisafi milli hâsıla özellikleri 0,345 bilgi kazancı oranı ile İnsani Gelişim Endeksi üzerinde en etkili olan etmenler olarak gözlemlendi.

2014 yılında önceki yıllara göre bu özelliklerin sıralamasında değişiklik meydana geldi. Tablo 3.2'de gösterildiği gibi önceki yıllarda etkili olan bebek ölüm hızı ve kişi başına düşen ilaç harcaması etkisini devam ettirirken, GSYH yerine anne ölüm oranı ön plana çıktı. Bebek ölüm hızı, kişi başına düşen ilaç harcaması ve anne ölüm oranı 0,345 kazanç oranı ile IGE üzerinde eşit etkiye sahiptir. 2015 ve 2016 yıllarında oluşan özellik sıralaması aynı çıktı.

**Tablo 3.2** 2011 - 2012 yıllar bilgi kazancı değerleri

| | 2011 | | 2012 | | 2013 | | 2014 | | 2015 | | 2016 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sıra | Özellik | Değer | Özellik | Değer | Özellik | Değer | Özellik | Değer | Özellik | Değer | Özellik | Değer |
| 1 | GSYH | 0.345 | İlaç harcaması | 0.345 | İlaç harcaması | 0.345 | İlaç harcaması | 0.345 | İlaç harcaması | 0 | İlaç harcaması | 0 |
| 2 | BÖO | 0.345 | BÖO | 0.345 | BÖO | 0.345 | AÖO | 0.345 | Hastane sayısı | 0 | Hastane sayısı | 0 |
| 3 | AÖO | 0 | GSYH | 0.345 | GSYH | 0.345 | BÖO | 0.345 | Sağlık Harcaması | 0 | Sağlık Harcaması | 0 |
| 4 | HYS | 0 | HYS | 0 | HYS | 0 | Yaşam beklentisi | 0 | HYS | 0 | HYS | 0 |
| 5 | Sağlık Harcaması | 0 | Sağlık Harcaması | 0 | Sağlık Harcaması | 0 | Medikal Teknoloji | 0 | BÖO | 0 | BÖO | 0 |
| 6 | Hastane sayısı | 0 | AÖO | 0 | AÖO | 0 | Sağlık Harcaması | 0 | Medikal Teknoloji | 0 | Medikal Teknoloji | 0 |
| 7 | Yaşam beklentisi | 0 | Yaşam beklentisi | 0 | Yaşam beklentisi | 0 | HYS | 0 | Yaşam beklentisi | 0 | Yaşam beklentisi | 0 |
| 8 | Medikal Teknoloji | 0 | Medikal Teknoloji | 0 | Medikal Teknoloji | 0 | Hastane sayısı | 0 | AÖO | 0 | AÖO | 0 |
| 9 | İlaç harcaması | 0 | Hastane sayısı | 0 | Hastane sayısı | 0 | GSYH | 0 | GSYH | 0 | GSYH | 0 |

### 3.2.2. Korelasyon tabanlı öznitelik seçimi bulguları

Özellik ile seçilen çıktı değişkeninin arasındaki korelasyona yapılan bu sıralama işleminde çıkış değişkeni İnsani Gelişim Endeksi olarak belirlendi. Veri setinde bulunan dokuz niteliğin çıkış değişkeni ile arasındaki korelasyon hesaplanarak, bu korelasyon değerine göre sıralama yapıldı. Tablo 3.3'de 2011 yılında bu dokuz girdi değişkeninin İnsani Gelişim Endeksi'ne etkisinin korelasyon yöntemi ile belirlendiği sonuçlar gösterildi. İnsani gelişimde en büyük etkiyi bebek ölüm oranı özelliği göstermektedir. En az etkiyi ise medikal teknoloji yani sağlık araç-gereçleri gösterdi.

2012 yılına ait verilerin yer aldığı Tablo 3.3'te de IGE değeri üzerinde en büyük etkinin bebek ölüm oranının olduğu görüldü. Bir önceki yıldan farklı olarak doğumda yaşam beklentisi az bir farkla da olsa kişi başına düşen ilaç satışının önüne geçti. Diğer özelliklerin sıralamadaki yeri değişmedi ve medikal cihazlar 2012 yılında da en az etkiye sahip özellik oldu. 2013 yıl verilerinde İnsani Gelişim Endeksi ile güçlü bir korelasyona sahip olan özelliğin yine bebek ölüm oranı olduğu belirlendi. Bir önceki yıla ait veriler ile karşılaştırıldığında anne ölüm oranı ve kişi başına düzen ilaç harcamasının İnsani Yaşam Endeksi üzerinde doğumda yaşam beklentisinden daha etkili olduğu belirlendi. Önceki yıllarda da olduğu gibi en düşük etkiyi medikal teknoloji cihazlarının sayısı gösterdi. Bir önceki yıla göre büyük bir değişim yaşanmadığı 2014 yılında tüm özelliklerin çıkış değişkeni ile arasındaki ilişki sıralaması sabit kaldı.

2016 yılında da çıkış değişkeni ile 2011'den bu yana en yüksek ilişkiye sahip özellik olan bebek ölüm oranı ilk sıradaki yerini korudu. Medikal teknoloji makineleri 2011 -2016 yılları arasında çıkış değişkeni ile en az ilişkili özellik oldu. Hastane sayısı 2015 yılından önce IGE ile en az ilişkiye sahip olan değişken iken 2015 ve sonrasında diğer değişkenlere oranla çıkış değişkeni ile olan ilişkisi arttı.

2011 -2016 arasındaki tüm ilişkiler göz önüne alındığında bebek ölüm oranı, anne ölüm oranı ve kişi başı ilaç satışı İnsani Gelişim Endeksinin üzerinde en fazla etkiye sahip değişkenler olarak belirlendi. Bu yıllar arasında yaşanan medikal teknoloji cihazlarındaki değişim IGE' değeri üzerinde en az etkiye sahip değişken olduğu belirlendi.

**Tablo 3.3** 2011 - 2012 yılları korelasyon tabanlı özellik seçimi bulguları

| | 2011 | | 2012 | | 2013 | | 2014 | | 2015 | | 2016 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sıra | Özellik | Değer | Özellik | Değer | Özellik | Değer | Özellik | Değer | Özellik | Değer | Özellik | Değer |
| 1 | BÖO | 0.8458 | BÖO | 0.8461 | BÖO | 0.8561 | BÖO | 0.8549 | BÖO | 0.6888 | BÖO | 0.7511 |
| 2 | AÖO | 0.6527 | AÖO | 0.6651 | AÖO | 0.6154 | AÖO | 0.7458 | AÖO | 0.453 | AÖO | 0.6701 |
| 3 | İlaç Harcaması | 0.4632 | Yaşam beklentisi | 0.466 | İlaç Harcaması | 0.4761 | İlaç Harcaması | 0.4954 | Yaşam beklentisi | 0.3558 | İlaç Harcaması | 0.3677 |
| 4 | Yaşam beklentisi | 0.4617 | İlaç Harcaması | 0.4575 | Sağlık Harcaması | 0.3872 | Sağlık Harcaması | 0.4 | İlaç Harcaması | 0.3536 | Yaşam beklentisi | 0.3586 |
| 5 | Sağlık Harcaması | 0.3719 | Sağlık Harcaması | 0.383 | Yaşam beklentisi | 0.3534 | Yaşam beklentisi | 0.3726 | Hastane sayısı | 0.2971 | Hastane sayısı | 0.299 |
| 6 | GSYH | 0.3296 | GSYH | 0.3311 | GSYH | 0.3313 | GSYH | 0.3162 | GSYH | 0.2532 | GSYH | 0.2557 |
| 7 | HYS | 0.3209 | HYS | 0.3115 | HYS | 0.3011 | HYS | 0.2923 | HYS | 0.2467 | HYS | 0.2426 |
| 8 | Hastane sayısı | 0.2249 | Hastane sayısı | 0.2254 | Hastane sayısı | 0.2248 | Hastane sayısı | 0.2245 | Sağlık Harcaması | 0.2092 | Sağlık Harcaması | 0.2361 |
| 9 | Medikal Teknoloji | 0.0408 | Medikal Teknoloji | 0.0344 | Medikal Teknoloji | 0.0338 | Medikal Teknoloji | 0.0335 | Medikal Teknoloji | 0.0309 | Medikal Teknoloji | 0.0306 |

### 3.3. Kümeleme Çalışması Bulguları

Bu başlık altında OECD ülkeleri sağlık göstergelerinin kümeleme analizi sonuçlarına göre sınıflandırılması hedeflenmektedir. Elde edilen verilere göre OECD içerisinde ülkelerin OECD içerisinde hangi ülkelerle benzerlik gösterdiği belirlenecek. Kümeleme işleminde kullanılacak değişkenler Tablo 3.4'te verilmiştir. $X_1,\ldots,X_3$ kümeleme için kullanılacak belirleyici değişkenleri göstermektedir. Veri setinde yer alan IGE değeri ve IGE sınıfı kümeleme çalışmasında sonuca doğrudan etki ettiği için kullanılacak göstergeler arasına alınmadı.

**Tablo 3.4** Kümeleme işleminde kullanılacak değişkenler

| Değişkenler | Değer |
|---|---|
| X1 | Kişi başına GSYH (PPP $) |
| X2 | GSYH İçerisindeki Tüm Sağlık Harcamalarının Yüzdesi(%) |
| X3 | 1000 Kişi Başına Düşen Yatak Sayısı |
| X4 | Hastane Sayısı (1 000 kişi aşına) |
| X5 | Bebek Ölüm Oranı (1 000 canlı doğum başına) |
| X6 | Doğumda Yaşam Beklentisi |
| X7 | Anne Ölüm Oranı (100 000 canlı doğum başına) |
| X8 | Medikal Teknoloji (Makine Sayısı) |
| X9 | Kişi Başına Düşen İlaç Harcaması (ABD $) |

Çalışmada kullanılacak olan OECD sağlık veri setinde yer alan bu değişkenlere ait değerlerden 2011 yılına ait olan veriler Tablo 3.5'te sunuldu.

2016 yılına ait veriler de Tablo 3.6'da verilmiş olup 2012, 2013 ve 2014 yıllarına ait veriler EK 1 Tablo 1, EK 2 Tablo 2 ve EK 3 Tablo 3 olarak ekler kısmında verildi.

**Tablo 3.5** OECD ülkeleri 2011 yılı sağlık göstergeleri ve verileri

| Ülkeler | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 |
|---|---|---|---|---|---|---|---|---|---|
| Avustralya | 44419 | 8.5 | 3.79 | 1345 | 3.8 | 82 | 4.3 | 2324 | 473.8 |
| Avusturya | 44469 | 10 | 7.68 | 272 | 3.6 | 81.1 | 2.6 | 751 | 422.4 |
| Kanada | 41663 | 10.4 | 2.8 | 725 | 4.9 | 81.3 | 4.8 | 2207 | 699.8 |
| Şili | 20303 | 6.8 | 2.22 | 385 | 7.7 | 78.7 | 16.1 | 545 | 105.6 |
| Çekya | 28796 | 7 | 7.06 | 255 | 2.7 | 78 | 10.1 | 580 | 435.8 |
| Danimarka | 44408 | 10.2 | 5.08 | 1295 | 3.5 | 79.9 | 5.1 | 463 | 686.8 |
| Estonya | 24739 | 5.8 | 5.36 | 55 | 2.5 | 76.4 | 13.6 | 58 | 214.1 |
| Finlandiya | 40917 | 9 | 5.52 | 275 | 2.4 | 80.6 | 0 | 492 | 500.5 |
| Fransa | 37448 | 11.2 | 6.36 | 2681 | 3.3 | 82.3 | 8.4 | 1722 | 589.1 |
| Almanya | 42542 | 10.7 | 8.38 | 3278 | 3.6 | 80.5 | 4.7 | 5005 | 514.6 |
| Macaristan | 22894 | 7.5 | 7.19 | 173 | 4.9 | 75 | 10.2 | 402 | 257 |
| İzlanda | 40769 | 8.3 | 3.29 | 8 | 0.9 | 82.4 | 22.3 | 31 | 705.1 |
| İrlanda | 44870 | 10.7 | 2.62 | 98 | 3.5 | 80.8 | 2.7 | 274 | 593.9 |
| İsrail | 30551 | 7 | 3.14 | 87 | 3.5 | 81.7 | 1.2 | 206 | 421 |
| İtalya | 36183 | 8.8 | 3.52 | 1184 | 2.9 | 82.3 | 2.6 | 6582 | 456.8 |
| Japonya | 35775 | 10.6 | 13.4 | 8605 | 2.3 | 82.7 | 4.1 | 24966 | 684.9 |
| Güney Kore | 31228 | 6.3 | 9.53 | 3064 | 3 | 80.6 | 17.2 | 6150 | 398.5 |
| Letonya | 19798 | 5.6 | 5.88 | 70 | 6.6 | 73.7 | 5.4 | 145 | 197.3 |
| Litvanya | 22824 | 6.5 | 7.43 | 105 | 4.8 | 73.7 | 6.6 | 151 | 421 |
| Lüksemburg | 91814 | 6.1 | 5.28 | 13 | 4.3 | 81.1 | 0 | 41 | 549.2 |
| Meksika | 16547 | 5.7 | 1.41 | 4430 | 13.7 | 74.2 | 43 | 1835 | 59.9 |
| Hollanda | 46599 | 10.2 | 4.25 | 449 | 3.6 | 81.3 | 1.7 | 655 | 433.9 |
| Yeni Zelanda | 32667 | 9.5 | 2.82 | 164 | 5.2 | 81 | 11.3 | 291 | 147.2 |
| Polonya | 22576 | 6.2 | 6.63 | 968 | 4.7 | 76.8 | 2.3 | 1498 | 421 |
| Slovakya | 26051 | 7.4 | 6.05 | 140 | 4.9 | 76.1 | 9.9 | 294 | 290.6 |
| Slovenya | 28931 | 8.6 | 4.62 | 29 | 2.9 | 80.1 | 0 | 112 | 318.3 |
| İspanya | 31872 | 9.1 | 3.05 | 763 | 3.1 | 82.6 | 3 | 2692 | 367.5 |
| İsviçre | 56184 | 10.8 | 4.87 | 300 | 3.8 | 82.8 | 3.7 | 685 | 822 |
| Türkiye | 19445 | 4.7 | 2.62 | 1453 | 11.6 | 74.6 | 15.8 | 2779 | 114 |
| İngiltere | 37146 | 8.4 | 2.88 | 1747 | 4.2 | 81 | 6.6 | 1986 | 328.8 |
| ABD | 49811 | 16.4 | 2.97 | 5724 | 6.1 | 78.7 | 15.1 | 57084 | 421 |

**Tablo 3.6** OECD ülkeleri 2016 yılı sağlık göstergeleri ve verileri

| Ülkeler | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 |
|---|---|---|---|---|---|---|---|---|---|
| Avustralya | 50263 | 9.2 | 3.84 | 1352 | 3.1 | 82.5 | 3.9 | 3230 | 469.8 |
| Avusturya | 51791 | 10.4 | 7.42 | 273 | 3.1 | 81.7 | 5.7 | 808 | 380.6 |
| Kanada | 45109 | 10.8 | 2.58 | 722 | 4.5 | 82 | 6.3 | 2414 | 589.5 |
| Şili | 22788 | 8.5 | 2.12 | 356 | 7 | 80.2 | 9 | 774 | 113 |
| Çekya | 35234 | 7.2 | 6.66 | 260 | 2.8 | 79.1 | 7.1 | 593 | 337.9 |
| Danimarka | 50869 | 10.2 | 4.9 | 1322 | 3.1 | 80.9 | 3.2 | 529 | 572.8 |
| Estonya | 30911 | 6.5 | 4.76 | 30 | 2.3 | 77.8 | 14.4 | 69 | 243.7 |
| Finlandiya | 44016 | 9.4 | 3.97 | 262 | 1.9 | 81.5 | 5.6 | 559 | 466.3 |
| Fransa | 42067 | 11.5 | 6.06 | 3065 | 3.7 | 82.6 | 7.6 | 2636 | 541 |
| Almanya | 49516 | 11.1 | 8.06 | 3100 | 3.4 | 81.1 | 2.9 | 5736 | 485.2 |
| Macaristan | 27171 | 7.1 | 7 | 168 | 3.9 | 76.2 | 11.8 | 437 | 180.6 |
| İzlanda | 52340 | 8.2 | 3.13 | 8 | 0.7 | 82.3 | 0 | 32 | 626.4 |
| İrlanda | 70616 | 7.4 | 2.97 | 86 | 3 | 81.8 | 6.2 | 300 | 482.9 |
| İsrail | 37475 | 7.3 | 2.99 | 84 | 3.1 | 82.5 | 2.2 | 251 | 371 |
| İtalya | 39178 | 8.9 | 3.17 | 1090 | 2.8 | 83.3 | 2.8 | 6971 | 415.6 |
| Japonya | 41138 | 10.8 | 13.11 | 8442 | 2 | 84.1 | 3.7 | 26454 | 479.7 |
| Güney Kore | 37143 | 7.3 | 11.98 | 3788 | 2.8 | 82.4 | 8.4 | 7132 | 444 |
| Letonya | 25879 | 6.2 | 5.72 | 65 | 3.7 | 74.7 | 23.1 | 163 | 208.4 |
| Litvanya | 30334 | 6.6 | 6.69 | 93 | 4.5 | 74.8 | 6.5 | 177 | 371 |
| Lüksemburg | 107775 | 5.5 | 4.81 | 12 | 3.8 | 82.8 | 0 | 37 | 409.3 |
| Meksika | 18969 | 5.5 | 1.39 | 4474 | 12.1 | 75.2 | 36.7 | 2467 | 43.3 |
| Hollanda | 51340 | 10.3 | 3.44 | 534 | 3.5 | 81.6 | 3.5 | 643 | 295.5 |
| Yeni Zelanda | 38784 | 9.3 | 2.73 | 159 | 4.27 | 81.7 | 9.7 | 330 | 159 |
| Polonya | 27406 | 6.5 | 6.64 | 1064 | 4 | 78 | 2.4 | 1678 | 371 |
| Slovakya | 30825 | 7 | 5.78 | 133 | 5.4 | 77.3 | 6.9 | 341 | 294.5 |
| Slovenya | 33198 | 8.5 | 4.49 | 29 | 2 | 81.3 | 5 | 118 | 250 |
| İspanya | 36581 | 9 | 2.97 | 764 | 2.7 | 83.4 | 3.7 | 2967 | 281.6 |
| İsviçre | 64324 | 12.2 | 4.55 | 283 | 3.6 | 83.7 | 4.6 | 734 | 773.2 |
| Türkiye | 26330 | 4.3 | 2.75 | 1510 | 9.9 | 78 | 14.7 | 3545 | 91.4 |
| İngiltere | 43509 | 9.7 | 2.57 | 1922 | 3.8 | 81.2 | 6.6 | 2317 | 381.8 |
| ABD | 57884 | 17.1 | 2.77 | 5534 | 5.9 | 78.7 | 11.51 | 62435 | 371 |

Detayları materyal ve yöntem başlığı altında verilen OECD sağlık veri seti üzerinde min-max normalizasyonu işlemi yapıldı. Daha sonra 0 - 1 aralığına çekilen bu dokuz özniteliğe ait 31 örnek Öklid uzaklık yöntemi kullanılarak örnek noktalar arasındaki uzaklıklar belirlendi.

Bu işlem 2011-2016 yılları arasındaki tüm veri setler için gerçekleştirildi. Uzaklıkları belirlenen noktaları bir kümeye atamadan önce küme sayısının belirlenmesi gerekmektedir. Burada ortalama Silhouette genişliğinden yararlanılarak. Küme sayısı belirlendi.

### 3.3.1. Hiyerarşik kümeleme

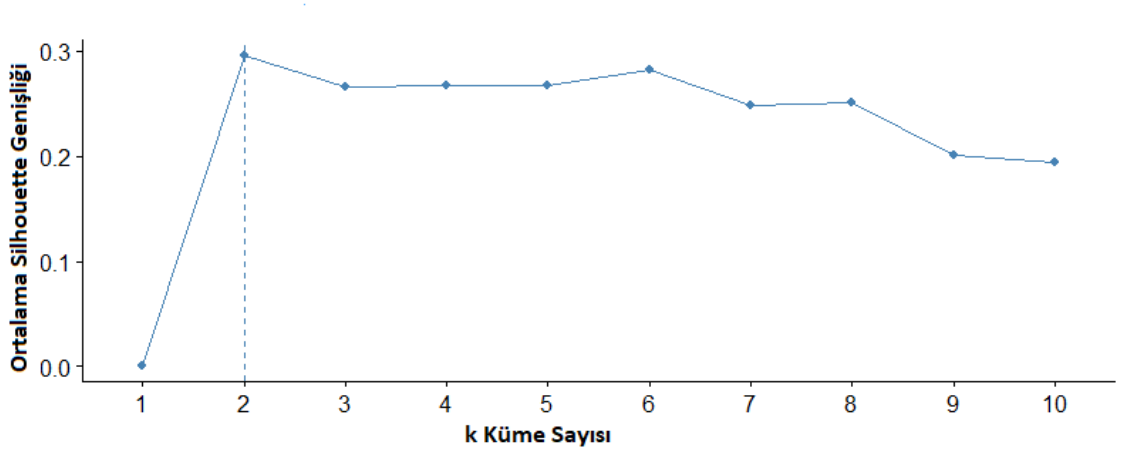
2011 ve 2016 yılları arasında OECD ülkelerinin kümeleme sonuçlarında yaşanan değişimleri görebilmek için ayrıştırıcı hiyerarşik kümeleme ile 2011 ve 2016 yıllarında ülkelerin hangi kümeye dahil olduğu belirlendi. Bunun için ilk olarak optimum küme sayısı belirlendi. Küme sayısını belirlemek için Ortalama Silhouette Genişliği değerleri kullanıldı.

R programlamada her bir bağlantı türü için küme sayısı iki, üç dört ve beş olan tüm kümelerin Ortalama Silhouette Genişliği değeri hesaplandı(Tablo 3.7). Hiyerarşik kümelemede tüm bağlantı yöntemleri için en uygun küme sayısının iki olduğu belirlendi.

**Tablo 3.7** 2011 Yılı Verilerinin Ortalama Silhoutte Genişliği Değerleri

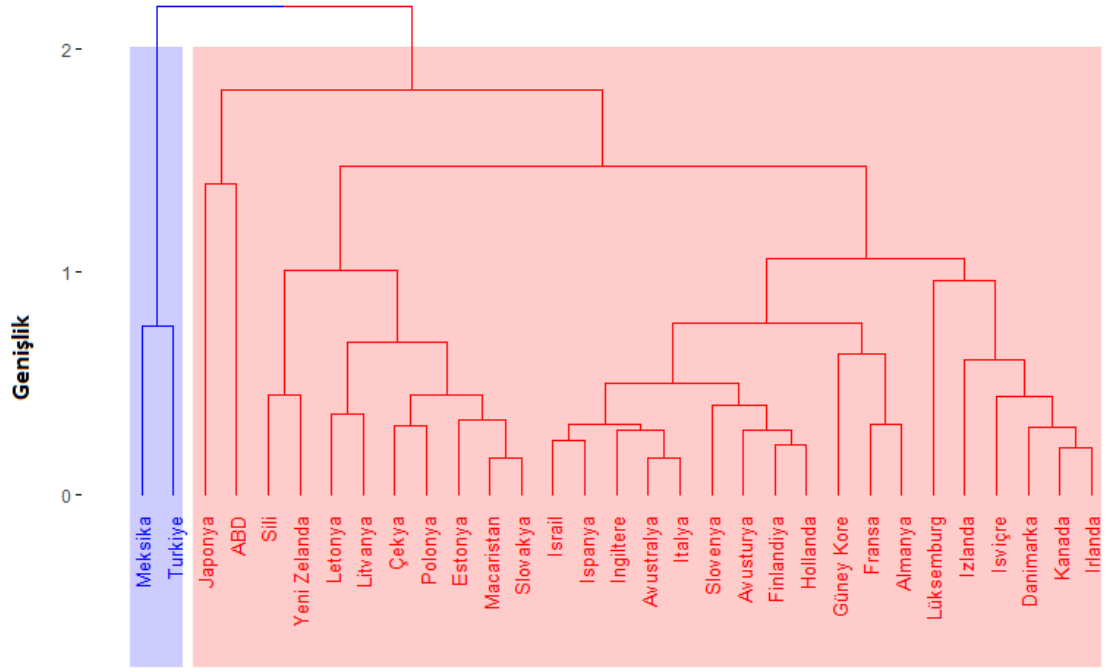
| Bağlantı Türü | Küme sayısı = 2 | Küme sayısı = 3 | Küme sayısı = 4 | Küme sayısı = 5 |
|---|---|---|---|---|
| Tam Bağlantı | 0.42 | 0.4 | 0.33 | 0.33 |
| Ortalama Bağlantı | 0.44 | 0.41 | 0.38 | 0.33 |
| Tek Bağlantı | 0.43 | 0.41 | 0.38 | 0.21 |
| Ward Yöntemi | 0.33 | 0.3 | 0.31 | 0.24 |

Bu işlemin haricinde R programlama dilinde bulunan fviz_nbclust fonksiyonu ile veri setinin Ortalama Silhouette Genişliği incelendiğinde Şekil 3.7'deki grafik elde edildi. Bu grafikde de en uygun k küme değerinin iki olduğu belirlendi.




**Şekil 3.7** 2011 Yılı Verilerine ait Ortalama Silhouette Genişliği

Hiyerarşik kümelemede sonuçların gösterildiği dendrogramda yatay eksen ülkeleri dikey eksen ise oluşan kümelerin birbirine olan yakınlığını belirtmektedir.




**Şekil 3.8** Tam Bağlantı Yöntemi ile Hiyerarşik Kümeleme (2011 Yılı)

K değerinin belirlenmesi ile tüm hiyerarşik kümeleme işlemleri için küme değeri iki olarak uygulandı. Tam bağlantı yönteminin kullanıldığı Şekil 3.8'deki hiyerarşik

kümeleme ağacına göre oluşan kümeler Tablo 38'de gösterildi. Küme-1'de iki ülke yer aldı, Küme-2'de 29 ülke yer aldı. Dikey eksen 0-2 aralığına kadar genişledi. Bu genişlik değeri arttıkça kümeler birbirinden uzaklaşmaktadır.

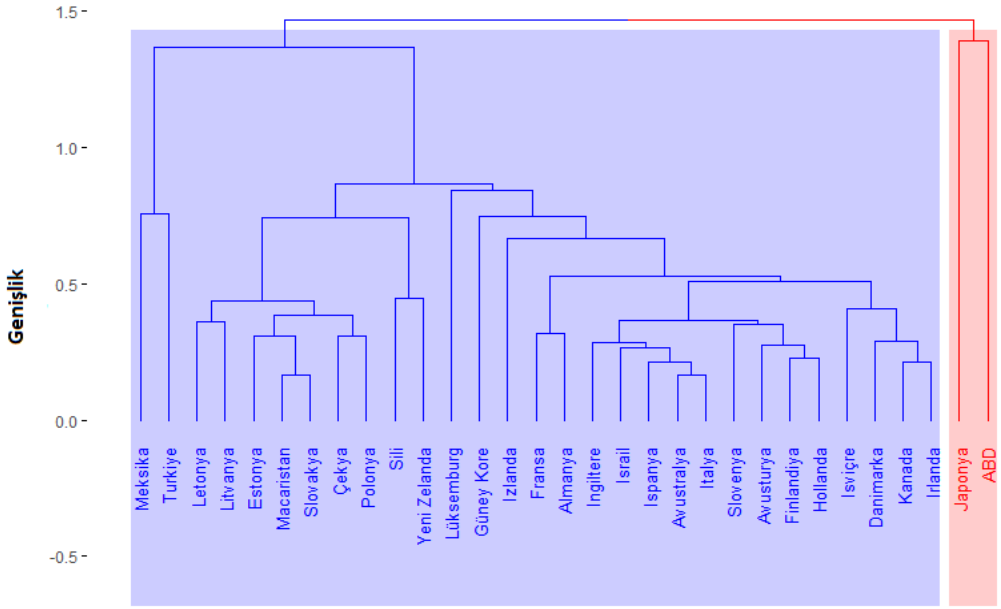
Tam bağlantı yöntemi kullanılarak yapılan kümeleme işleminin sonucu IGE değerlerinin gerçek kümeleri ile birebir aynı çıktı.

**Tablo 3.8** Tam bağlantı yöntemi ile hiyerarşik kümeleme sonucu oluşan kümeler

| Küme-1 | Küme-2 | |
|---|---|---|
| Türkiye | Japonya | İtalya |
| Meksika | ABD | Slovenya |
| | Şili | Avusturya |
| | Yeni Zelanda | Finlandiya |
| | Letonya | Hollanda |
| | Litvanya | Güney Kore |
| | Çekya | Fransa |
| | Polonya | Almanya |
| | Estonya | Lüksemburg |
| | Macaristan | İzlanda |
| | Slovakya | İsviçre |
| | İsrail | Danimarka |
| | İspanya | Kanada |
| | İngiltere | İrlanda |
| | Avustralya | |

2011 yılı OECD sağlık verileri üzerinde ortalama bağlantı yöntemi ile elde edilen hiyerarşik kümeleme sonuçları Şekil 3.9'da verildi. Küme-1 içerisinde 29 ülke Küme-2 içerisinde iki ülke bulunmaktadır.

Ortalama bağlantı yönteminin kullanıldığı hiyerarşik kümeleme ağacına göre oluşan kümeler Tablo 3.9'da gösterildi. Kümeleme işleminin sonucunda ABD ve Japonya diğer ülkelerden ayrıştılar. Dikey eksen 0 - 1.5 aralığına kadar genişledi. Ortalama bağlantı yöntemi kullanılarak elde edilen hiyerarşik kümeleme sonucunda oluşan iki kümenin arasındaki mesafenin tam bağlantı yöntemine göre daha az olduğu gözlemlendi.

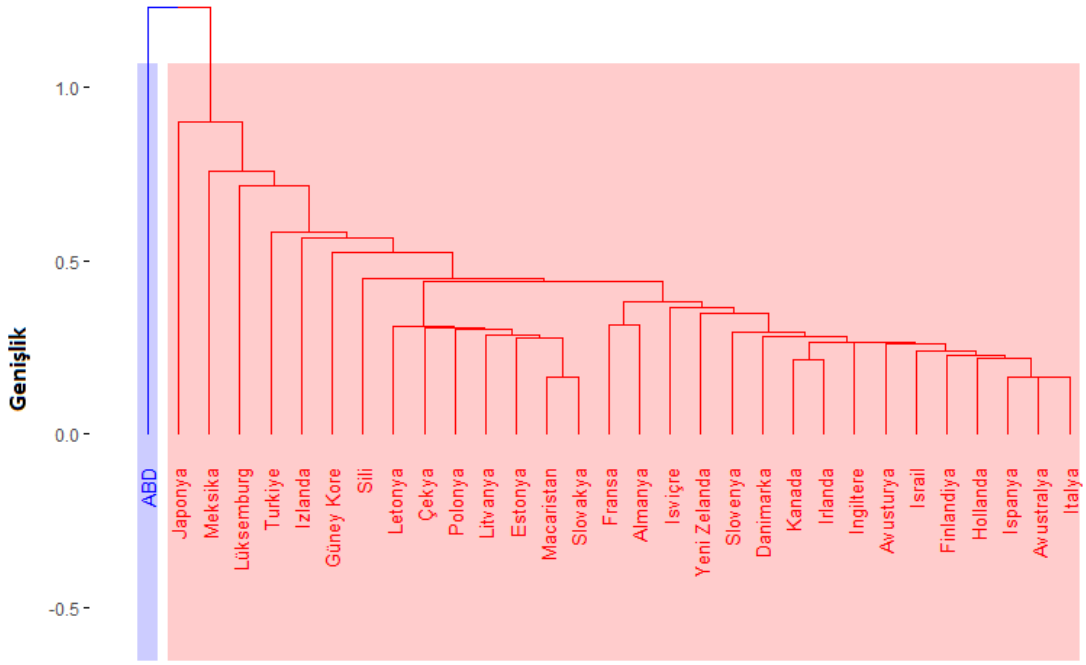



**Şekil 3.9** Ortalama Bağlantı Yöntemi ile Hiyerarşik Kümeleme (2011 Yılı)

**Tablo 3.9** Ortalama bağlantı yöntemi ile hiyerarşik kümeleme sonucu oluşan kümeler

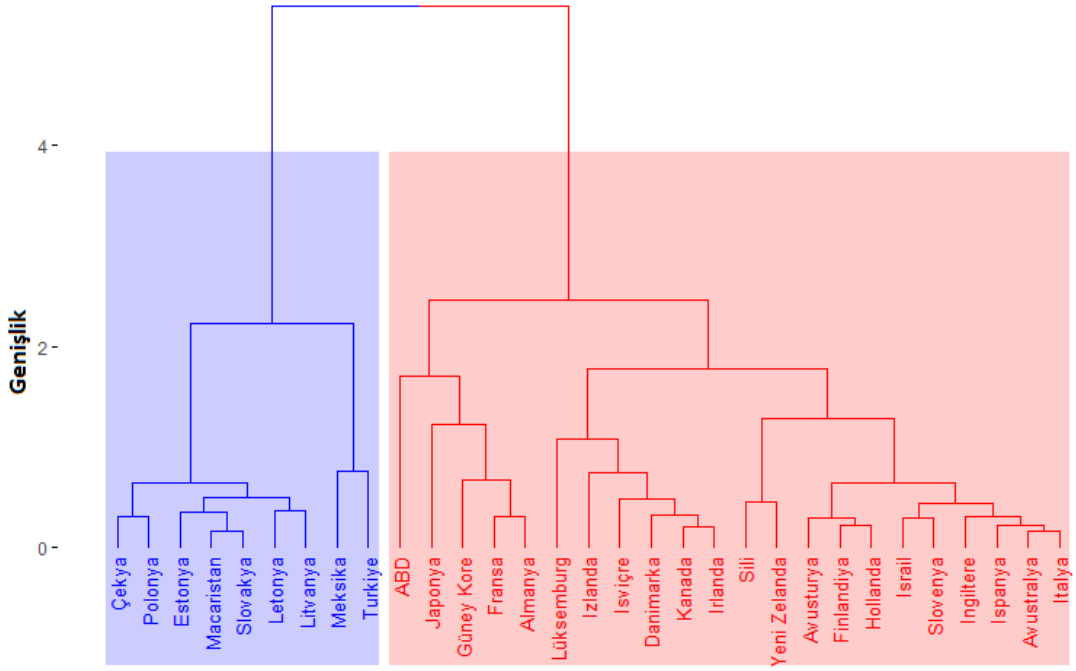
| Küme-1 | Küme-2 | |
|---|---|---|
| Japonya | Türkiye | İtalya |
| ABD | Meksika | Slovenya |
| | Şili | Avusturya |
| | Yeni Zelanda | Finlandiya |
| | Letonya | Hollanda |
| | Litvanya | Güney Kore |
| | Çekya | Fransa |
| | Polonya | Almanya |
| | Estonya | Lüksemburg |
| | Macaristan | İzlanda |
| | Slovakya | İsviçre |
| | İsrail | Danimarka |
| | İspanya | Kanada |
| | İngiltere | İrlanda |
| | Avustralya | |

Tek bağlantı yöntemi ile elde edilen hiyerarşik kümeleme sonuçları Şekil 3.10'da verildi. Küme-1 içerisinde 30 ülke Küme-2 içerisinde bir ülke bulunmaktadır. Tek bağlantı yönteminin kullanıldığı hiyerarşik kümeleme dendrogramına göre ABD diğer ülkelerden ayrılarak bir grup oluşturdu. Dikey eksen 0 – 1 aralığına kadar genişledi. Tek bağlantı yöntemi kullanılarak elde edilen hiyerarşik kümeleme sonucunda oluşan iki kümenin arasındaki mesafenin tam bağlantı ve ortalama bağlantı yöntemine göre daha az olduğu gözlemlendi. Bu da burada oluşan kümeler arasında farkın az olduğunuz gösteriyor.




**Şekil 3.10** Tek bağlantı yöntemi ile Hiyerarşik Kümeleme (2011 Yılı)

Ward bağlantı yönteminin kullanıldığı Şekil 3.11'deki dendrogramda genişlik 0-4 arasında olmuştur. Tek bağlantı, ortalama bağlantı ve tam bağlantı yöntemlerine kıyasla birbirinden daha uzak kümeler elde edildi. Küme-1 içerisinde birbirine en benzer ülkelerin Slovakya ve Macaristan, Küme-2 içerisinde birbirine en fazla benzeyen ülkelerin Avusturalya ve İtalya olduğu belirlendi.



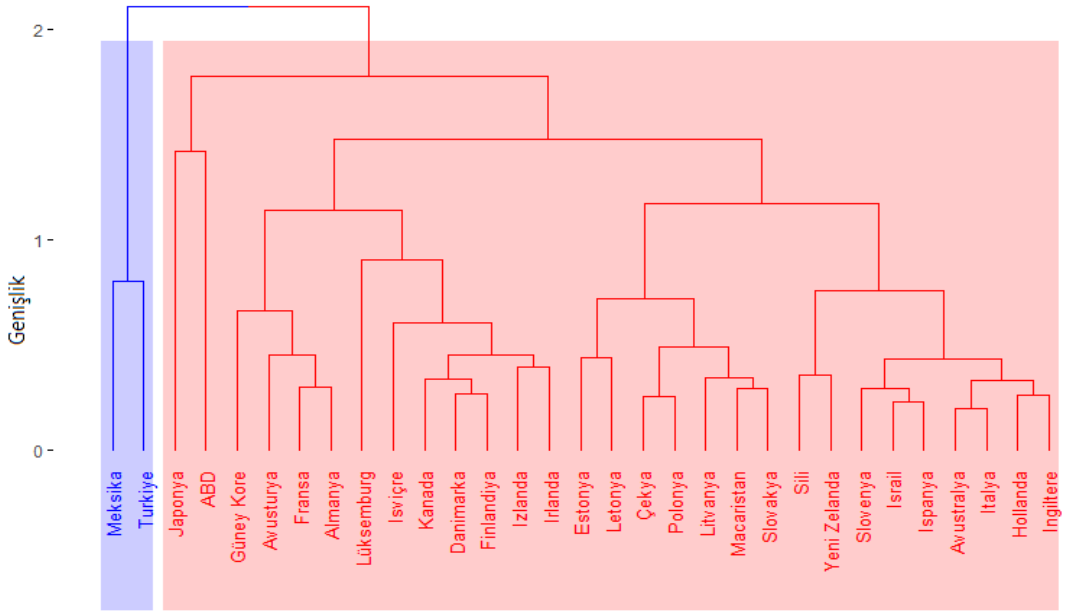

**Şekil 3.11** Ward bağlantı yöntemi ile Hiyerarşik Kümeleme (2011 Yılı)

Ward bağlantı yöntemi ile oluşan kümeler Tablo 3.10'de verildi. Oluşan iki kümeden dokuz tanesi Küme-1, geriye kalan 22 tanesi de Küme-2'ye dahil oldu.

**Tablo 3.10** Ward bağlantı yöntemi ile hiyerarşik kümeleme sonucu oluşan kümeler

| Küme-1 | Küme-2 | |
|---|---|---|
| Türkiye | Japonya | İtalya |
| Meksika | ABD | Slovenya |
| Letonya | Şili | Avusturya |
| Litvanya | Yeni Zelanda | Finlandiya |
| Çekya | İsrail | Hollanda |
| Polonya | İspanya | Güney Kore |
| Estonya | İngiltere | Fransa |
| Macaristan | Avustralya | Almanya |
| Slovakya | Danimarka | Lüksemburg |
| | Kanada | İzlanda |
| | İrlanda | İsviçre |

2016 yılı OECD sağlık verilerinin kümelenmesi için kullanılacak hiyerarşik kümelemede k küme değeri iki olarak uygulandı. Tam bağlantı yönteminin kullanıldığı 3.12'deki hiyerarşik kümeleme ağacına göre oluşan kümeler Tablo 3.11'te gösterildi.



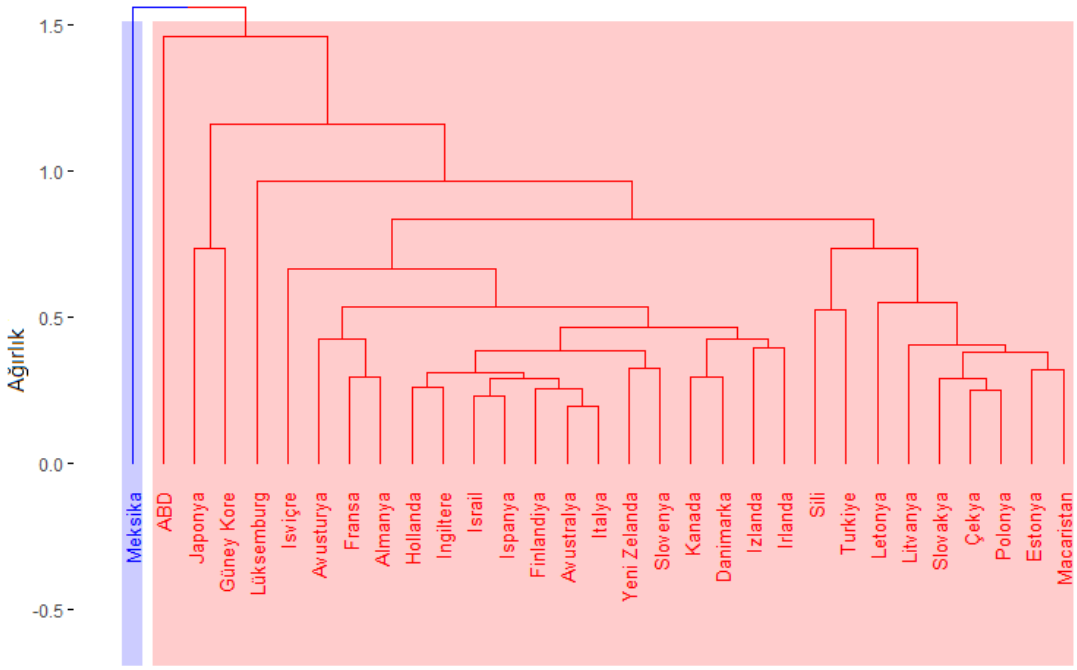

**Şekil 3.12** Tam Bağlantı Yöntemi ile Hiyerarşik Kümeleme (2016 Yılı)

Küme-1'de iki ülke yer aldı, Küme-2'de 29 ülke yer aldı. Dikey eksen 0-2 aralığına kadar genişledi. Bu genişlik değeri arttıkça kümeler birbirinden uzaklaşmaktadır.

**Tablo 3.11** Tam bağlantı yöntemi ile hiyerarşik kümeleme sonucu oluşan kümeler

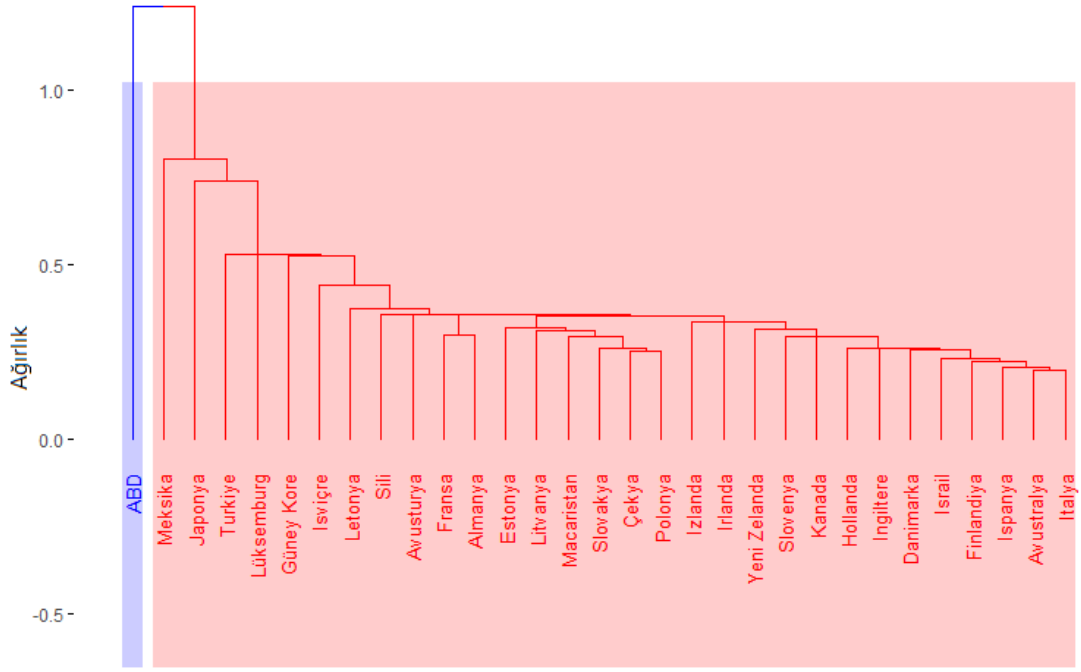
| Küme-1 | Küme-2 | | |
|---|---|---|---|
| Türkiye | Japonya | İtalya | İsrail |
| Meksika | ABD | Slovenya | İspanya |
| | Şili | Avusturya | İngiltere |
| | Yeni Zelanda | Finlandiya | Avustralya |
| | Letonya | Hollanda | Danimarka |
| | Litvanya | Güney Kore | Kanada |
| | Çekya | Fransa | İrlanda |
| | Polonya | Almanya | |
| | Estonya | Lüksemburg | |
| | Macaristan | İzlanda | |
| | Slovakya | İsviçre | |

Ortalama bağlantı yönteminin kullanıldığı hiyerarşik kümeleme dendrogramına göre oluşan kümeler Şekil 3.13'te gösterildi. Kümeleme işleminin sonucunda Meksika diğer ülkelerden ayrıştı. Dikey eksen 0 - 1.5 aralığına kadar genişledi. Ortalama bağlantı yöntemi kullanılarak elde edilen hiyerarşik kümeleme sonucunda oluşan iki kümenin arasındaki mesafenin tam bağlantı yöntemine göre daha az olduğu gözlemlendi. Oluşan küme sonucu veri setindeki ülkelerin IGE sınıfları ile aynı şekilde kümelendi.



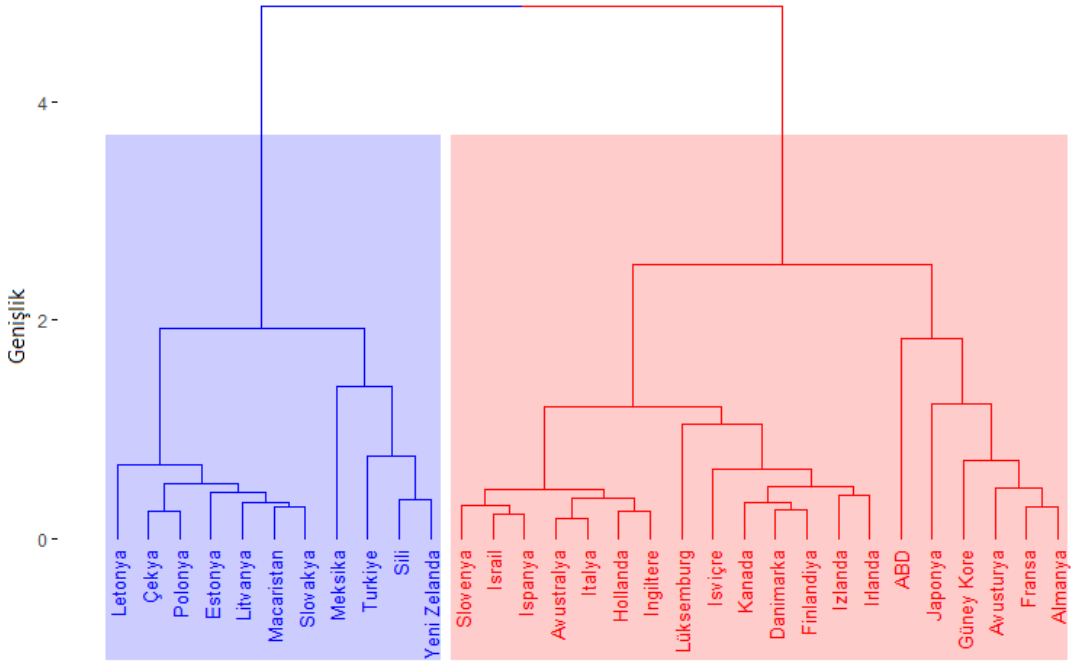

**Şekil 3.13** Ortalama Bağlantı Yöntemi ile Hiyerarşik Kümeleme (2016 Yılı)

Tek bağlantı yönteminin kullanıldığı hiyerarşik kümeleme dendrogramına göre oluşan kümeler Şekil 3.14'te gösterildi. Kümeleme işleminin sonucunda ABD diğer ülkelerden ayrıştı.




**Şekil 3.14** Tek Bağlantı Yöntemi ile Hiyerarşik Kümeleme (2016 Yılı)

Ward bağlantı yöntemi ile oluşan kümeler Şekil 3.15'te verildi. Oluşan iki kümeden 11 tanesi Küme-1, geriye kalan 20 tanesi de Küme-2'ye dahil oldu. İki küme arasındaki genişlik 0-4 aralığındadır. Bu diğer tüm yöntemlere kıyasla en büyük değerdir.




**Şekil 3.15** Ward Yöntemi ile Hiyerarşik Kümeleme (2016 Yılı)

2016 yılına ait verilere göre ward yöntemi ile oluşturulmuş olan kümeler Tablo 3.12'de gösterildi. 2011 yılından farklı olarak Şili ve Yeni Zelanda Küme-1'e geçti.

**Tablo 3.12** Ward bağlantı yöntemi ile hiyerarşik kümeleme sonucu oluşan kümeler

| Küme-1 | Küme-2 | |
|---|---|---|
| Türkiye | Japonya | İtalya |
| Meksika | ABD | Slovenya |
| Letonya | İsrail | Avusturya |
| Litvanya | İspanya | Finlandiya |
| Çekya | İngiltere | Hollanda |
| Polonya | Avustralya | Güney Kore |
| Estonya | Danimarka | Fransa |
| Macaristan | Kanada | Almanya |
| Slovakya | İrlanda | Lüksemburg |
| Şili | İsviçre | İzlanda |
| Yeni Zelanda | | |

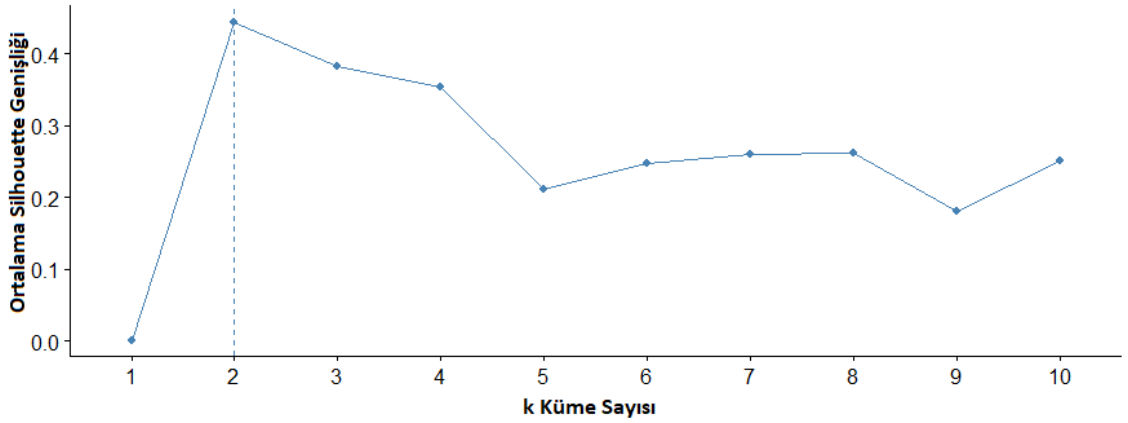
### 3.3.2. K-Means kümeleme

R programlama dilinde yukarıdaki adımlara göre çalışan "kmeans" fonksiyonu kullanıldı. K-means algoritmasında kullanılacak veri seti nümerik data frame, nümerik matris veya nümerik vektör olabilir. Bu algoritmada maksimum iterasyon sayısı da belirtilebilir ancak bu durum kesinleşmemiş kümeler elde edilmesine yol açabilir. İter.max parametresi ile kmeans fonksiyonuna maksimum iterasyon değeri atanabilir. Bu fonksiyon sonuçları list formatında döndürmektedir. Algoritma sonucunda dönen list içerisinde clusters, centers, totts, withiss, tot.withiss, betweenss, size, iter ve ifault bilgileri yer alır.

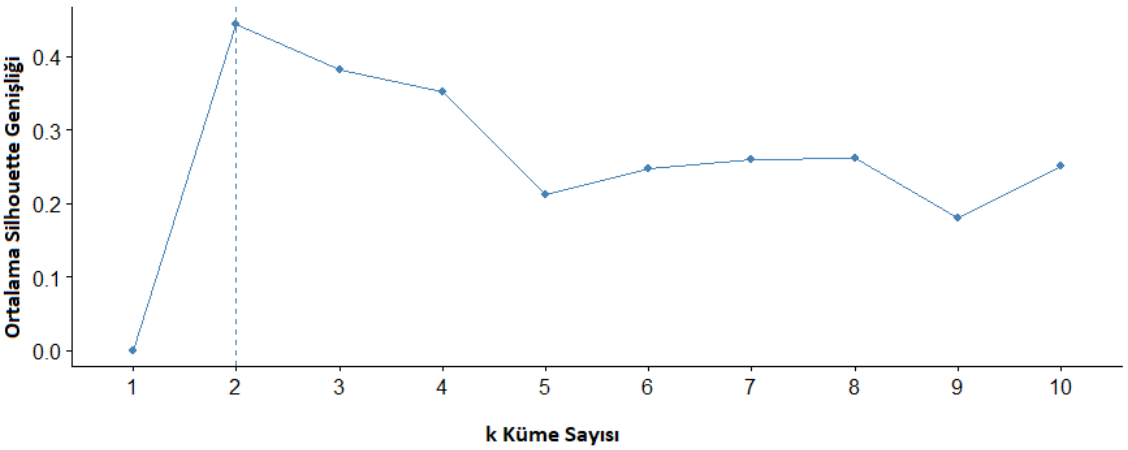
- **cluster**: her noktanın dahil olduğu kümeyi gösteren bir tamsayı vektörüdür.
- **centers**: Küme merkezlerinin bulunduğu matristir
- **totss**: Toplam kareler toplamı
- **withinss**: Küme içi kareler toplamının vektörel temsili
- **tot.withinss**: withinss vektöründeki değerlerin toplamı

- **betweenss**: Kümeler arasındaki kareler toplamı (totss - tot.withinss).

K-means kümeleme işleminin ilk adımı küme sayısının belirlenmesidir. En uygun küme sayısının belirlenmesi için birçok yöntem mevcuttur. Ortalama Silhouette Genişliği de K-Means kümelemede uygun küme sayısının belirlenmesinde sıkça kullanılan yöntemlerden bir tanesidir. Bu yöntemin kullanılabilmesi için öncelikle kümeleme yapılacak veri seti ve uygulanacak yöntemin bilgisine ihtiyaç bulunmaktadır. R Studio'da fviz_nbclust fonksiyonu yardımıyla 2011 – 2016 yılları veri setlerinin her biri için Ortalama Silhouette Genişliği kullanılarak küme sayısı belirlendi. 2011 – 2016 yıllarının tümünde en uygun küme sayısı iki olarak belirlendi. Şekil 3.16'da 2011 yılına ait verilerin ortalama silhouette genişliğine göre optimum küme sayısı gösterildi. Şekil 3.17'de ise 2016 yılına ait verilerin ortalama Silhouette genişliğine göre optimum küme sayısı gösterildi.

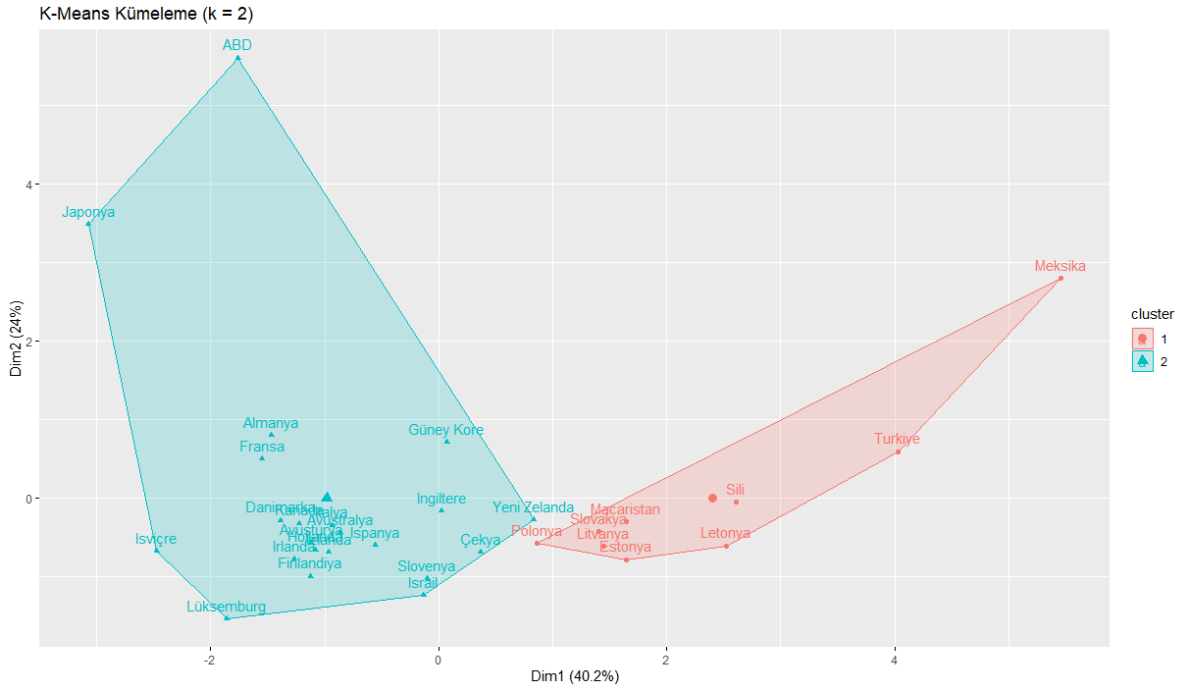



**Şekil 3.16** Ortalama Silhouette Genişliği (2011 Yılı)




**Şekil 3.17** Ortalama Silhouette Genişliği (2016 Yılı)

Küme sayısının belirlenmesinden sonra kümeleme çalışmasında kullanılacak değişkenler üzerinde min – max normalizasyonu yapıldı. K-means yöntemi kullanılarak 2011 yılına ait verilerin kümelenmesi sonucu oluşan kümeler R programlamada fviz_cluster fonksiyonu ile görselleştirildi. Kümeleme sonucu 2D boyuta çevirerek görselleştirildi. Varyansa en çok etki eden boyutlar Dim1 ve Dim 2 olarak gösterildi. Dim 1 ilaç harcamasını, Dim 2 hastane sayısını göstermektedir. Şekil 3.18'de 2011 yılına ait kümeleme sonuçları yer almaktadır. Kırmızı ile gösterilen küme Küme-1'i mavi ile gösterilen küme Küme-2'yi temsil etmektedir. Elde edilen iki merkez tabanlı kümede yer alan ülkeler Tablo 3.13'de verilmiştir.




**Şekil 3.18** 2011 yılı verilerinin K-means kümeleme ile kümeleme sonucu

Küme-1'de dokuz ülke Küme-2'de 22 ülke bulunmaktadır. Kümeleme sonucu 2D boyuta çevirerek görselleştirildi. Dim1 değeri %40.2 ve Dim2 değeri %24 olarak hesaplandı.

İşlemler tek bir iterasyonda yapıldı. Withinss yani küme içi kareler toplamının değerleri Küme-1 için 2.339155 ve Küme-2 için 6.614235 olarak hesaplandı. Betweenss değeri yani kümeler arası kareler toplamı 4.894838 olarak bulundu. totss (toplam kareler toplamı) değeri 13.84823 olarak hesaplandı.

4.894838 / 13.84823 = 0.3534631

Toplam varyans yüzdesi %35.3 olarak hesaplandı.

**Tablo 3.13** Kmeans kümeleme sonucu oluşan kümeler

| Küme-1 | Küme-2 | |
|---|---|---|
| Türkiye | Japonya | İtalya |
| Meksika | ABD | Slovenya |
| Letonya | İsrail | Avusturya |
| Litvanya | İspanya | Finlandiya |
| Şili | İngiltere | Hollanda |
| Polonya | Avustralya | Güney Kore |
| Estonya | Danimarka | Fransa |
| Macaristan | Kanada | Lüksemburg |
| Slovakya | İrlanda | İzlanda |
| | İsviçre | Yeni Zelanda |
| | Çekya | Almanya |

Değişkenlere ait küme ortalamaları Tablo 3.14'te verildi. Bu değerlere göre Küme-1'de en yüksek ortalama değere sahip değişken yaşam beklentisi iken Küme-2'de en yüksek ortalama değerine sahip değişken bebek ölüm oranı oldu.

**Tablo 3.14** Değişkenlerin küme ortalamaları (2011 yılı)

| Değişken | Küme-1 | Küme-2 |
|---|---|---|
| GSYIH | 0.34669847 | 0.09701709 |
| Sağlık Harcaması | 0.4233631 | 0.1750000 |
| Yatak Sayısı | 0.2979035 | 0.3038396 |
| Hastane Sayısı | 0.18441795 | 0.09571971 |
| Bebek Ölüm Oranı | 0.2175021 | 0.4263158 |
| Yaşam Beklentisi | 0.7821682 | 0.2585106 |
| Anne Ölüm Oranı | 0.1337875 | 0.3613079 |
| Medikal Teknoloji | 0.09611946 | 0.01590308 |
| İlaç Harcaması | 0.5439036 | 0.2495958 |

2016 yılına ait K-means kümeleme ile elde edilen sonuçlar Şekil 3.19'da gösterildiği şekilde R programlamada görselleştirildi. Burada merkez tabanlı iki küme oluştu ancak aynı zamanda bitişik kümeler elde edildi.


**Şekil 3.19** 2016 yılı verilerinin K-means Kümeleme ile Kümeleme Sonucu

Birinci Küme-1'de dokuz ülke Küme-2'de 22 ülke bulunmaktadır. Dim1 değeri %39.7 ve Dim2 değeri %24.5 olarak hesaplandı.

İşlemler tek bir iterasyonda yapıldı. Withinss yani küme içi kareler toplamının değerleri Küme-1 için 6.306011 ve Küme-2 için 2.597196 olarak hesaplandı. Betweenss değeri yani kümeler arası kareler toplamı 4.0277 olarak bulundu. totss (toplam kareler toplamı) değeri 12.93091 olarak hesaplandı.

4.0277 / 12.93091 = 0.3534631

Toplam varyans yüzdesi %31.1 olarak hesaplandı. Tablo 3.15'te 2016 yılına ait verilerin kümelenmesi ile oluşan kümeler belirtildi.

**Tablo 3.15** Kmeans kümeleme sonucu oluşan kümeler (2016 yılı)

| Küme-1 | | Küme-2 |
|---|---|---|
| Japonya | İtalya | Türkiye |
| ABD | Slovenya | Meksika |

| | | |
|---|---|---|
| İsrail | Avusturya | Letonya |
| İspanya | Finlandiya | Litvanya |
| İngiltere | Hollanda | Şili |
| Avustralya | Güney Kore | Polonya |
| Danimarka | Fransa | Estonya |
| Kanada | Lüksemburg | Macaristan |
| İrlanda | İzlanda | Slovakya |
| İsviçre | Yeni Zelanda | Çekya |
| Almanya | | |

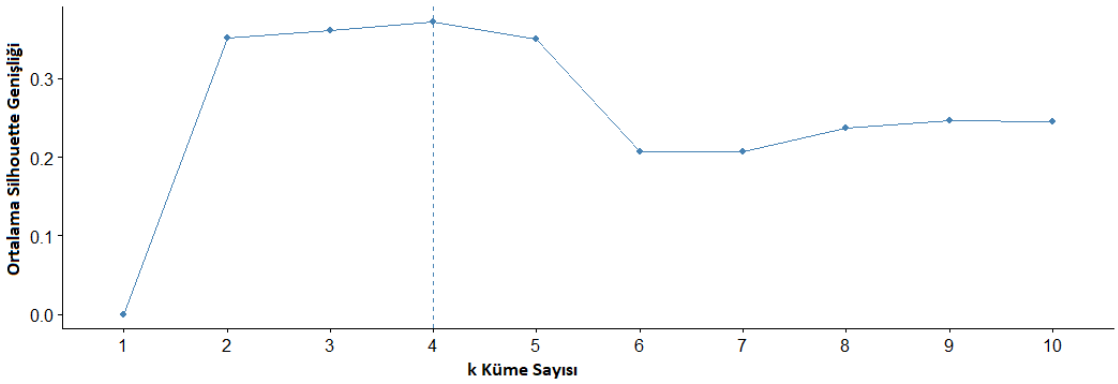
Değişkenlere ait küme ortalamaları Tablo 3.16'da verildi. Bu değerlere göre Küme-1'de en yüksek ortalama değere sahip değişken yaşam beklentisi iken Küme-2'de en yüksek ortalama değerine sahip değişken bebek ölüm oranı oldu.

**Tablo 3.16** Değişkenlerin küme ortalamaları (2016 yılı)

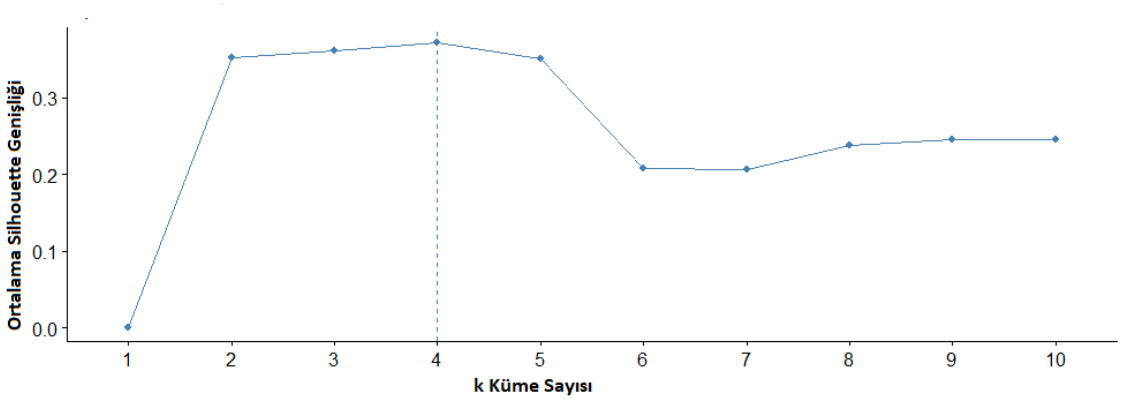
| Değişken | Küme-1 | Küme-2 |
|---|---|---|
| GSYIH | 0.34669847 | 0.09701709 |
| Sağlık Harcaması | 0.4233631 | 0.1750000 |
| Yatak Sayısı | 0.2979035 | 0.3038396 |
| Hastane Sayısı | 0.18441795 | 0.09571971 |
| Bebek Ölüm Oranı | 0.2175021 | 0.4263158 |
| Yaşam Beklentisi | 0.7821682 | 0.2585106 |
| Anne Ölüm Oranı | 0.1337875 | 0.3613079 |
| Medikal Teknoloji | 0.09611946 | 0.01590308 |
| İlaç Harcaması | 0.5439036 | 0.2495958 |

### 3.3.3. K-Medoids kümeleme

OECD sağlık veri setinin kullanılacağı K-medoids yönteminde k küme sayısı değerini belirlemek için ortalama Silhouette genişliğinden yararlanıldı. Şekil 3.20'de 2011 yılın ait ortalama Silhouette genişliği grafiği verildi. Şekil 3.21'te 2016 yılın ait ortalama Silhouette genişliği grafiği verildi. Her ikisinde de optimal küme değeri dört olarak belirlendi. Bu nedenle k değeri dört olarak atandı.

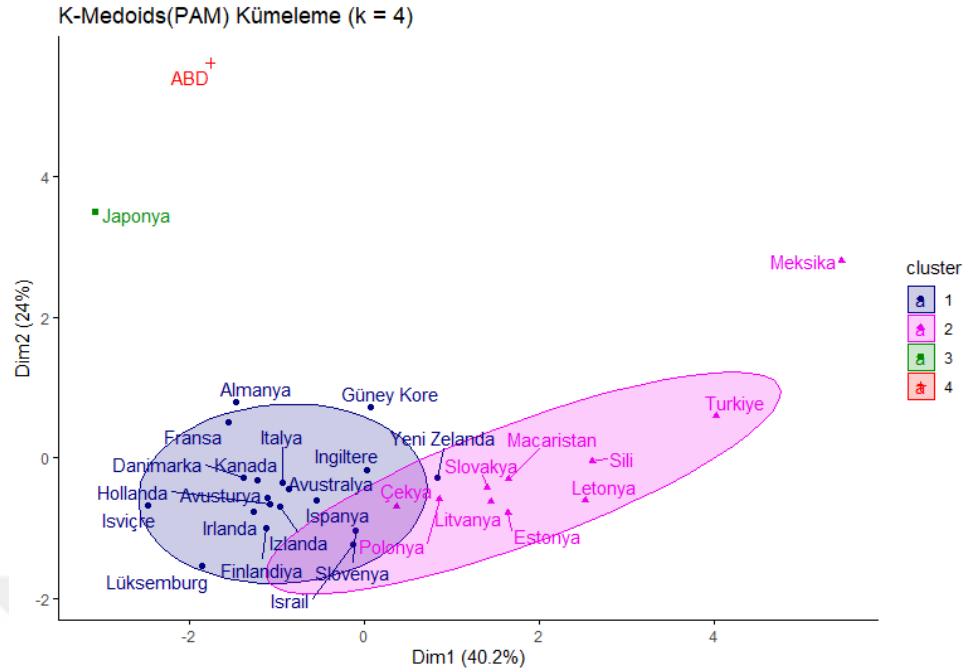



**Şekil 3.20** Ortalama Silhouette Genişliği (2011 Yılı)




**Şekil 3.21** Ortalama Silhouette Genişliği (2016 Yılı)

Küme sayısının belirlenmesinden sonra kümeleme çalışmasında kullanılacak değişkenler üzerinde min – max normalizasyonu yapıldı. K-medoids yöntemi kullanılarak 2011 yılına ait verilerin kümelenmesi sonucu oluşan kümeler R programlamada fviz_cluster fonksiyonu ile görselleştirildi. Şekil 3.22'de 2011 yılına ait kümeleme sonuçları yer almaktadır. Mavi ile gösterilen küme Küme-1'i, pembe ile gösterilen küme Küme-2'yi, yeşil ile temsil edilen küme Küme-3'ü ve kırmızı ile gösterilen küme Küme-4'ü temsil etmektedir. Kümeleme sonucu 2D boyuta çevirerek görselleştirildi. Varyansa en çok etki eden boyutlar Dim1 ve Dim 2 olarak gösterildi. Dim 1 ilaç harcamasını, Dim 2 hastane sayısını göstermektedir. Elde edilen dört merkez tabanlı kümede yer alan ülkeler Tablo 3.17'de verildi.




**Şekil 3.22** 2011 yılı verilerinin K-medoids Kümeleme ile Oluşan Kümeler

ABD ve Japonya tek başlarına birer küme oluşturdu. Küme-2'nin medoidi Slovakya'dır. Küme-1'in medoidi Avusturalya'dır. Dim1 değeri %40.2 ve Dim2 değeri %24 olarak bulundu.

**Tablo 3.17** K-medoids kümeleme sonucu oluşan kümeler (2011 yılı)

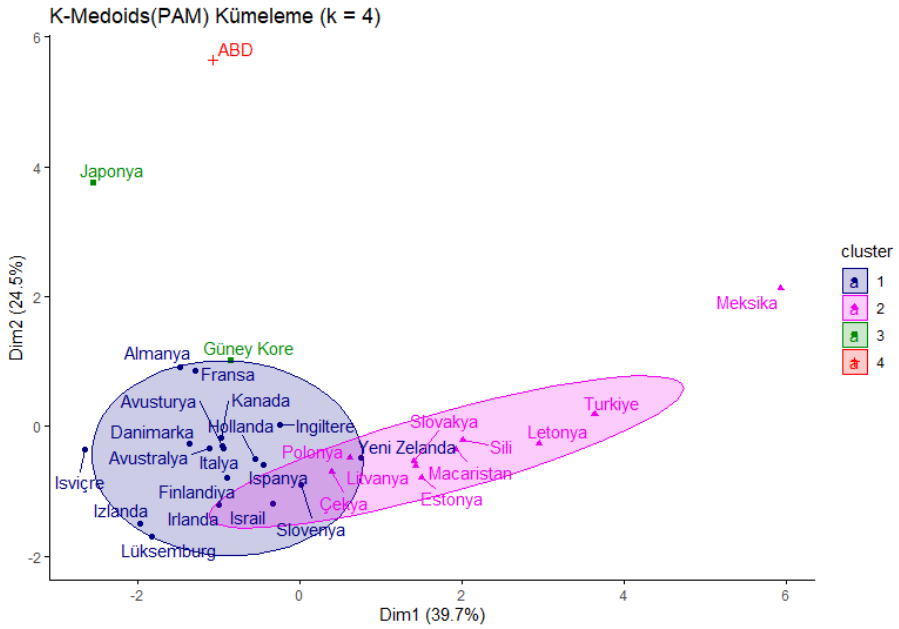
| Küme-1 | | Küme-2 | Küme-3 | Küme-4 |
|---|---|---|---|---|
| İsrail | Slovenya | Türkiye | Japonya | ABD |
| İspanya | Avusturya | Meksika | | |
| İngiltere | Finlandiya | Letonya | | |
| Avustralya | Hollanda | Litvanya | | |
| Danimarka | Fransa | Şili | | |
| Kanada | Güney Kore | Polonya | | |
| İrlanda | İtalya | Estonya | | |
| İsviçre | | Macaristan | | |
| Almanya | | Slovakya | | |
| Lüksemburg | | Çekya | | |
| İzlanda | | | | |
| Yeni Zelanda | | | | |

2011 yılı verilerine göre PAM algoritması ile yapılan kümelemede küme medoidleri Avusturalya, Slovakya, Japonya ve ABD oldu. K-medoids kümelemede kullanılacak veriler üzerinde min-max normalizasyonu yapıldığı için değerler 0-1 aralığındadır. Japonya hastane yatak sayısı ve hastane sayısı değişkenlerinde en yüksek değere sahip ülke olduğu için bu değişkenlerin değeri bir oldu. ABD sağlık harcaması ve medikal teknoloji değişkenlerinde en yüksek değere sahip olduğu için bu değişkenlerin değeri bir oldu.

**Tablo 3.18** 2011 yılı verilerine göre küme medoidleri

| Gösterge | Avusturalya | Slovakya | Japonya | ABD |
|---|---|---|---|---|
| GSYH | 0.37030837 | 0.12627048 | 0.25546388 | 0.44194667 |
| Sağlık Harcaması | 0.32478632 | 0.23076923 | 0.5042735 | 1.000000000 |
| Hastane Yatak Sayısı | 0.1984987 | 0.3869892 | 1.000000000 | 0.1301084 |
| Hastane Sayısı | 0.155519367 | 0.015354193 | 1.000000000 | 0.664883099 |
| Bebek Ölüm Oranı | 0.2265625 | 0.3125000 | 0.109375 | 0.40625 |
| Yaşam Beklentisi | 0.91208791 | 0.26373626 | 0.98901099 | 0.54945055 |
| Anne Ölüm Oranı | 0.1000000 | 0.23023256 | 0.09534884 | 0.35116279 |
| Medikal Teknoloji | 0.0401907 | 0.004609749 | 0.437049761 | 1.000000000 |
| İlaç Harcaması | 0.54310458 | 0.30271618 | 0.82010235 | 0.47382233 |

Şekil 3.23'te 2016 yılına ait kümeleme sonuçları yer almaktadır. Mavi ile gösterilen küme Küme-1'i, pembe ile gösterilen küme Küme-2'yi, yeşil ile temsil edilen küme Küme-3'ü ve kırmızı ile gösterilen küme Küme-4'ü temsil etmektedir.




**Şekil 3.23** 2016 yılı verilerinin K-medoids Kümeleme ile Oluşan Kümeler

Elde edilen dört merkez tabanlı kümede yer alan ülkeler Tablo 3.19'de verildi. Güney Kore 2011 yılı kümeleme sonuçlarında Küme-1 içerisinde yer alırken 2016 yılı verilerine göre Küme-3'te yer aldı.

**Tablo 3.19** K-medoids kümeleme sonucu oluşan kümeler (2016 yılı)

| Küme-1 | | Küme-2 | Küme-3 | Küme-4 |
|---|---|---|---|---|
| İsrail | Slovenya | Türkiye | Japonya | ABD |
| İspanya | Avusturya | Meksika | Güney Kore | |
| İngiltere | Finlandiya | Letonya | | |
| Avustralya | Hollanda | Litvanya | | |
| Danimarka | Fransa | Şili | | |
| Kanada | İzlanda | Polonya | | |
| İrlanda | Yeni Zelanda | Estonya | | |
| İsviçre | İtalya | Macaristan | | |
| Almanya | | Slovakya | | |
| Lüksemburg | | Çekya | | |

Tablo 3.20'de gösterildiği gibi 2016 yılı verilerine göre PAM algoritması ile yapılan kümelemede küme medoidleri Avusturalya, Slovakya, Japonya ve ABD oldu. K-medoids kümelemede kullanılacak veriler üzerinde min-max normalizasyonu yapıldığı için

değerler 0-1 aralığındadır. Japonya hastane yatak sayısı, hastane sayısı ve yaşam beklentisi değişkenlerinde en yüksek değere sahip ülke olduğu için bu değişkenlerin değeri bir oldu. ABD sağlık harcaması ve medikal teknoloji değişkenlerinde en yüksek değere sahip olduğu için bu değişkenlerin değeri bir oldu.

**Tablo 3.20** 2016 yılı verilerine göre küme medoidleri

| Gösterge | Avusturalya | Slovakya | Japonya | ABD |
|---|---|---|---|---|
| GSYH | 0.3523861 | 0.1335045 | 0.2496340 | 0.4382024 |
| Sağlık Harcaması | 0.3828125 | 0.2109375 | 0.5078125 | 1.0000000 |
| Hastane Yatak Sayısı | 0.2090444 | 0.3745734 | 1.0000000 | 0.1177474 |
| Hastane Sayısı | 0.15935499 | 0.01482096 | 1.00000000 | 0.65520512 |
| Bebek Ölüm Oranı | 0.2105263 | 0.4122807 | 0.1140351 | 0.4561404 |
| Yaşam Beklentisi | 0.8297872 | 0.2765957 | 1.0000000 | 0.4255319 |
| Anne Ölüm Oranı | 0.1062670 | 0.1880109 | 0.1008174 | 0.3136240 |
| Medikal Teknoloji | 0.051247536 | 0.004951685 | 0.423409131 | 1.000000000 |
| İlaç Harcaması | 0.5843266 | 0.3441567 | 0.5978901 | 0.4489656 |

Uzaklık hesaplama yöntemi olarak Öklid uzaklığı kullanılarak maksimum uzaklık ve ortalama uzaklık değerleri hesaplandı. Tablo 3.21'de 2011 OECD sağlık verilerine uygulanan K-medoids kümeleme algoritması sonuçlarının istatistiki değerleri gösterildi. Küme ayrışması(Separation) değeri bir kümedeki gözlem ile başka bir kümedeki gözlem arasındaki en kısa mesafeyi göstermektedir. Küme çapı(diameter) bir kümedeki iki gözlem arasındaki maksimum farklılığı göstermektedir.

**Tablo 3.21** Kmedoids yöntemi ile kümeleme istatistikleri (2011 yılı)

| Kümeler | Küme boyutu | Maksimum Uzaklık | Ortalama Uzaklık | Küme Çapı | Küme Ayrışması |
|---|---|---|---|---|---|
| Küme-1 | 19 | 0.7146573 | 0.390970 | 1.060059 | 0.4396775 |
| Küme-2 | 10 | 1.2794481 | 0.440517 | 1.496262 | 0.4396775 |
| Küme-3 | 1 | 0.0000000 | 0.0000000 | 0.0000000 | 0.8992760 |
| Küme-4 | 1 | 0.0000000 | 0.0000000 | 0.0000000 | 1.2309757 |

Japonya ve ABD bulundukları kümede tek eleman oldukları için maksimum uzaklık, ortalama uzaklık ve küme çapı hesaplanamadı. Küme-2 küme çapı en büyük kümedir. Küme-1 ve Küme-2'nin küme ayrışma değeri aynıdır. Yani Küme-1'e en kısa mesafede olan başka kümedeki gözlem Küme-2'de aynı şekilde Küme-2'ye en kısa mesafede olan başka kümedeki gözlem Küme-1'de yer almaktadır.

Tablo 3.22'de 2016 OECD sağlık verilerine uygulanan K-medoids kümeleme algoritması sonuçlarının istatistiki değerleri gösterildi. ABD Küme-4'te tek olduğu için maksimum uzaklık, ortalama uzaklık ve küme çapı hesaplanamadı.

**Tablo 3.22** Kmedoids yöntemi ile kümeleme istatistikleri (2016 yılı)

| Kümeler | Küme boyutu | Maksimum Uzaklık | Ortalama Uzaklık | Küme Çapı | Küme Ayrışması |
|---|---|---|---|---|---|
| Küme-1 | 18 | 0.7487484 | 0.3519924 | 0.9936303 | 0.3512948 |
| Küme-2 | 10 | 1.2692653 | 0.4574893 | 1.4679867 | 0.3512948 |
| Küme-3 | 2 | 0.3684032 | 0.7368064 | 0.7368064 | 0.5235589 |
| Küme-4 | 1 | 0.0000000 | 0.0000000 | 0.0000000 | 1.238533 |

Küme-2'nin küme çapı en büyük kümedir. Aynı zamanda küme içi maksimum uzaklığa sahip küme de Küme-2 oldu. Küme boyunun küçük olması ve elemanlarının uzak olması nedeniyle küme içi ortalama uzaklığın en fazla olduğu küme Küme-3 olmuştur. Küme-3'e kıyasla Küme-1 ve Küme-2 içinde yer alan ülkeler kendi kümelerindeki yer alan ülkelere yakın özellikler taşımakta olduğu belirlendi.

### 3.4. Sınıflandırma Çalışması Bulguları

Bu başlık altında OECD ülkeleri sağlık göstergeleri üzerinde YSA, lojistik regresyon ve Naive Bayes veri madenciliği yöntemleri ile sınıflandırma yapılması hedeflenmektedir. Elde edilen verilere göre OECD içerisindeki ülkelerin sahip olduğu verilere göre hangi sınıfta yer alması gerektiği belirlenecektir. Sınıf değeri olarak IGE sınıfı kullanıldı. Kümeleme işleminde kullanılacak değişkenler Tablo 3.23'te verildi. $X_1,\ldots,X_3$ kümeleme için kullanılacak belirleyici değişkenleri göstermektedir. Y ise sınıf değişkeni olan IGE sınıfı değerini göstermektedir.

**Tablo 3.23** Sınıflandırma işleminde kullanılacak değişkenler

| Değişkenler | Değer |
|---|---|
| X1 | Kişi başına GSYH (PPP $) |
| X2 | GSYH İçerisindeki Tüm Sağlık Harcamalarının Yüzdesi(%) |
| X3 | 1000 Kişi Başına Düşen Yatak Sayısı |
| X4 | Hastane Sayısı (1 000 kişi aşına) |
| X5 | Bebek Ölüm Oranı (1 000 canlı doğum başına) |
| X6 | Doğumda Yaşam Beklentisi |
| X7 | Anne Ölüm Oranı (100 000 canlı doğum başına) |
| X8 | Medikal Teknoloji (Makine Sayısı) |
| X9 | Kişi Başına Düşen İlaç Harcaması (ABD $) |
| Y | IGE Sınıfı |

OECD sağlık veri setinde yer alan IGE sınıfı "Yüksek" ve "Çok Yüksek" olmak üzere iki sınıftan oluşmaktadır. Bu çalışmada bu değerler Yüksek = 0 ve Çok Yüksek =1 olacak şekilde kullanıldı. 2011-2016 yıllarına ait tüm veriler bir CSV dosyasında birleştirildi. Veri seti toplamda 186 kayıt içeriyor. Bunlardan %5'i "Yüksek" sınıfında %95'i "Çok Yüksek" sınıfında yer aldı.

### 3.4.1. Yapay sinir ağı ile sınıflandırma bulguları

YSA ile sınıflandırma işlemine başlamadan önce OECD sağlık verisinin ne kadarının eğitim için kullanılacağı ne kadarının da test verisi olarak ayrılacağı belirlenmelidir. OECD sağlık veri seti aşağıdaki kodlar ile R programlamada eğitim ve test seti olarak ayrıldı. R programlamada dplyr kütüphanesi içerisinde yer alan sample() fonksiyonu olasılıklı örneklem oluşturmak için kullanılır. Fonksiyon içerisinde yer alan prob değişkeni olasılık ağırlıklarını ayarlamak için kullanılır. Negatif değer almaz, hepsi sıfır olamaz ve verilen değerlerin toplamının bire eşit olması zorunluluğu yoktur. Rastgele bir şekilde bir ve iki değerleri atanır. Bir olanlar eğitim seti, iki olanlar test seti olacak şekilde ayırılır. Veri seti aşağıdaki R programlama kodları ile eğitim ve test seti olarak ayrıldı.

```
set.seed(1234)
```

```
ind <- sample(2, nrow(data), replace=TRUE, prob = c(0.8, 0.2))

egitim <- data[ind==1,]

test <- data[ind==2,]
```

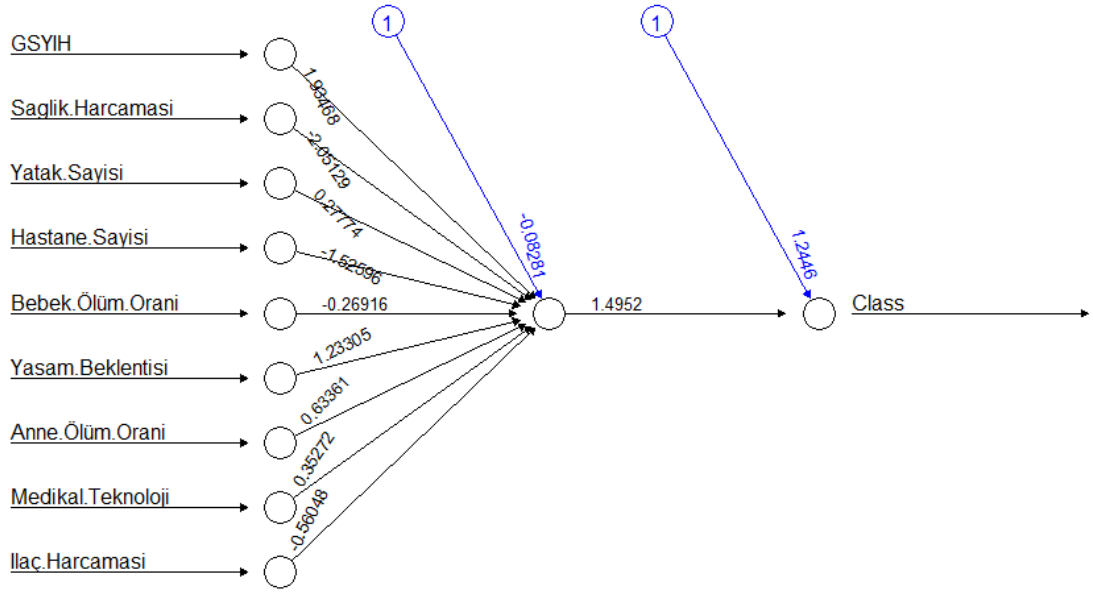
İki ayrı veri seti oluşturulduktan sonra YSA modeli kuruldu. Bu modeli oluştururken Tablo 2.23'te verilen değişkenler kullanıldı. R programlamada yer alan neuralnet kütüphanesi kullanıldı. Hata fonksiyonu olarak da detayları hata fonksiyonu başlığı altında verilen toplam karesel hata (SSE) yöntemi kullanıldı.

```
#Yapay Sinir Ağı ile Sınıflandırma

library(neuralnet)

ysa_model <- neuralnet(IGESinif~GSYIH + Saglik.Harcamasi + Yatak.Sayisi + Hastane.Sayisi + Bebek.Ölüm.Orani + Yasam.Beklentisi + Anne.Ölüm.Orani + Medikal.Teknoloji + Ilaç.Harcamasi, data = egitim, hidden = 1, err.fct = "sse", linear.output = FALSE)
```

Kurulan model R programlama dili ile görselleştirildi. Hata değeri 30.179411 olarak hesaplandı.




**Şekil 3.24** Kurulan YSA modeli

YSA modeli kurulduktan ve eğitim aşaması bittikten sonra tahmin aşamasına geçildi. R programlamada yapay sinir ağlarında tahmin yapabilmek için predict() fonksiyonu kullanıldı. Test verisinden sınıf değerleri çıkarılarak bir tahmin yapıldı.

```
tahmin <- predict(ysa_model, test[,-1])
```

Tahmin yapıldıktan sonra tahminin sınıflandırması işlemi yapıldı. İnsani gelişim endeksi bileşenleri bağlığı altında IGE değerlerinin sınıflarını sınırları verildi. 0,8 ve üzeri "Çok Yüksek" IGE değerine denk gelir. İkili lojistik regresyon kullanıldığı için tahmin edilen değer 0,8'ten büyükse bir yani "Çok Yüksek" sınıfında 0,8'den küçükse sıfır değerinin alır yani "Yüksek" sınıfındadır.

```
tahminSınıf <- ifelse(tahmin >0.8,1,0)
```

Test verisi üzerinde bu sınıflandırma işlemi yapıldıktan sonra gerçek test sınıfları ile tahmin edilen sınıflar bir tabloda karşılaştırıldı.

```
tablo <- table(Tahmin = tahminSınıf, GercekDeğer=test$ IGESinif)
```

Tahmin sonuçları Tablo 3.24'te karmaşıklık matrisi üzerinde gösterildi. Test seti 32 örneklemden oluşmaktadır. Sınıfların tahmin edilmesi %100 başarı ile gerçekleşti.

**Tablo 3.24** YSA sınıflandırıcısının karmaşıklık matrisi

| | Tahmin Edilen Negatif | Tahmin Edilen Pozitif |
|---|---|---|
| Gerçek Negatif | DN = 3 | YP = 0 |
| Gerçek Pozitif | YN = 0 | DP = 29 |

Karmaşıklık matrisi üzerinden sınıflandırmayı yorumlamak için doğruluk, hassasiyet, hata oranı, F1-skor ve duyarlılık değerleri R programlama ile hesaplandı ve Tablo 3.25'deki sonuçlar elde edildi. Bu değerlerin hesaplanma yöntemleri karmaşıklık matrisi başlığı altında verildi.

**Tablo 3.25** YSA karmaşıklık matrisi değerlendirme parametreleri

| Doğruluk | Hassasiyet | Hata Oranı | F1-Skor | Duyarlılık |
|---|---|---|---|---|
| 1 | 1 | 0 | 1 | 1 |

Tüm sınıflar doğru tahmin edildiği için hata oranı sıfır olarak bulundu. Doğruluk, hassasiyet, duyarlılık ve F1-Skor 1 olarak bulundu.

### 3.4.2. Lojistik regresyon ile sınıflandırma bulguları

Lojistik regresyon ile sınıflandırma işlemine başlamadan önce OECD sağlık verisinin ne kadarının eğitim için kullanılacağı ne kadarının da test verisi olarak ayrılacağı belirlenmelidir. OECD sağlık veri seti aşağıdaki kodlar ile R programlamada eğitim ve test seti olarak ayrıldı. R programlamada dplyr kütüphanesi içerisinde yer alan sample() fonksiyonu olasılıklı örneklem oluşturmak için kullanılır. Fonksiyon içerisinde yer alan prob değişkeni olasılık ağırlıklarını ayarlamak için kullanılır. Negatif değer almaz, hepsi sıfır olamaz ve verilen değerlerin toplamının bire eşit olması zorunluluğu yoktur. Rastgele bir şekilde bir ve iki değerleri atanır. Bir olanlar eğitim seti, iki olanlar test seti olacak şekilde ayırılır. Veri seti aşağıdaki R programlama kodları ile eğitim ve test seti olarak ayrıldı.

```
set.seed(333)

ind <- sample(2,nrow(data), replace = T , prob = c(0.8,0.2))

train <- data[ind==1,]

test <- data[ind==2,]
```

Veri seti ayırma işlemi gerçekleştikten sonraki aşama Lojistik regresyon modelinin kurulmasıdır. Bu modeli oluştururken Tablo 2.23'te verilen değişkenler kullanıldı. R programlamada yer alan glm() fonksiyonu lojistik regresyon için kullanılmaktadır. Aşağıda bu çalışma için R programlamada oluşturulan LRmodel isimli lojistik regresyon modeli bulunmaktadır. Family olarak binomial verilmesinin nedeni ikili(binary) lojistik regresyon yapılacak olmasıdır.

```
LRmodel <- glm(IGESinif~ GSYIH + Saglik.Harcamasi + Ilaç.Harcamasi + Medikal.Teknoloji + Anne.Ölüm.Orani + Yasam.Beklentisi + Bebek.Ölüm.Orani + Hastane.Sayisi + Yatak.Sayisi, data = egitim, family = 'binomial')

summary(LRmodel)
```

Modelin eğitimi eğitim seti üzerinden yapıldıktan sonra test seti üzerinde aşağıdaki R programlama kodu ile tahmin yapıldı. Tahmin işlemi için predict fonksiyonu kullanıldı.

Parametre olarak oluşturulan lojistik regresyon modeli, test seti ve type değeri verildi. İkili lojistik regresyon modelinde type olarak "response" kullanılması uygundur çünkü "response" olasılıkları döndürür.

```
tahmin <- predict(LRmodel, test, type= 'response')
```

Tahmin yapıldıktan sonra tahmin'in sınıflandırması işlemi yapıldı. Bunun için aşağıdaki karar sistemi kuruldu. İnsani gelişim endeksi bileşenleri bağlığı altında IGE değerlerinin sınıflarını sınırları verilmiştir. 0,8 ve üzeri "Çok Yüksek" IGE değerine denk gelir. İkili lojistik regresyon kullanıldığı için tahmin edilen değer 0,8'ten büyükse bir yani "Çok Yüksek" sınıfında 0,8'den küçükse sıfır değerinin alır yani "Yüksek" sınıfındadır.

```
tahminSınıf <- ifelse(tahmin >0.8,1,0)
```

Test verisi üzerinde bu sınıflandırma işlemi yapıldıktan sonra gerçek test sınıfları ile tahmin edilen sınıflar bir tabloda karşılaştırıldı.

```
tablo <- table(Tahmin = tahminSınıf, GercekDeğer=test$ IGESinif)
```

Elde edilen sonuçlara göre Tablo 3.26'da verilen lojistik regresyon sınıflandırmasına ait karmaşıklık matrisi oluşturuldu. Test seti 40 örneklemden oluşmaktadır.

**Tablo 3.26** Lojistik regresyon karmaşıklık matrisi

| | Tahmin Edilen Negatif | Tahmin Edilen Pozitif |
|---|---|---|
| Gerçek Negatif | DN = 1 | YP = 0 |
| Gerçek Pozitif | YN = 2 | DP = 37 |

Karmaşıklık matrisi üzerinden lojistik regresyon sınıflandırmasını yorumlamak için doğruluk, hassasiyet, hata oranı, F1-skor ve duyarlılık değerleri R programlama ile hesaplandı. Karmaşıklık matrisi parametreleri Tablo 3.27'de verildi.

**Tablo 3.27** Lojistik regresyon karmaşıklık matrisi değerlendirme parametreleri

| Doğruluk | Hassasiyet | Hata Oranı | F1-Skor | Duyarlılık |
|---|---|---|---|---|
| 0,95 | 1 | 0,05 | 0.9736 | 0.9487 |

Kurulan modelin sınıflandırma doğruluk oranı %95 olarak belirlendi. Bu oran tahmin için iyi bir orandır. Sadece iki örneğim hatalı tespit edilmesinden dolayı %5 olarak bulundu. Ancak lojistik regresyon ile sınıflandırma başarı olarak YSA'nın gerisinde kaldı.

### 3.4.3. Naive Bayes sınıflandırıcı ile sınıflandırma bulguları

Olasılıksal bir sınıflandırıcı olan Naive Bayes her bir özelliği sonuca etkisini olasılıksal olarak hesaplar ve ona göre bir sınıflandırma yapar. Çalışmanın bu kısmında kullanılacak değişkenler sınıflandırma çalışması bulguları başlığı altında Tablo 3.23'te verildi. Veriler ilk olarak sample() fonksiyonu yardımıyla test ve eğitim seti olarak iki gruba ayrıldı. Burada sınıflandırılacak verilerden 159 tanesi eğitim verisi 27 tanesi test verisi olacak şekilde ayrıldı. Veri seti ayrıldıktan sonra Naive Bayes sınıflandırma modeli kuruldu. Modelde tablo 3.23 verilen tüm değişkenler kullanıldı. R programlamada Naive Bayes ile sınıflandırma naive_bayes() fonksiyonunun yer aldığı naivebayes kütüphanesi kullanıldı.

```
library(naivebayes)

NBmodel <- naive_bayes(IGESinif~ GSYIH + Saglik.Harcamasi + Ilaç.Harcamasi + Medikal.Teknoloji + Anne.Ölüm.Orani + Yasam.Beklentisi + Bebek.Ölüm.Orani + Hastane.Sayisi + Yatak.Sayisi, data = egitim)
```

Kurulan model eğitildikten sonra test verisi üzerinde tahmin yapıldı. Sonuçlar bir tabloya kaydedildi. Bu tablo karmaşıklık matrisi değerlerini içerir.

```
tahmin <- predict(NBmodel, test)

tablo1 <- table (tahmin, test$ IGESinif)
```

Elde edilen sonuçlara göre Tablo 3.28'de verilen Naive Bayes sınıflandırmasına ait karmaşıklık matrisi oluşturuldu. Test seti 27 örneklemden oluşmaktadır.

**Tablo 3.28** Naive Bayes sınıflandırıcı karmaşıklık matrisi

| | Tahmin Edilen Negatif | Tahmin Edilen Pozitif |
|---|---|---|
| Gerçek Negatif | DN = 4 | YP = 2 |
| Gerçek Pozitif | YN = 0 | DP = 21 |

Karmaşıklık matrisi üzerinden Naive Bayes sınıflandırıcısının sonuçlarını yorumlamak için doğruluk, hassasiyet, hata oranı, F1-skor ve duyarlılık değerleri R programlama ile hesaplandı. Karmaşıklık matrisi parametreleri Tablo 3.29'de verildi.

**Tablo 3.29** Naive Bayes sınıflandırıcının karmaşıklık matrisi değerlendirme parametreleri

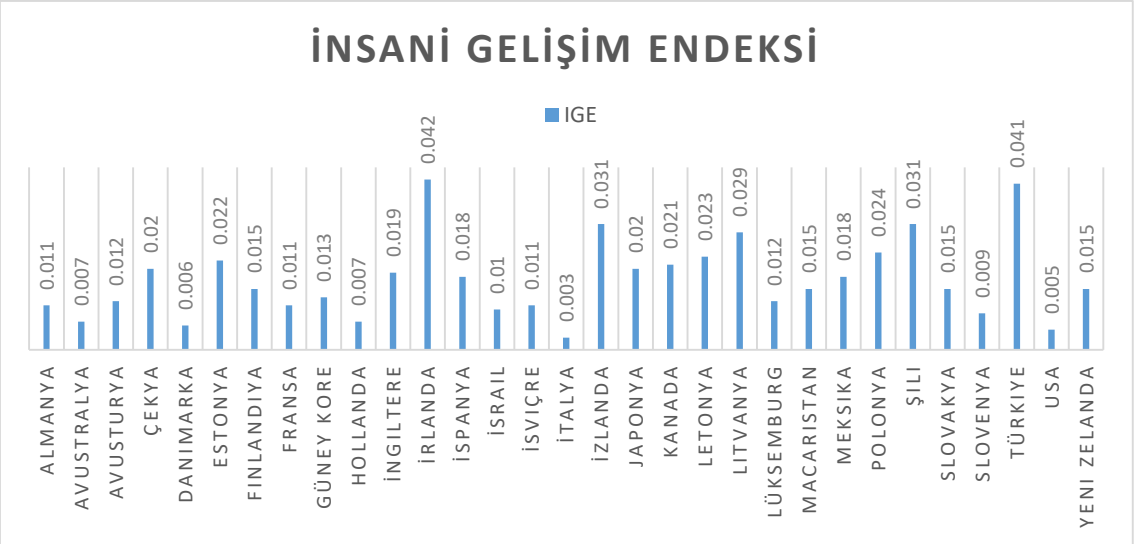
| Doğruluk | Hassasiyet | Hata Oranı | F1-Skor | Duyarlılık |
|---|---|---|---|---|
| 0.9259 | 0.9130 | 0.0741 | 0.9545 | 1 |

Kurulan modelin sınıflandırma doğruluk oranı %93 olarak belirlendi. Bu oran tahmin için iyi bir orandır. Sadece iki örneğin hatalı tespit edilmesinden dolayı %7 olarak bulundu. Ancak Naive Bayes sınıflandırıcı tahmin sonuçlarında lojistik regresyon ve YSA yönteminin gerisinde kaldı.

## 3.5. İnsani Gelişim Alanında En Fazla Gelişim Gösteren On Ülke

Ülkelerin insani gelişim indeksleri her yıl Birleşmiş Milletler Gelişim Programı(UNDP) tarafından ilan edilmektedir. Ülkeler sağlık, gelir ve eğitim parametrelerine etki eden değerler sonucunda IGE'de ilerleme veya gerileme yaşayabilmektedir. İnsani gelişim endeksinin sağlık boyutu hesaplanırken doğumda yaşam beklentisi yani ortalama ömür parametresi temel alınır. Bu çalışmada OECD ülkelerinde 2011-2016 yılları arasındaki IGE ve doğumda yaşam beklentisi üzerindeki değişimi gösterildi.

OECD ülkelerin 2011-2016 yılları arasındaki İnsani Gelişim Endeksi değişimi Şekil 3.26'da gösterildi. En fazla değişim yaşandığı on ülke Şekil 3.27'de gösterildiği gibi İrlanda, Türkiye, Şili, İzlanda, Litvanya, Polonya, Letonya, Estonya Kanada ve Çek Cumhuriyeti gösterdi. İrlanda bu altı yıl içerisinde IGE değerinde 0,042'lik bir artış yaşadı, onu 0,041 oranında bir artış ile Türkiye izledi. En düşük İnsani gelişim 0,003 ile İtalya tarafından gerçekleştirildi.



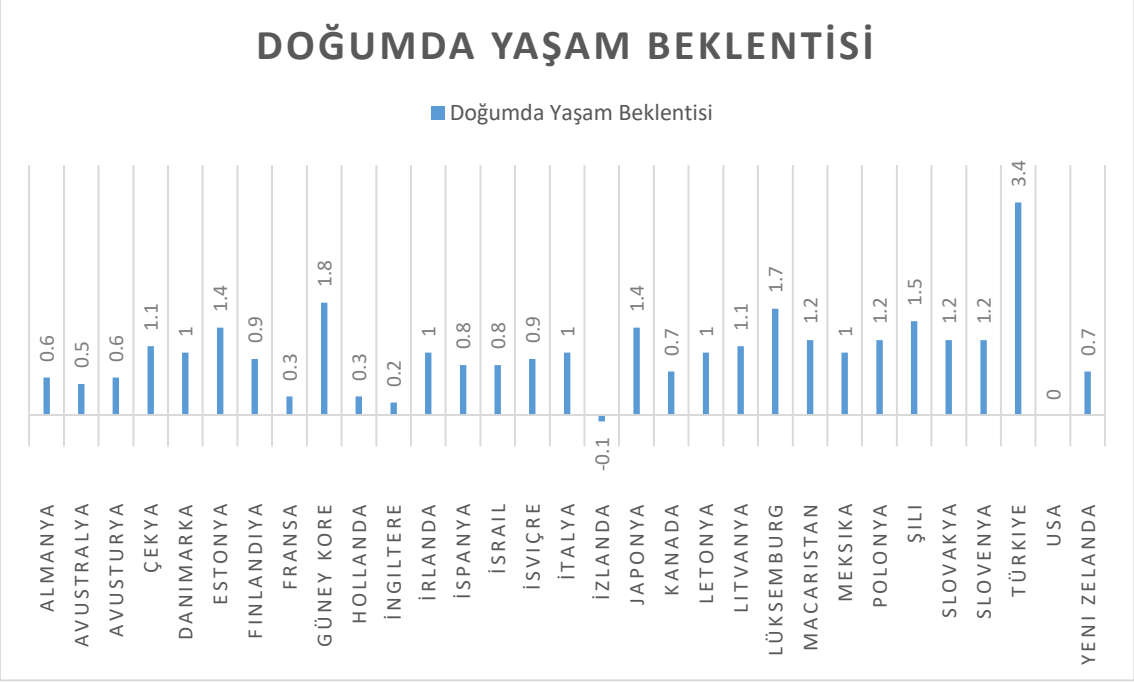

**Şekil 3.25** OECD ülkelerinin 2011-2016 yılları arası IGE değişimi


**Şekil 3.26** 2011-2016 yılları arasından en fazla insani gelişim gösteren on ülke

IGE değeri parametrelerinden biri olan sağlık üzerinde doğrudan etkisi olan doğumda yaşam beklentisini incelediğimizde ise Türkiye'nin 2011-2016 yılları arasında 3,4 yıl artış ile diğer OECD ülkelerinden açık ara farkla birinci olduğunu görüldü. Doğumda yaşam beklentisi üzerinde birçok etmen etkilidir. Türkiye son yıllarda sağlık alanında yaptığı yatırımlarla ve iyi eğitilmiş sağlık çalışanları ile bu artış trendini yakalayabildi. Türkiye'yi sırasıyla Güney Kore, Lüksemburg, Şili, Estonya, Japonya, Macaristan, Polonya, Slovakya ve Slovenya takip etti.


**Şekil 3.27** 2011-2016 yılları arasından doğumda yaşam beklentisini en fazla arttıran 10 ülke

2011-2016 yılları arasında doğumda yaşam beklentisinde en iyi gelişim gösteren 10 ülke Şekil 3.28'de gösterildi. Türkiye, Polonya, Estonya ve Şili hem IGE değerine hem de doğumda yaşam beklentisine göre en iyi gelişim gösteren ülkeler oldu. İzlanda'da 2011-2016 yılları arasında doğumda yaşam beklentisinde 0.1 yaş gerileme görüldü. Bu altı yıl içerinde OECD ülkelerinin doğumda yaşam beklentisi değerlerindeki değişim Şekil 3.29'da gösterildi.

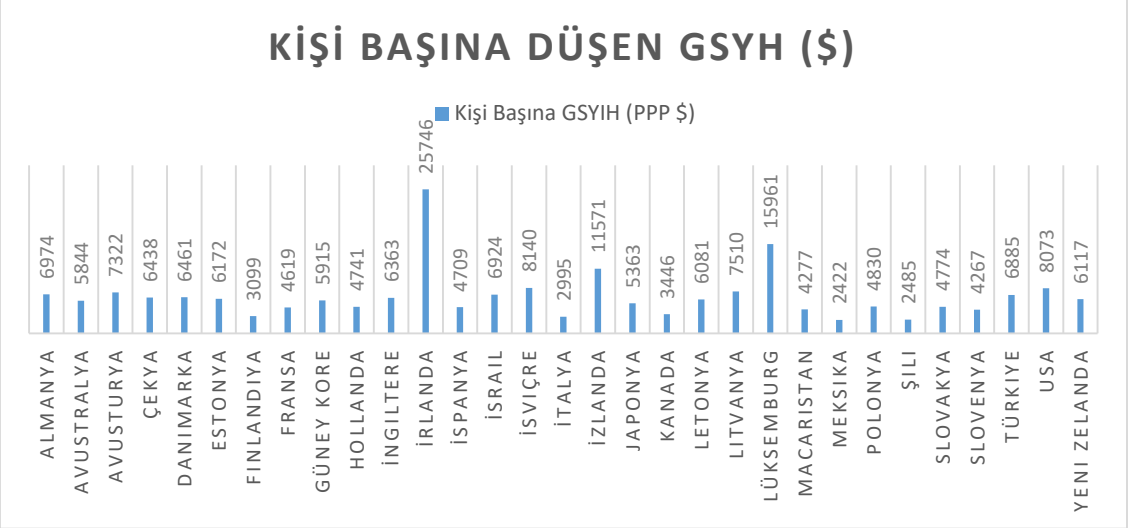



**Şekil 3.28** OECD ülkelerinin 2011-2016 yılları arası doğumda yaşam beklentisi değişimi

Hem IGE'nin gelir boyutuna doğrudan etkisi olduğu için hem de IGE'nin sağlık boyutuna dolaylı yoldan etkisi olduğu için Kişi başına düşen milli gelir kavramı insani gelişim

açısından büyük önem teşkil etmektedir. IGE'nin gelir endeksini oluşturan kişi başına düşen GSYH değerinin 2011-2016 yılları arasındaki değişim miktarı Şekil 3.31'de gösterildi. Bu sonuçlara göre en iyi gelişim gösteren 10 ülke Şekil 3.30'da gösterildi.


**Şekil 3.29** 2011-2016 yılları kişi başına milli gelirini en fazla arttıran 10 ülke




**Şekil 3.30** OECD ülkelerinin 2011-2016 yılları arası Kişi başına düşen milli gelir değişimi

OECD ülkelerini 2011-2016 yılları arasında kişi başına düşen mili gelir değişimi açısından karşılaştırdığımızda İrlanda 25746$ artış ile büyük bir gelişim gösterdi. Lüksemburg 15961$'lık artış ile ikinci sırada yer aldı. Bu iki ülkeyi İzlanda, İsviçre, ABD, Litvanya, Avusturya, Almanya, İsrail ve Türkiye izlemiştir. Türkiye bu altı yıl içerisinde kişi başına düşen milli gelirini en çok arttıran 10. Ülke olmasına rağmen 2016 yılında diğer OECD ülkeleri arasında en az kişi başına düşen milli gelire sahip dördüncü ülkedir.

## 4. SONUÇLAR

Ülkelerin gelişmişlik düzeyini belirlemek için kullanılan İnsani Gelişim Endeksi (HDI) ölçütü eğitim, sağlık ve gelir olmak üzere üç parametre göz önünde bulundurularak hesaplanır. OECD ülkelerinin IGE değerleri üzerinde etkili olan dokuz faktör; medikal teknoloji, hastane yatak sayısı, hastane sayısı, doğumda yaşam beklentisi, bebek ölüm oranı, anne ölüm oranı, kişi başına düşen milli gelir, sağlık harcaması ve kişi başına düşen ilaç satışı alanlarından oluşmaktadır. Birleşmiş Milletler, Dünya Bankası, Eurostsat olmak üzere birçok farklı kaynaktan derlenen OECD ülkeleri sağlık verilerinden seçilen bu dokuz faktöre ait değerler OECD sitesinde yayınlanan verilerden alınarak bir veri seti hazırlandı. 37 OECD üye ülkesinin 31 tanesi ve yıl aralığı olarak 2011 - 2016 yılları arasındaki veriler dâhil edilen veri seti üzerinde eğri uydurma ve ortalama değer atama veri önişleme teknikleri uygulanarak eksik veriler tamamlandı. Her bir yıla ait paralel koordinat sistemi eksen sıralaması yapılarak görselleştirildi. 2011 yılında bebek ölüm oranında ve IGE değerinde iki sınıf tamamen ayrıştı. 2011 yılına benzer şekilde 2012 ve 2013'de de IGE değeri ve bebek ölüm oranında iki sınıf birbirinden farklı aralıkta yer aldı. 2014 yılına bebek ölüm oranı ve IGE değerindeki ayrışmaya ilaç harcaması göstergesi de dahil oldu.

Çalışmanın ilk aşaması olan sonuçların paralel koordinat sisteminde gösterilmesi işlemi yapıldı. Veri setindeki veriler "Çok Yüksek" ve "Yüksek" olmak üzere iki sınıfa ayrıldığı için paralel koordinat sisteminde bu iki sınıf farklı renklerde görselleştirildi. Görselleştirme işlemi R programlama dili ile yapıldı. Veri seti içerisinde her bir yıl için 31 kayıt bulunmaktadır. 2011,2012 ve 2013 yılı verilerinin her birinde veri setinin % 0.067'si "Yüksek" sınıfına ve %0.933'ü "Çok Yüksek" sınıfına aittir. 2015 ve 2016 yıllarında Türkiye "Çok Yüksek" sınıfına geçtiği için bu yılların her birinde veri setinin % 0.033'ü "Yüksek" ve 0.967'si "Çok Yüksek" olarak değişti. 2015 yılında Türkiye'nin İGE değerini arttırması sonucu "Çok Yüksek" sınıfına geçmesi ile birlikte Meksika "Yüksek" sınıfında kalan tek ülke oldu. 2015 yılına ait paralel koordinat sistemi incelendiğinde Meksika'nın 2015 yılında 31 ülke içerisinde en yüksek bebek ölüm oranı ve en düşük IGE değeri, ilaç harcaması, GSYH ve yatak sayısı sayısına sahip olduğu görüldü. 2016 yılında ise 2015 yılındaki sonuçtan farklı olarak Meksika anne ölüm oranı bakımından da en kötü ülke konumuna geldi.

Çalışmanın ikinci kısmında 37 OECD ülkesi için dokuz göstergenin yer aldığı veri seti özellik seçimi için hazır hale getirildi. Medikal teknoloji, hastane yatak sayısı, hastane sayısı, doğumda yaşam beklentisi, bebek ölüm oranı, anne ölüm oranı, kişi başına GSYH, sağlık harcaması ve kişi başına düşen ilaç harcaması özniteliklerine göre özellik seçimi işlemi yapıldı. Çıkış değişkeni olarak Birlemiş Milletler'in UNDP programı kapsamında her bir yıl için yayınladığı IGE değerleri alındı. Bilgi kazancı tabanlı özellik seçimi yöntemi ve korelasyon tabanlı özellik seçimi yöntemi hazırlanan bu veri setine uygulandı. Her bir yıl için ayrı ayrı hesaplanan bu değerler sonucunda 2011-2016 yılları arasında OECD ülkeleri İnsani Gelişim Endeksine etki eden özelliğin bebek ölüm oranı olduğu belirlendi. Bebek ölüm oranından sonra en etkili olan özellik anne ölüm oranı olarak belirlendi. 2011-2016 yılları arasında korelasyon tabanlı özellik seçimi yöntemi ile incelenen özelliklerden en az etkili olan özelliğin medikal teknoloji olduğu belirlendi. Yapılan öznitelik seçme çalışmasının sonuçları literatürde yer alan çalışmalarla karşılaştırıldığında bebek ölüm oranı ve anne ölüm oranının insani gelişime etkisi literatürde de belirtildiği gibi güçlü oldu. Ancak medikal teknoloji literatürdeki çalışmaların aksine yapılan çalışmada insani gelişime en az katkıyı sağlayan öznitelik oldu.

Veri madenciliğinde kullanılan bir diğer önemli yöntem de kümeleme analizidir. Veri madenciliği yöntemlerini araştırıldığı bu çalışmanın üçüncü bölümünde OECD ülkeleri sağlık verileri üzerinde hiyerarşik kümeleme, K-means ve K-medoids ile kümeleme yapıldı. Hiyerarşik kümelemede ve K-means kümelemede optimum küme sayısı ortalama Silhouette genişliği kullanılarak iki olarak, K-medoids kümeleme yöntemi için ise dört olarak belirlendi. Hiyerarşik kümelemede tam, ortalama, tek ve ward bağlantı yöntemi kullanılarak sınıflandırma yapıldı. Bir yıla ait verilerde her bir yöntemden elde edilen kümeler birbirinden farklı oldu. Hiyerarşik kümeler dendgrogram üzerindeki genişliklere göre yorumlandı. 2011 ve 2016 yılı verilerine göre kümeler arası mesafenin en fazla olduğu yöntem ward yöntemi oldu. Ward bağlantı yöntemi ile kullanılarak oluşturulan kümeler birbirine en az benzeyen kümeler oldu. Ancak ward yöntemi ile elde edilen kümeler IGE sınıfı değerine göre oluşan kümelerden farklı oldu. 2011 yılı verilerinde tam bağlantı yöntemi ile yapılan hiyerarşik kümelemede elde edilen kümeler IGE sınıfına göre oluşan kümelerle aynı oldu. 2016 yılı verilerinde ise ortalama bağlantı yöntemi ile yapılan hiyerarşik kümelemede elde edilen kümeler IGE sınıfına göre oluşan kümelerle

aynı oldu. 2011 ve 2016 yılı verileri üzerinde K-means kümeleme ile elde edilen küme sonuçlarına göre sadece Çekya bulunduğu kümeyi değiştirdi. Diğer tüm ülkeler sabit kaldı. K-means kümeleme sonuçlarına göre 2011 yılı verilerinde küme içi kareler toplamının değerleri Küme-1 için 2.339155 ve Küme-2 için 6.614235 olarak hesaplandı. 2016 yılı verilerinde küme içi kareler toplamının değerleri Küme-1 için 6.306011 ve Küme-2 için 2.597196 olarak hesaplandı. K-means kümelemede veriler dört kümeye ayrıldı. PAM algoritması ile yapılan kümelemede 2011 ve 2016 yılı verilerine göre küme medoidleri Avusturalya, Slovakya, Japonya ve ABD oldu. 2011 yılı verilerine göre ABD ve Japonya tek başlarına bir küme oldular. Küme-1 19 ülke Küme-2 10 ülkeden oluştu. Türkiye Küme-2 içerisinde yer aldı. 2016 yılında Güney Kore Küme-1'den Japonya'nın olduğu Küme-3'e geçti. ABD 2016 yılı verilerinde de tek başına bir küme oluşturdu. Küme-2'nin küme çapı en büyük küme oldu. Japonya hastane yatak sayısı, hastane sayısı ve yaşam beklentisi değişkenlerinde en yüksek değere sahip ülke olduğu için diğerler kümelerden ayrıldı. ABD sağlık harcaması ve medikal teknoloji değişkenlerinde en yüksek değere sahip olduğu için diğer kümelerden ayrıldı. 2011 ve 2016 yılları K-medoids kümeleme sonucuna göre Küme-2 içerisinde yer alan ülkeler diğer ülkelere nispeten belirlenen değişkenlerde daha kötü değerlere sahiptirler.

YSA, lojistik regresyon ve Naive Bayes yöntemlerinin karşılaştırıldığı çalışmada YSA rastgele seçilmiş 32 örnekten oluşan test seti üzerinde %100 doğru tahmin ile bu üç algoritma içerişinde en başarılı yöntem olarak tespit edildi. Lojistik regresyon rastgele seçilen 40 örnekten sadece iki örneğin sonucunu yanlış tahmin ederek %95 doğruluk oranında sınıf tespiti yaptı. Naive Bayes algoritması 27 örneklemden %93'ini doğru tahmin etti. Bu sonuçlara göre OECD sağlık verileri üzerinde en iyi tahmini YSA yaptı. Onu %95 doğrulukla Lojistik regresyon yöntemi izledi. Naive Bayes başarılı sınıflandırma yapmasına rağmen bu iki yöntemin gerisinde kaldı. Yapılan sınıflandırma sonucu literatürde yer alan sınıflandırma çalışmaları ile karşılaştırıldı. Çalışma sonunda elde edilen algoritma performansları literatürde yer alan çalışmalarla benzerlik gösterdi. YSA literatür çalışmalarında da belirtildiği gibi doğru tahmin etme performansı bakımından diğer yöntemlere karşı üstünlük sağladı.

Çalışmanın son bölümünde insani gelişim alanında en fazla gelişim gösteren on ülke araştırıldı. 2011-2016 yılları arasındaki altı yıllık dönemde İrlanda ve Türkiye IGE

değerini en fazla arttıran ülke oldu. Türkiye ve İrlanda'yı sırayla Şili, İzlanda, Litvanya, Polonya, Letonya, Estonya Kanada ve Çekya izledi.

Doğumda yaşam beklentisi bakımından karşılaştırıldığında Türkiye bu altı yıllık süreçte doğumda yaşam beklentisi değerini 3,4 yıl gibi büyük bir oranda arttırdı. Türkiye'yi 1,8 yıl ile Güney Kore izlemiştir. Doğumda yaşam beklentisi IGE değerinin sağlık göstergesini doğrudan etkileyen bir gösterge olduğu için doğumda yaşam beklentisinde yaşanan artış doğrudan insani gelişime etki etmektedir. OECD sağlık veri setinde yer alan ve IGE'nin gelir endeksine doğrudan etki eden kişi başına GSYH değerine göre sırasıyla İrlanda, Lüksemburg, İzlanda, İsviçre, ABD, Litvanya, Avusturya, Almanya, İsrail ve Türkiye kişi başına düşen GSYH oranını en fazla arttıran on ülke oldu.

# EKLER

**EK 4 Tablo 1** OECD Ülkeleri 2012 Yılı Sağlık Göstergeleri ve Verileri

| Ülkeler | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 |
|---|---|---|---|---|---|---|---|---|---|
| Avustralya | 43879 | 8.7 | 3.75 | 1347 | 3.3 | 82.1 | 5.2 | 2509 | 471.6 |
| Avusturya | 46478 | 10.2 | 7.67 | 277 | 3.2 | 81 | 1.3 | 761 | 398.3 |
| Kanada | 42247 | 10.4 | 2.78 | 720 | 4.8 | 81.5 | 5.7 | 2239 | 699.8 |
| Şili | 21447 | 7 | 2.17 | 384 | 7.4 | 78.7 | 17.2 | 582 | 111 |
| Çekya | 29051 | 7 | 6.93 | 252 | 2.6 | 78.2 | 5.5 | 576 | 407 |
| Danimarka | 44809 | 10.2 | 5.06 | 1302 | 3.4 | 80.1 | 0 | 482 | 631.6 |
| Estonya | 26141 | 5.8 | 5.53 | 56 | 3.6 | 76.5 | 7.1 | 60 | 215.4 |
| Finlandiya | 40873 | 9.3 | 5.3 | 263 | 2.4 | 80.7 | 3.4 | 503 | 470.2 |
| Fransa | 37684 | 11.3 | 6.34 | 2657 | 3.3 | 82.1 | 8.7 | 1917 | 534.1 |
| Almanya | 43360 | 10.8 | 8.34 | 3229 | 3.3 | 80.6 | 4.6 | 5040 | 486.3 |
| Macaristan | 23148 | 7.5 | 7 | 176 | 4.9 | 75.2 | 10 | 401 | 199.1 |
| İzlanda | 41928 | 8.2 | 3.25 | 8 | 1.1 | 83 | 0 | 31 | 605.2 |
| İrlanda | 46278 | 10.7 | 2.54 | 95 | 3.5 | 80.9 | 2.8 | 281 | 571.2 |
| İsrail | 31707 | 7.1 | 3.1 | 86 | 3.6 | 81.8 | 5.3 | 215 | 394.3 |
| İtalya | 36003 | 9 | 3.42 | 1156 | 2.9 | 82.3 | 2.1 | 6654 | 411.6 |
| Japonya | 37214 | 10.8 | 13.35 | 8565 | 2.2 | 83.2 | 4.8 | 25708 | 685.3 |
| Güney Kore | 32097 | 6.4 | 10.25 | 3298 | 2.9 | 80.9 | 9.9 | 6506 | 384.4 |
| Letonya | 21298 | 5.4 | 5.89 | 66 | 6.3 | 73.9 | 20.5 | 146 | 186.6 |
| Litvanya | 24646 | 6.3 | 7.43 | 105 | 3.9 | 74 | 9.8 | 168 | 394.3 |
| Lüksemburg | 91527 | 6.6 | 5.15 | 13 | 2.5 | 81.5 | 16.6 | 40 | 510.7 |
| Meksika | 17220 | 5.8 | 1.41 | 4421 | 13.3 | 74.4 | 42.3 | 2171 | 61 |
| Hollanda | 47272 | 10.5 | 4.25 | 499 | 3.7 | 81.2 | 3.4 | 607 | 353.4 |
| Yeni Zelanda | 32912 | 9.7 | 2.83 | 161 | 4.7 | 81.2 | 11.3 | 291 | 169.8 |
| Polonya | 23542 | 6.2 | 6.63 | 1038 | 4.6 | 76.9 | 1 | 1546 | 394.3 |
| Slovakya | 26940 | 7.6 | 5.91 | 137 | 5.8 | 76.2 | 3.6 | 304 | 262.4 |
| Slovenya | 29048 | 8.8 | 4.54 | 29 | 1.6 | 80.2 | 9.2 | 111 | 283.8 |
| İspanya | 31725 | 9.1 | 2.99 | 759 | 3.1 | 82.5 | 2.2 | 2799 | 304.1 |
| İsviçre | 57850 | 11.1 | 4.79 | 298 | 3.6 | 82.8 | 8.5 | 698 | 791.8 |
| Türkiye | 20473 | 4.5 | 2.66 | 1483 | 11.6 | 74.6 | 15.2 | 3142 | 100.1 |
| İngiltere | 38309 | 8.3 | 2.81 | 1747 | 4 | 81 | 6.4 | 2045 | 341.5 |
| ABD | 51541 | 16.3 | 2.93 | 5723 | 6 | 78.8 | 13.3 | 57911 | 394.3 |

**EK 1 Tablo 2** OECD Ülkeleri 2013 Yılı Sağlık Göstergeleri ve Verileri

| Ülkeler | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 |
|---|---|---|---|---|---|---|---|---|---|
| Avustralya | 47761 | 8.8 | 3.74 | 1359 | 3.6 | 82.2 | 1.9 | 2825 | 493.4 |
| Avusturya | 47937 | 10.3 | 7.64 | 278 | 3.1 | 81.2 | 1.3 | 763 | 412.6 |
| Kanada | 44211 | 10.3 | 2.71 | 722 | 5 | 81.7 | 6 | 2274 | 679 |
| Şili | 22353 | 7.4 | 2.16 | 389 | 7 | 79.5 | 15.2 | 655 | 116 |
| Çekya | 30496 | 7.8 | 6.7 | 253 | 2.5 | 78.3 | 1.9 | 574 | 387.4 |
| Danimarka | 46743 | 10.2 | 5.01 | 1320 | 3.5 | 80.4 | 3.6 | 508 | 647.6 |
| Estonya | 27596 | 6 | 5.01 | 31 | 2.1 | 77.3 | 7.3 | 64 | 232.7 |
| Finlandiya | 41493 | 9.5 | 4.87 | 259 | 1.8 | 81.1 | 1.7 | 502 | 496.4 |
| Fransa | 39528 | 11.4 | 6.28 | 3192 | 3.6 | 82.3 | 8 | 2047 | 541 |
| Almanya | 44994 | 10.9 | 8.28 | 3183 | 3.3 | 80.6 | 4.1 | 5051 | 519 |
| Macaristan | 24498 | 7.3 | 7.04 | 173 | 5 | 75.7 | 14.7 | 416 | 194.4 |
| İzlanda | 44153 | 8.2 | 3.22 | 8 | 1.8 | 82.1 | 0 | 30 | 657.9 |
| İrlanda | 47936 | 10.3 | 2.56 | 93 | 3.6 | 81 | 4.3 | 288 | 562.7 |
| İsrail | 34160 | 7.1 | 3.09 | 86 | 3.1 | 82.1 | 4.7 | 222 | 398.6 |
| İtalya | 36068 | 9 | 3.31 | 1135 | 2.9 | 82.8 | 2.8 | 6697 | 428.6 |
| Japonya | 39008 | 10.8 | 13.3 | 8540 | 2.1 | 83.4 | 4 | 25708 | 554.4 |
| Güney Kore | 32616 | 6.6 | 10.92 | 3450 | 3 | 81.4 | 11.5 | 6692 | 408.5 |
| Letonya | 22691 | 5.4 | 5.8 | 65 | 4.4 | 74.1 | 24.7 | 152 | 202.9 |
| Litvanya | 26680 | 6.1 | 7.28 | 99 | 3.7 | 74.1 | 6.7 | 173 | 398.6 |
| Lüksemburg | 95246 | 5.7 | 5.17 | 13 | 3.9 | 81.9 | 16.4 | 39 | 515.7 |
| Meksika | 17462 | 5.9 | 1.44 | 4436 | 13 | 74.6 | 38.2 | 2212 | 60.1 |
| Hollanda | 49243 | 10.6 | 4.18 | 509 | 3.8 | 81.4 | 2.3 | 596 | 339.8 |
| Yeni Zelanda | 36074 | 9.4 | 2.78 | 160 | 5 | 81.4 | 16.8 | 307 | 170.4 |
| Polonya | 24423 | 6.4 | 6.61 | 1085 | 4.6 | 77.1 | 1.9 | 1707 | 398.6 |
| Slovakya | 27969 | 7.5 | 5.8 | 136 | 5.5 | 76.5 | 1.8 | 307 | 301.5 |
| Slovenya | 29980 | 8.8 | 4.55 | 29 | 2.9 | 80.4 | 4.8 | 109 | 290.7 |
| İspanya | 32453 | 9 | 2.96 | 764 | 2.7 | 83.2 | 4.2 | 2873 | 303.9 |
| İsviçre | 60109 | 11.3 | 4.68 | 293 | 3.9 | 82.9 | 2.4 | 726 | 791 |
| Türkiye | 22205 | 4.4 | 2.65 | 1517 | 10.8 | 78 | 15.7 | 3243 | 98.9 |
| İngiltere | 39985 | 9.8 | 2.76 | 1645 | 3.9 | 81.1 | 6.4 | 2068 | 356 |
| ABD | 53046 | 16.3 | 2.89 | 5686 | 6 | 78.8 | 12.7 | 58287 | 398.6 |

**EK 2 Tablo 3** OECD Ülkeleri 2014 Yılı Sağlık Göstergeleri ve Verileri

| Ülkeler | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 |
|---|---|---|---|---|---|---|---|---|---|
| Avustralya | 47639 | 9 | 3.79 | 1322 | 3.4 | 82.4 | 4 | 2964 | 456.5 |
| Avusturya | 48814 | 10.4 | 7.58 | 279 | 3 | 81.6 | 8.6 | 780 | 432.7 |
| Kanada | 45628 | 10.1 | 2.67 | 720 | 4.7 | 81.8 | 6 | 2344 | 649.2 |
| Şili | 22688 | 7.8 | 2.11 | 363 | 7.2 | 79.7 | 13.5 | 774 | 111.4 |
| Çekya | 32265 | 7.7 | 6.68 | 257 | 2.4 | 78.9 | 6.4 | 547 | 371.1 |
| Danimarka | 47905 | 10.2 | 4.98 | 1313 | 4 | 80.8 | 8.8 | 508 | 675.9 |
| Estonya | 29108 | 6.1 | 5.01 | 30 | 2.7 | 77.2 | 0 | 65 | 251.7 |
| Finlandiya | 41750 | 9.5 | 4.53 | 258 | 2.2 | 81.3 | 5.3 | 526 | 514.7 |
| Fransa | 40144 | 11.6 | 6.2 | 3111 | 3.5 | 82.8 | 6.7 | 2225 | 541 |
| Almanya | 47011 | 11 | 8.23 | 3138 | 3.2 | 81.2 | 4.1 | 5332 | 549.9 |
| Macaristan | 25605 | 7.1 | 6.98 | 174 | 4.5 | 75.9 | 6.6 | 420 | 193 |
| İzlanda | 45713 | 8.3 | 3.16 | 8 | 2.1 | 82.9 | 0 | 30 | 619.7 |
| İrlanda | 51126 | 9.7 | 2.57 | 91 | 3.3 | 81.4 | 1.5 | 279 | 545.1 |
| İsrail | 34228 | 7.1 | 3.08 | 85 | 3.1 | 82.2 | 2.8 | 236 | 398.9 |
| İtalya | 36195 | 9 | 3.21 | 1121 | 2.8 | 83.2 | 1.2 | 6803 | 434 |
| Japonya | 39183 | 10.8 | 13.21 | 8493 | 2.1 | 83.7 | 3.3 | 26454 | 488.9 |
| Güney Kore | 33587 | 6.8 | 11.59 | 3672 | 3 | 81.8 | 11 | 6688 | 443 |
| Letonya | 23839 | 5.5 | 5.66 | 64 | 3.8 | 74.3 | 14 | 157 | 210.2 |
| Litvanya | 28156 | 6.2 | 7.22 | 94 | 3.9 | 74.7 | 3.3 | 172 | 398.9 |
| Lüksemburg | 100934 | 5.6 | 5.05 | 12 | 2.8 | 82.3 | 0 | 40 | 496.9 |
| Meksika | 18168 | 5.6 | 1.43 | 4395 | 12.5 | 74.8 | 38.9 | 2318 | 55.6 |
| Hollanda | 49233 | 10.6 | 4.09 | 505 | 3.6 | 81.8 | 2.9 | 661 | 340.5 |
| Yeni Zelanda | 37061 | 9.4 | 2.75 | 162 | 5.7 | 81.5 | 5.1 | 316 | 176.5 |
| Polonya | 25298 | 6.2 | 6.63 | 1096 | 4.2 | 77.7 | 2.1 | 1555 | 398.9 |
| Slovakya | 28992 | 6.9 | 5.79 | 134 | 5.8 | 76.9 | 3.6 | 334 | 322.5 |
| Slovenya | 30873 | 8.5 | 4.54 | 29 | 1.8 | 81.2 | 4.8 | 110 | 278.1 |
| İspanya | 33544 | 9 | 2.97 | 763 | 2.8 | 83.3 | 2.1 | 2886 | 322.6 |
| İsviçre | 61902 | 11.5 | 4.58 | 289 | 3.9 | 83.3 | 5.9 | 692 | 790.6 |
| Türkiye | 23983 | 4.3 | 2.68 | 1528 | 11.1 | 78 | 15 | 3288 | 92.9 |
| İngiltere | 41269 | 9.8 | 2.73 | 1568 | 3.9 | 81.4 | 6.7 | 2089 | 407.4 |
| ABD | 54993 | 16.4 | 2.83 | 5627 | 5.8 | 78.9 | 12.25 | 59340 | 398.9 |

**EK 3 Tablo 4** OECD Ülkeleri 2015 Yılı Sağlık Göstergeleri ve Verileri

| Ülkeler | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 |
|---|---|---|---|---|---|---|---|---|---|
| Avustralya | 47351 | 9.3 | 3.82 | 1331 | 3.2 | 82.5 | 2.6 | 3093 | 387.5 |
| Avusturya | 49955 | 10.4 | 7.54 | 278 | 3.1 | 81.3 | 4.7 | 791 | 375.8 |
| Kanada | 44671 | 10.6 | 2.61 | 719 | 4.5 | 81.9 | 7.1 | 2388 | 591.8 |
| Şili | 22593 | 8.3 | 2.14 | 363 | 6.9 | 79.9 | 15.5 | 774 | 109.2 |
| Çekya | 33701 | 7.2 | 6.67 | 256 | 2.5 | 78.7 | 6.3 | 593 | 325.5 |
| Danimarka | 49071 | 10.2 | 4.93 | 1321 | 3.7 | 80.8 | 1.7 | 516 | 572.8 |
| Estonya | 29444 | 6.4 | 4.96 | 30 | 2.5 | 77.7 | 0 | 63 | 223.6 |
| Finlandiya | 42528 | 9.7 | 4.35 | 268 | 1.7 | 81.6 | 3.6 | 544 | 451 |
| Fransa | 40841 | 11.5 | 6.13 | 3089 | 3.7 | 82.4 | 7.8 | 2525 | 541 |
| Almanya | 47684 | 11.1 | 8.13 | 3108 | 3.3 | 80.7 | 3.3 | 5613 | 479.9 |
| Macaristan | 26668 | 7 | 6.99 | 168 | 4.2 | 75.7 | 5.5 | 425 | 169.6 |
| İzlanda | 48857 | 8.1 | 3.12 | 8 | 2.2 | 82.5 | 0 | 31 | 559.2 |
| İrlanda | 69147 | 7.3 | 2.92 | 88 | 3.4 | 81.5 | 1.5 | 297 | 468.2 |
| İsrail | 35450 | 7.1 | 3.03 | 84 | 3.1 | 82.1 | 3.9 | 241 | 359 |
| İtalya | 36909 | 9 | 3.2 | 1115 | 2.9 | 82.6 | 3.4 | 6901 | 413 |
| Japonya | 40406 | 10.9 | 13.17 | 8480 | 1.9 | 83.9 | 4.4 | 26454 | 438.6 |
| Güney Kore | 35761 | 7 | 11.61 | 3678 | 2.7 | 82.1 | 8.7 | 6803 | 426.4 |
| Letonya | 24834 | 5.7 | 5.69 | 67 | 4.1 | 74.6 | 55.2 | 162 | 186 |
| Litvanya | 28824 | 6.5 | 6.97 | 95 | 4.2 | 74.5 | 9.5 | 168 | 359 |
| Lüksemburg | 103788 | 5.5 | 4.93 | 12 | 2.8 | 82.4 | 0 | 38 | 412.3 |
| Meksika | 18438 | 5.8 | 1.39 | 4456 | 12.5 | 75 | 34.6 | 2403 | 48.5 |
| Hollanda | 50302 | 10.3 | 3.52 | 517 | 3.3 | 81.6 | 3.5 | 648 | 292.2 |
| Yeni Zelanda | 37158 | 9.3 | 2.71 | 165 | 4.3 | 81.7 | 9.7 | 319 | 149.9 |
| Polonya | 26529 | 6.3 | 6.63 | 1067 | 4 | 77.6 | 1.6 | 1693 | 359 |
| Slovakya | 29932 | 6.8 | 5.75 | 134 | 5.1 | 76.7 | 1.8 | 349 | 283.1 |
| Slovenya | 31640 | 8.5 | 4.51 | 29 | 1.6 | 80.9 | 5 | 111 | 239.7 |
| İspanya | 34939 | 9.1 | 2.98 | 765 | 2.7 | 82.9 | 3.6 | 2937 | 272.3 |
| İsviçre | 63939 | 11.9 | 4.58 | 288 | 3.9 | 83 | 6.9 | 713 | 775.8 |
| Türkiye | 25728 | 4.1 | 2.68 | 1533 | 10.2 | 78 | 14.6 | 3389 | 86 |
| İngiltere | 42522 | 9.7 | 2.61 | 1882 | 3.9 | 81 | 4.5 | 2203 | 415 |
| ABD | 56770 | 16.7 | 2.8 | 5564 | 5.9 | 78.7 | 11.86 | 61173 | 359 |

# KAYNAKLAR

[1] Witten, I. H., Frank, E., Hall, M. A. (2005). Practical machine learning tools and techniques. Morgan Kaufmann, 578.

[2] Glymour, C., Madigan, D., Pregibon, D., Smyth, P. (1997). Statistical themes and lessons for data mining. Data mining and knowledge discovery, 1(1), 11-28.

[3] Gorunescu, F. (2011). Data Mining: Concepts, models and techniques (Vol. 12). Springer Science & Business Media.

[4] Piateski, G., Frawley, W. (1991). Knowledge discovery in databases. MIT press.

[5] Nasereddin, H. H. (2011). Stream Data Mining. International Journal of Web Application, 3(2), 90-97.

[6] Altındiş, S. (2018). Büyük Verinin Sağlık Hizmetleri Kalitesindeki Rolü. Sakarya Tıp Dergisi, 8(2), 205-213.

[7] Klugman, J. (2010). Human development report 2010. United Nations Development Programme, Palgrave Macmillan.

[8] Maddison, A. (2010). Historical Statistics of World Economy: 1-2008 AD. Paris: Organization for Economic Cooperation and Development.

[9] Riley, J.C. (2005). Poverty and life expctancy. Cambridge, UK: Cambridge University Press.

[10] Jahan, S (2016). Human Development Report, Technical Notes 2016.

[11] Alijanzadeh, M., Asefzadeh, S., Zare, M., Ali, S. (2016). Correlation between human development index and infant mortality rate worldwide. Biotechnology and Health Sciences, 3(1).

[12] Bilas, V., Franc, S., Bošnjak, M. (2014). Determinant factors of life expctancy at birth in the European Union countries. Collegium antropologicum, 38(1), 1-9.

[13] Potter, L. B. (1991). Socioeconomic determinants of white and black males' life expctancy differentials, 1980. Demography, 28(2), 303-321.

[14] Hassan, F. A., Minato, N., Ishida, S., Nor, N. M. (2016). Social environment determinants of life expctancy in developing countries: a panel data analysis. Global Journal of Health Science, 9(5), 105.

[15] Delavari, S., Zandian, H., Rezaei, S., Moradinazar, M., Delavari, S., Saber, A., Fallah, R. (2016). life expctancy and its Socioeconomic Determinants in Iran. Electronic physician, 8(10), 3062.

[16] Lin, R. T., Chen, Y. M., Chien, L. C., Chan, C. C. (2012). Political and social determinants of life expctancy in less developed countries: a longitudinal study. BMC Public Health, 12(1), 85.

[17] Balan, C., Jaba, E. (2011). Statistical analysis of the determinants of life expctancy in Romania. Romanian Journal of Regional Science, 5(2), 25-38.

[18] Bayın, G. (2016). Doğuşta ve ileri yaşta beklenen yaşam sürelerine etki eden faktörlerin belirlenmesi. Türkiye Aile Hekimliği Dergisi, 20(3), 93-103.

[19] Gilligan, A. M., Skrepnek, G. H. (2014). Determinants of life expctancy in the Eastern Mediterranean Region. Health policy and planning, 30(5), 624-637.

[20] Folland, S., Goodman, A. C., Stano, M. (2007). The economics of health and health care (Vol. 6). Upper Saddle River, NJ: Pearson Prentice Hall.

[21] Mathers, C. D., Stevens, G. A., Boerma, T., White, R. A., Tobias, M. I. (2015). Causes of international increases in older age life expectancy. The Lancet, 385(9967), 540-548.

[22] Akın, A., Ersoy, K. (2012). 2050'ye doğru nüfusbilim ve yönetim: sağlık sistemine bakış. TÜSİAD.

[23] Karabulut, K. (2010). Sağlık Harcamaları Ve Göstergelerinin Karşılaştırılması. Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi, 13(1).

[24] Remuzzi, A., Remuzzi, G. (2020). COVID-19 and Italy: what next?. The Lancet.

[25] Greer, S. L., Greer, S. L. (2009). The politics of European Union health policies. McGraw-Hill Education (UK).

[26] Koçak, O., Tiryaki, D. (2011). Sosyal devlet anlayışında sağlık politikalarının önemi ve sağlıkta dönüşüm programının değerlendirilmesi: Yalova örneği.

[27] Proksch, D., Busch-Casler, J., Haberstroh, M. M., Pinkwart, A. (2019). National health innovation systems: Clustering the OECD countries by innovative output in healthcare using a multi indicator approach. Research Policy, 48(1), 169-179.

[28] Karakış, E., GÜLDEN, T. (2019). OECD Ülkelerinin Ekonomik Özgürlüklerine Göre Kümeleme Analizi ile Sınıflandırılması. Cumhuriyet Üniversitesi İktisadi ve İdari Bilimler Dergisi, 20(2), 297-316.

[29] KANGALLI, S. G., Uyar, U., Buyrukoğlu, S. (2014). OECD Ülkelerinde Ekonomik Özgürlük: Bir Kümeleme Analizi. Journal of Alanya Faculty of Business/Alanya Isletme Fakültesi Dergisi, 6(3).

[30] Mylevaganam, S. (2017). The Analysis of Human Development Index (HDI) for Categorizing the Member States of the United Nations (UN). Open Journal of Applied Sciences, 7(12), 661-690.

[31] YAKUT, E., GUNDUZ, M., DEMİRCİ, A. (1999). Comparison of classification success of Human Development Index by using ordered logistic regression analysis and artificial neural network methods. Journal of Applied Quantitative Methods, 159, 15.

[32] do Carvalhal Monteiro, R. L., Pereira, V., Costa, H. G. (2018). A multicriteria approach to the human development index classification. Social Indicators Research, 136(2), 417-438.

[33] Burmaoğlu, S., Oktay, E. (2015). A Statistical Classification Study of Countries' Human Development Level by Discriminant Analysis. Ataturk University Journal of Economics & Administrative Sciences, 29(4).

[34] RENÇBER, Ö. F., METE S. (2018). Reclassification Of Countries According To Human Development Index: An Application With Ann And Anfis Methods. Business & Management Studies: An International Journal, 6(3), 228-252.

[35] Fish, K. E., Barnes, J. H., AikenAssistant, M. W. (1995). Artificial neural networks: a new methodology for industrial market segmentation. Industrial Marketing Management, 24(5), 431-438.

[36] Schumacher, M., Roßner, R., Vach, W. (1996). Neural networks and logistic regression: Part I. Computational Statistics & Data Analysis, 21(6), 661-682.

[37] Zhang, G., Hu, M. Y., Patuwo, B. E., Indro, D. C. (1999). Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. European journal of operational research, 116(1), 16-32.

[38] Yi, H. C., You, Z. H., Zhou, X., Cheng, L., Li, X., Jiang, T. H., Chen, Z. H. (2019). ACP-DL: a deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation. Molecular Therapy-Nucleic Acids, 17, 1-9.

[39] Alam, S., Dobbie, G., Koh, Y. S., Riddle, P., Rehman, S. U. (2014). Research on particle swarm optimization based clustering: a systematic review of literature and techniques. Swarm and Evolutionary Computation, 17, 1-13.

[40] Permai, S. D., Tanty, H., Rahayu, A. (2016, October). Geographically weighted regression analysis for human development index. In AIP Conference Proceedings (Vol. 1775, No. 1, p. 030045). AIP Publishing LLC.

[41] Ottenbacher, K. J., Linn, R. T., Smith, P. M., Illig, S. B., Mancuso, M., Granger, C. V. (2004). Comparison of logistic regression and neural network analysis applied to predicting living setting after hip fracture. Annals of epidemiology, 14(8), 551-559.]

[42] Dreiseitl, S., Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. Journal of biomedical informatics, 35(5-6), 352-359.

[43] Dreiseitl, S., Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. Journal of biomedical informatics, 35(5-6), 352-359.

[44] Radhimeenakshi, S. (2016, March). Classification and prediction of heart disease risk using data mining techniques of Support Vector Machine and Artificial Neural Network. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 3107-3111). IEEE.

[45] Saritas, M. M., Yasar, A. (2019). Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification. International Journal of Intelligent Systems and Applications in Engineering, 7(2), 88-91.

[46] Kharya, S., Agrawal, S., Soni, S. (2014). Naive Bayes classifiers: A probabilistic detection model for breast cancer. International Journal of Computer Applications, 92(10), 0975-8887.

[47] Yang, Y., Webb, G. I. (2002, August). A comparative study of discretization methods for naive-bayes classifiers. In Proceedings of PKAW (Vol. 2002).

[48] Ng, A. Y., Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In Advances in neural information processing systems (pp. 841-848).

[49] Halloran, J. (2009). Classification: Naive Bayes vs logistic regression. In Technical report. University of Hawaii.

[50] Wright, R. E. (1995). Logistic regression.

[51] Lin, Y. P., Chu, H. J., Wu, C. F., Verburg, P. H. (2011). Predictive ability of logistic regression, auto-logistic regression and neural network models in empirical land-use change modeling–a case study. International Journal of Geographical Information Science, 25(1), 65-87.

[52] Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of clinical epidemiology, 49(11), 1225-1231.

[53] Xue, J. H., Titterington, D. M. (2008). Comment on "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes". Neural processing letters, 28(3), 169.

[54] Caruana, R., Niculescu-Mizil, A. (2006, June). An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning (pp. 161-168).

[55] Han, J., Pei, J., Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[56] Drineas, P., Mahoney, M. W., Muthukrishnan, S., Sarlós, T. (2011). Faster least squares approximation. Numerische mathematik, 117(2), 219-249.

[57] Zielesny, A. (2011). From Curve Fitting to Machine Learning (Vol. 18). Springer.

[58] Alasadi, S. A., Bhaya, W. S. (2017). Review of data preprocessing techniques in data mining. Journal of Engineering and Applied Sciences, 12(16), 4102-4107.

[59] García, S., Luengo, J., Herrera, F. (2015). Data preprocessing in data mining (pp. 195-243). Cham, Switzerland: Springer International Publishing.

[60] Jain, A., Nandakumar, K., Ross, A. (2005). Score normalization in multimodal biometric systems. Pattern recognition, 38(12), 2270-2285.

[61] Forman, G. 2003. An Extensive Empirical Study of Feature Selection Metrics for Text Classification, Journal of Machine Learning Research, 3, 1289–1305.

[62] Ladha, L., Deepa, T. 2011. Feature Selection Methods And Algorithms, International Journal on Computer Science and Engineering, 3(5), 1787-1797.

[63] Cichosz, P. (2014). Data mining algorithms: explained using R. John Wiley & Sons.

[64] Bhosale, D., Ade, R. (2014). Feature selection based classification using naive bayes, j48 and support vector machine. International Journal of Computer Applications, 99(16), 14-18.

[65] Ahmed, A. B. E. D., Elaraby, I. S. (2014). Data mining: A prediction for student's performance using classification method. World Journal of Computer Application and Technology, 2(2), 43-47.

[66] Inselberg, A. and Dimsdale, B. (1990). Parallel coordinates: A tool for visualizing multidimensional geometry. In VIS ’90: Proceedings of the 1st conference on visualization, pages 361–378. Los Alamitos, CA: IEEE Computer Society.

[67] Fua, Y., Ward, M., and Rundensteiner, E. (1999). Hierarchical parallel coordinates for exploration of large datasets. In VIS ’99: Proceedings of the conference on Visualization’99, pages 43–50. Los Alamitos, CA: IEEE Computer Society Press.

[68] Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., Tatham, R. L. (1998). Multivariate data analysis (Vol. 5, No. 3, pp. 207-219). Upper Saddle River, NJ: Prentice hall

[69] Abonyi, J. ve Feil, B. (2007), Cluster Analysis For Data Mining And System Identification, 1. Baskı, Berlin: Birkhauser Verlag AG.

[70] Xu, R., Wunsch, D. (2008). Clustering (Vol. 10). John Wiley & Sons.

[71] Wu, W., Xiong, H., Shekhar, S. (2013). Clustering and information retrieval (Vol. 11). Springer Science & Business Media.

[72] Sharma, S. (1996). Applied Multivariate Techniques, New York: John Wiley and Sons Inc..

[73] Kaufman, L., Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis (Vol. 344). John Wiley & Sons

[74] Kaur, M., Garg, S. K. (2014). Survey on clustering techniques in data mining for software engineering. International Journal of Advanced and Innovative Research, 3, 238-243.

[75] He, J., Zhao, G., Zhang, H. L., Ramamohanarao, K., Pang, C. (2014, December). An effective clustering algorithm for auto-detecting well-separated clusters. In 2014 IEEE International Conference on Data Mining Workshop (pp. 867-874). IEEE.

[76] Kaur, M., Garg, S. K. (2014). Survey on clustering techniques in data mining for software engineering. International Journal of Advanced and Innovative Research, 3, 238-243.

[77] Gan, G., Ma, C., Wu, J. (2007). Data clustering: theory, algorithms, and applications (Vol. 20). Siam.

[78] Masciari, E., Mazzeo, G. M., Zaniolo, C. (2013, April). A new, fast and accurate algorithm for hierarchical clustering on euclidean distances. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 111-122). Springer, Berlin, Heidelberg.

[79] Liu, H., Motoda, H. (2012). Feature selection for knowledge discovery and data mining (Vol. 454). Springer Science & Business Media.

[80] Li, X., Zaiane, O. R., Li, Z. (2006, January). Advanced data mining and applications. In Proceedings of Second International Conference, ADMA (pp. 14-16).

[81] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

[82] Zhao, Q. (2012). Cluster validity in clustering methods. Publications of the University of Eastern Finland.

[83] Everitt, B. S., Landau, S., Leese, M., Stahl, D. (2011). Cluster Analysis. John Wiley & Sons, Ltd. 5. Baskı.

[84] Gan, G., Ma, C., Wu, J. (2007). Data clustering: theory, algorithms, and applications (Vol. 20). Siam.

[85] Stuetzle, W., Nugent, R. (2010). A generalized single linkage method for estimating the cluster tree of a density. Journal of Computational and Graphical Statistics, 19(2), 397-418.

[86] Murtagh, F., Contreras, P. (2017). Algorithms for hierarchical clustering: an overview, II. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7(6), e1219.

[87] Sokal, R. R., Michener, C. D. (1958). A statistical method for evaluating systematic relationships. University of Kansas Scientific Bulletin, 38, 1409–1438.

[88] Ferreira, L., Hitchcock, D. B. (2009). A comparison of hierarchical methods for clustering functional data. Communications in Statistics-Simulation and Computation, 38(9), 1925-1949.

[89] MacQueen, J. (1965, January). On convergence of k-means and partitions with minimum average variance. In Annals of mathematical statistics (Vol. 36, No. 3, p. 1084).

[90] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.

[91] Reddy, C. K., Vinzamuri, B. (2018). A survey of partitional and hierarchical clustering algorithms. In Data Clustering (pp. 87-110). Chapman and Hall/CRC.

[92] Hothorn, T., Everitt, B. S. (2014). A handbook of statistical analyses using R. CRC press.

[93] Vathy-Fogarassy, Á., Abonyi, J. (2013). Graph-based clustering and data visualization algorithms. London: Springer.

[94] Steinley, D. (2008). Stability analysis in K-means clustering. British Journal of Mathematical and Statistical Psychology, 61(2), 255-273.

[95] B. G. Mirkin (2005). Clustering for Data Mining: A Data Recovery Approach, volume 3. CRC Press, Boca Raton, FL.

[96] Park, H. S., Jun, C. H. (2009). A simple and fast algorithm for K-medoids clustering. Expert systems with applications, 36(2), 3336-3341.

[97] Aggarwal, C. C., Reddy, C. K. (2014). Data clustering. Algorithms and applications. Chapman&Hall/CRC Data mining and Knowledge Discovery series, Londra.

[98] Dinçer, Ş. E. (2006). Veri madenciliğinde K-means algoritması ve tıp alanında uygulanması (Master's thesis, Kocaeli Universitesi, Fen Bilimleri Enstitusu).

[99] Bhuvaneswari, R., Kalaiselvi, K. (2012). Naive Bayesian classification approach in healthcare applications. International Journal of Computer Science and Telecommunications, 3(1), 106-112.

[100] Ersöz, F. (2015). Veri Madenciliği Teknikleri ve Uygulamaları. Dijital Basımevi, 72.

[101] Ahire, J. (2018). Artificial Neural Networks: the Brain behind AI. Lulu. com.

[102] Ciaburro, G., Venkateswaran, B. (2017). Neural networks with R: smart models using CNN, RNN, deep learning, and artificial intelligence principles.

[103] Adisa, J. A., Ojo, S. O., Owolawi, P. A., Pretorius, A. B. (2019, November). Financial Distress Prediction: Principle Component Analysis and Artificial Neural Networks. In 2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC) (pp. 1-6). IEEE.

[104] Du, K. L., Swamy, M. N. (2013). Neural networks and statistical learning. Springer Science & Business Media.

[105] Aydemir, E. (2018). Weka ile Yapay Zeka, Seçkin Yayınevi, Ankara.

[106] Langfelder, P., Zhang, B., Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics, 24(5), 719-720.

[107] Ghatak, A. (2019). Deep Learning with R (pp. 1-245). Springer.

[108] Hosner, D. W., Lemeshow, S. (1989). Applied logistic regression. New York: Jhon Wiley & Son.

[109] Gülpınar, V. (2013). Yapay sinir ağları işleyişinin sosyal ağ analizi yardımı ile çözümlenmesi.

[110] Priddy, K. L., Keller, P. E. (2005). Artificial neural networks: an introduction (Vol. 68). SPIE press.

[111] Ataseven, B. (2013). Yapay sinir ağları ile öngörü modellemesi.

[112] Derksen, S., Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. British Journal of Mathematical and Statistical Psychology, 45(2), 265-282.

[113] Gamgam, H., Altunkaynak, B. (2017). SPSS uygulamalı regresyon analizi (lojistik regresyon eğri uydurma-tahmin)(2. Baskı). Ankara: Seçkin Yayıncılık.

[114] Demirtas E. A., Anagun A.S., Koksal G., (2009). Determination Of Optimal Product Styles By Ordinal Logistic Regression Versus Conjoint Analysis For Kitchen Faucets. International Journal of Industrial Ergonomics, 39, 866–875

[115] Bakır, B., Batmaz, I., Güntürkün, F. A., İpekçi, İ. A., Köksal, G., Özdemirel, N. E. (2006). Defect cause modeling with decision tree and regression analysis. World Acad Sci Eng Technoly, 24, 1-4.

[116] Berenson, M. L. ve Levine, D. M. (1996), Basic Business Statistics: Concepts and Applications, Sixth Edition, Prentice-Hall International, New Jersey

[117] Coenen, F. (2011). Data mining: past, present and future. The Knowledge Engineering Review, 26(1), 25-29.

[118] EMC Education Services. (2015). Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. Wiley.

[119] Burduk, R., Trajdos, P. (2013, September). Construction of sequential classifier using confusion matrix. In IFIP International Conference on Computer Information Systems and Industrial Management (pp. 401-407). Springer, Berlin, Heidelberg.

[120] Kumar, A., Paul, A. (2016). Mastering text mining with R. Packt Publishing Ltd.

# ÖZGEÇMİŞ

**Adı Soyadı** : Kübra ÇOŞAR

**Doğum Yeri ve Tarihi** : Fatih/İSTANBUL, 04.02.1992

**Ana Dili** : Türkçe

**Yabancı Dili** : İngilizce

**E-Posta** : kubracosar34@gmail.com

**Öğrenim Durumu**

| Derece | Bölüm/Program | Üniversite/Lise | Mezuniyet Yılı |
|---|---|---|---|
| Lise | Bilişim Tekolojileri / Veritabanı Dalı | Selçuk Anadolu Meslek Lisesi | 2010 |
| Lisans | Yazılım Mühendisliği | Fırat Üniversitesi | 2016 |
| Yüksek Lisans | Bilgisayar Mühendisliği | Marmara Üniversitesi | 2020 |

**İş Deneyimi**

| Yıl | Firma/Kurum | Görevi |
|---|---|---|
| 2018 - | İ2i Bilişim Danışmanlık Teknoloji Hiz. ve Paz. Tic. A.Ş. | Yazılım Mühendisi |