

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

**GELİR VE YAŞAM KOŞULLARI VERİSİNİN VERİ
MADENCİLİĞİ YAKLAŞIMI İLE İNCELENMESİ VE
MODEL PERFORMANSLARININ DEĞERLENDİRİLMESİ**

Olgun ÖZDEMİR

DOKTORA TEZİ

İstatistik Anabilim Dalı

İstatistik Programı

Danışman

Doç. Dr. İbrahim DEMİR

Şubat, 2020

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

**GELİR VE YAŞAM KOŞULLARI VERİSİNİN VERİ
MADENCİLİĞİ YAKLAŞIMI İLE İNCELENMESİ VE MODEL
PERFORMANSLARININ DEĞERLENDİRİLMESİ**

Olgun ÖZDEMİR tarafından hazırlanan tez çalışması 17.02.2020 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı, İstatistik Programı **DOKTORA TEZİ** olarak kabul edilmiştir.

Doç. Dr. İbrahim DEMİR
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Doç. Dr. İbrahim DEMİR, Danışman
Yıldız Teknik Üniversitesi

Prof. Dr. Ali Hakan BÜYÜKLÜ, Üye
Yıldız Teknik Üniversitesi

Prof. Dr. Adnan MAZMANOĞLU, Üye
Yıldız Teknik Üniversitesi

Doç. Dr. Reşit ÇELİK, Üye
Yıldız Teknik Üniversitesi

Doç. Dr. Ozan KOCADAĞLI, Üye
Mimar Sinan Güzel Sanatlar Üniversitesi

Danışmanım Doç. Dr. İbrahim DEMİR sorumluluğunda tarafımca hazırlanan Gelir ve Yaşam Koşulları Verisinin Veri Madenciliği Yaklaşımı ile İncelenmesi ve Model Performanslarının Değerlendirilmesi başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Olgun ÖZDEMİR

İmza



Aileme

TEŞEKKÜR

Bu tez çalışması süresince bilgi ve tecrübeleriyle beni yönlendiren değerli hocam Doç. Dr. İbrahim DEMİR'e teşekkürlerimi sunarım. Çalışma süresince katkı ve desteklerini esirgemeyen sevgili arkadaşlarım Türkiye İstatistik Kurumu Uzmanı Emrah Güler ve Türkiye İstatistik Kurumu Uzmanı Zafer Ağyar'a teşekkürü bir borç bilirim. Maddi ve manevi desteklerinden dolayı her zaman yanımda olan aileme sabır ve anlayışlarından dolayı sonsuz şükranlarımı sunarım.

Olgun ÖZDEMİR



İÇİNDEKİLER

SİMGE LİSTESİ	viii
KISALTMA LİSTESİ	ix
ŞEKİL LİSTESİ	x
TABLO LİSTESİ	xi
ÖZET	xiii
ABSTRACT	xv
1 GİRİŞ	1
1.1 Literatür Özeti.....	1
1.2 Tezin Amacı	10
1.3 Orjinal Katkı.....	11
2 VERİ MADENCİLİĞİ	13
2.1 Veri Madenciliği Kavramı	13
2.2 Veri Madenciliği Tarihçesi	14
2.3 Veri Madenciliği Süreci	14
2.4 Veri Madenciliği Modelleri	16
2.4.1 Tahmin Edici Modeller.....	16
2.4.2 Tanımlayıcı Modeller	17
2.5 Model Değerlendirme Ölçütleri	17
2.5.1 Doğruluk Oranı	18
2.5.2 Duyarlılık	19
2.5.3 Kesinlik	19
2.5.4 F-ölçütü.....	19
2.5.5 Kappa Katsayısı	19
2.5.6 Akaike Bilgi Kriteri (AIC)	20
2.5.7 Bayes Bilgi Kriteri (BIC)	20

2.5.8 Entropi.....	20
3 MATERYAL VE YÖNTEM	22
3.1 Çalışmada Kullanılan Veri Seti ve Özellikleri.....	22
3.2 Veri Ön İşleme.....	22
3.3 Kullanılan Yöntemler	26
3.4 Değişken Seçimi.....	26
3.5 Modelleme	33
3.5.1 Gizli Sınıf Analizi.....	34
3.5.2 k-modlar Kümeleme Algoritması	36
3.5.3 Rastgele Orman (Random Forest)	36
4 UYGULAMA	40
4.1 Gizli Sınıf Analizi ile Sonuçlar	40
4.2 k-modlar Algoritması ile Sonuçlar.....	43
4.3 RO Algoritması ile Sonuçlar	48
4.3.1 Eğitim Verisi için RO Modeli	49
4.3.2 Eğitim Verisi için Hata Matrisi ve Tahmin	49
4.3.3 Test Verisi için Hata Matrisi ve Tahmin.....	50
4.3.4 Rastgele Ormanın Hata Oranı	51
4.3.5 RO Modelinin İyileştirilmesi (Tune işlemi).....	52
4.3.6 Ağaçlar için Düğüm (Node) Sayısı.....	55
4.3.7 Değişken Önemliliği.....	56
4.3.8 Kısmi Bağımlılık Grafikleri	57
4.3.9 RO Modelinden Tek Bir Ağacın Seçilerek Elde Edilmesi.....	60
5 SONUÇ VE SONUÇLAR	62
KAYNAKÇA	68
A GYKA ANKETİ	77
A.1 GYKA Hakkında Genel Açıklama.....	77
A.2 GYKA Fert Kesit Veri Seti Yapısı.....	85

A.3 GYKA Fert Kayıt Kesit Veri Seti Yapısı	99
A.4 GYKA Hane Kesit Veri Seti Yapısı.....	101
B ALGORİTMA SONUÇLARI VE R KODLARI	111
B.1 Gizli Sınıf Olasılıkları	111
B.2 10 Sınıflı GSA Model Çözümünde Gizli Sınıf Tanımları.....	114
B.3 10 Kümeli k-modlar Model Çözümünde Küme Tanımları.....	120
B.4 RO Algoritması R Kodu ve Çıktıları.....	125
B.5 Boruta Algoritması ile Değişken Seçim İşlemi ve R Kodu.....	144
B.6 LCAvarsel Algoritması ile Değişken Seçim İşlemi ve R Kodu.....	151
TEZDEN İZLENİLMİŞ YAYINLAR	156



SİMGE LİSTESİ

P_e	Beklenen Sınıfların Oranı
π	Çarpım Sembolü
η	Denklem Sabiti
$\ln$	Doğal Logaritma
S	Entropi
$\forall$	Evrensel Niceleyici
τ	Gözlemin Gizli Sınıfta Bulunma Olasılığı
λ	Gözlemin Gizli Sınıfta Bulunma Olasılığı (Logaritmik Olasılık)
P_0	Gözlenen Sınıfların Oranı
κ	Kappa Sabiti
$\log$	Logaritma
$\hat{L}$	Log-benzerlik Değeri
L	Maksimum Olabilirlik Değeri
n_{min}	Minimum Düğüm Sayısı
Z	Mod Vektörü Kümesi
p_i	Olasılık Değeri
n	Örneklem Büyüklüğü
k	Parametre Sayısı
X	Rastgele Değişken
$\hat{C}$	Rastgele Orman Ağacının Sınıf Tahmini
T_b	Rastgele Orman Ağaçları Topluluğu
$\hat{f}$	Rastgele Orman Regresyon Tahmin Fonksiyonu
Σ	Toplam Sembolü
d	Uzaklık Fonksiyonu

KISALTMA LİSTESİ

AIC	Akaike information criterion
BIC	Bayesian information criterion
CHAID	Chi-squared automatic interaction detection
CART	Classification and regression tree
CAPI	Computer assisted personel interview
CRISP-DM	Cross industry standard process for data mining
EU-SILC	European Union statistics on income and living conditions
GYKA	Gelir ve yaşam koşulları araştırması
GSA	Gizli sınıf analizi
GSM	Global system for mobile communications
HİA	Hanehalkı işgücü araştırması
İMKB	İstanbul menkul kıymetler borsası
İBBS	İstatistiki bölge birimleri sınıflandırması
KOBİ	Küçük ve orta büyüklükteki işletmeler
MRT	Multivariate regression trees
NASS	National Agricultural Statistics Service
NSI	National Statistics Institutes
OOB	Out of bag
RO	Rastgele orman
SILC	Survey on income and living conditions
TÜİK	Türkiye İstatistik Kurumu
UİE	Ulusal istatistik enstitüleri
VM	Veri madenciliği
VTBK	Veri tabanlarında bilgi keşfi

ŞEKİL LİSTESİ

Şekil 2.1	Veri madenciliğinin diğer alan ve disiplinler ile ilişkisi	13
Şekil 2.2	Veri tabanlarında bilgi keşfi süreci (VTBK) ([92])	14
Şekil 2.3	CRISP-DM VM süreci ([93])	15
Şekil 3.1	RO ağaç yapısı ve işleyişi ([109])	38
Şekil 4.1	Seçilmiş 10 değişken ile GSA çıktıları	41
Şekil 4.2	k-modlar kümeleme analizi sonrası GYKA verisi için küme grafikleri.....	47
Şekil 4.3	AGE ve IG değişkenleri arasındaki verinin kümelere göre dağılımı	47
Şekil 4.4	RGN ve SEC değişkenleri arasındaki verinin kümelere göre dağılımı	48
Şekil 4.5	SEC ve IG değişkenleri arasındaki verinin kümelere göre dağılımı	48
Şekil 4.6	RO hata oranı.....	52
Şekil 4.7	En uygun m_{try} değerinin belirlenmesi.....	53
Şekil 4.8	Ağaçlar için düğüm sayıları.....	55
Şekil 4.9	Değişken önemliliği.....	56
Şekil 4.10	“EDU” değişkeni için sınıf 1 bazında kısmi bağımlılık	58
Şekil 4.11	“OCC” değişkeni için sınıf 1 bazında kısmi bağımlılık	58
Şekil 4.12	“EDU” değişkeni için sınıf 5 bazında kısmi bağımlılık	59
Şekil 4.13	“OCC” değişkeni için sınıf 5 bazında kısmi bağımlılık.....	59
Şekil 4.14	Tek bir ağaca ilişkin RO ağaç çıktısı.....	60
Şekil 4.15	RO modelinin ilk ağacı	61
Şekil 4.16	RO modelinin ikinci ağacı.....	61
Şekil 4.17	RO modelinin üçüncü ağacı	61
Şekil 4.18	RO modelinin dördüncü ağacı.....	61
Şekil B.1	Boruta ile elde edilen değişken önemliliği grafiği.....	148

TABLO LİSTESİ

Tablo 2.1	Hata matrisi örneği.....	18
Tablo 3.1	Fert geliri için 2015 yılına ilişkin ağırlıklı kantiller	24
Tablo 3.2	Excel programında basamak düzeltme için yapılan ayarlar.....	25
Tablo 3.3	Fert gelirine yönelik seçilen başlangıç değişkenleri.....	27
Tablo 3.4	GSA, k-modlar ve RO uygulamalarında kullanılan nihai veri setine ilişkin değişkenleri	30
Tablo 3.5	Değişken tipine göre gizli değişken modelleri	35
Tablo 4.1	Seçilmiş 10 değişken ile GSA'ya ilişkin komutlar.....	40
Tablo 4.2	GSA'nın sınıflara göre model uyum istatistikleri.....	41
Tablo 4.3	GSA sonucu elde edilen gizli sınıflar.....	42
Tablo 4.4	Seçilmiş 10 değişken ile k-modlar analizine ilişkin komutlar.....	43
Tablo 4.5	Küme büyüklükleri	44
Tablo 4.6	k-modlar kümelemesinin küme modları.....	44
Tablo 4.7	Kümeler için basit-eşleşme uzaklıkları	45
Tablo 4.8	Her kümede gelir grubuna göre fert sayıları.....	45
Tablo 4.9	Eğitim verisinin RO modelinde hata matrisi.....	49
Tablo 4.10	Eğitim verisinin tahmini için hata matrisi	49
Tablo 4.11	Test verisi tahmini için hata matrisi	50
Tablo 4.12	Test verisi sınıfları için duyarlılık değerleri	51
Tablo 4.13	Eğitim verisi için iyileştirilmiş modelin hata matrisi	53
Tablo 4.14	Eğitim verisinin tahmini için iyileştirilmiş modelin hata matrisi....	54
Tablo 4.15	Test verisi için iyileştirilmiş modelin hata matrisi.....	54
Tablo 4.16	Test verisi sınıfları için iyileştirilmiş modelin duyarlılık değerleri..	55
Tablo 4.17	Gini indeksindeki ortalama düşüş ve bulunma sıklığı	57
Tablo A.1	GYKA fert kesit veri seti değişkenleri ([112]).....	85
Tablo A.2	GYKA fert kayıt kesit veri seti değişkenleri ([112])	99
Tablo A.3	GYKA hane kesit veri seti değişkenleri ([112]).....	101
Tablo B.1	Değişken değerlerine ilişkin gizli sınıf olasılıkları	111
Tablo B.2	GSA model çözümünde gizli sınıf tanımları	114
Tablo B.3	k-modlar model çözümünde küme tanımları.....	120

Tablo B.4	Eğitim verisinin RO modeli için hata matrisi çıktısı	125
Tablo B.5	Eğitim verisinin tahmini için hata matrisi çıktısı.....	126
Tablo B.6	Eğitim verisinin tahmini için model istatistikleri çıktısı.....	127
Tablo B.7	Test verisi tahmini için hata matrisi çıktısı.....	128
Tablo B.8	Test verisi tahmini için model istatistikleri çıktısı.....	128
Tablo B.9	Eğitim verisi için iyileştirilmiş modelin hata matrisi çıktısı.....	130
Tablo B.10	Eğitim verisinin tahmini için iyileştirilmiş modelin hata matrisi çıktısı	131
Tablo B.11	Eğitim verisinin tahmini için iyileştirilmiş modelin model istatistikleri çıktısı.....	131
Tablo B.12	Test verisi için iyileştirilmiş modelin hata matrisi çıktısı.....	132
Tablo B.13	Test verisi için iyileştirilmiş modelin model istatistikleri çıktısı ...	133
Tablo B.14	Modelden seçilen tek bir ağaca ilişkin RO ağaç çıktısı	134
Tablo B.15	Boruta ile elde edilen onaylanmış değişkenlere ilişkin istatistikler.....	149
Tablo B.16	LCAvarsel ile değişken seçim işleminde swap algoritması sonuçları.....	152
Tablo B.17	LCAvarsel ile değişken seçim işleminde stepwise algoritması sonuçları.....	154

Gelir ve Yaşam Koşulları Verisinin Veri Madenciliği Yaklaşımı ile İncelenmesi ve Model Performanslarının Değerlendirilmesi

Olgun ÖZDEMİR

İstatistik Anabilim Dalı

Doktora Tezi

Danışman: Doç. Dr. İbrahim DEMİR

Ulusal İstatistik Enstitüleri (ÜİE) tarafından üretilen resmi veriler devletin ekonomik ve sosyal karar alma sürecinde önemli bir rol oynamaktadır. Resmi verilerin geleneksel istatistiksel yaklaşımlardan ziyade veri madenciliği (VM) yöntemleri ile ele alınması, yeni bilgiler ve gizli kalıpların ortaya çıkarılması bakımından önem kazanmaktadır. Resmi istatistikler için VM yöntemleri kullanışlı olmakla birlikte hala yeni VM yöntemlerinin keşfedilmesi devam etmektedir. Bu çalışmada, Türkiye İstatistik Kurumu (TÜİK) tarafından yürütülen 2015 yılı Gelir ve Yaşam Koşulları Araştırması (GYKA) verileri VM yöntemleri ile incelenmiştir. 36036 adet ferde ilişkin yatay kesit verileri ele alınmış olup fert gelirini en çok etkileyen değişkenler belirlenerek fertlerin refah durumu incelenmiştir. Fertlerin sosyoekonomik profillerinin belirlenmesi amacıyla gizli sınıf analizi (GSA) ve k-modlar kümeleme analizi kullanılmıştır. Bireylerin sosyoekonomik durumu kümeleme ve rastgele orman (RO) algoritma modelleri kullanılarak sınıflandırılmıştır. 10 sınıflı GSA modelinde yeni seçilen bir ferdin hangi olasılıkla hangi sınıfa dahil olacağı elde edilmiştir. Elde edilen gizli sınıfların en yüksek

olasılıkla aldıkları deęişken deęerlerine göre fertlerin gizli sınıf profil tanımları elde edilmiştir. k-modlar kümelemesi sonucu elde edilen 10 adet küme, küme modlarına göre tanımlanmış ve fertlerin küme profil tanımları elde edilerek sonuçları GSA sonuçlarıyla karşılaştırılmıştır. Kategorik deęişkenlerin ele alındığı bu çalışmada, GSA yönteminin k-modlar kümeleme yöntemine göre daha tutarlı sonuçlar sağladığı görölmüştür. Fert gelirinin dięer tüm dokuz adet girdi deęişkeninin fonksiyonu olarak seçildiğı RO modelinde deęişkenlerin önemlilikleri belirlenmiştir. Sırasıyla eğitim, meslek ve yaş deęişkenlerinin daha önemli olduğı ve RO modeline en fazla katkıyı sağladığı gözlenmiştir. Oldukça kapsamlı ve detaylı bir veri olan GYKA verisinde, VM yöntemlerinin uygulanabilmesi ve veriden anlamlı sonuçlar çıkarılması bakımından GSA ve RO gibi yöntemler uygun görünmektedir. Benzer VM süreçleri farklı resmi veriler için de anlamlı sonuçlar elde etmek amacıyla kullanılabilir.

Bu çalışmada ifade edilen görüş ve yorumlar hazırlayanın kendisine ait olup, TÜİK'i bağlamaz. Bu teze dayalı tüm çalışmalar için de aynı kural geçerlidir.

Anahtar Kelimeler: Veri madenciliğı, GYKA, gizli sınıf analizi, k-modlar, rastgele ormanlar

The Examination and The Model Performance Evaluation of The Income and Living Conditions Data within Data Mining Approach

Olgun ÖZDEMİR

Department of Statistics

Doctor of Philosophy Thesis

Advisor: Assoc. Prof. Dr. İbrahim DEMİR

Official data that are generated by the National Statistical Institutes (NSIs) has an important role to play in the economic and social decision-making phase of government. Tackling official data using data mining (DM) techniques is key to discovering new information and hidden patterns rather than traditional statistical approaches. While DM techniques are useful in the field of official statistics, the discovery of new DM techniques remains. Data from the 2015 Turkish Statistical Institute (TurkStat) Survey of Income and Living Conditions (SILC) were analyzed using DM techniques in this study. Cross-sectional data were processed for 36036 individuals and the variables influencing the individual income as well as individual welfare status were identified. Latent class analysis (LCA) and clustering of k-modes have been applied to assess the individuals' socioeconomic profiles. Models of clustering and random forest (RF) algorithms have been used to classify the socio-economic status of individuals. In the ten-class LCA model, what class of probability belongs to the newly selected individual has been determined. In the latent classes determined, the latent class profile descriptions

of the individuals were derived on the basis of the values provided by the variable values at the highest probability. Ten clusters that were obtained as a consequence of k-modes clustering were identified by cluster modes, and individual cluster profile descriptions were obtained, and their results were also compared to the LCA results. In this study, which considered categorical variables, it was observed that the LCA method yielded more consistent results than the method k-modes. The importance of the variables was calculated when the individual income was selected in the RF model as a component of all nine input variables. Variables of education, occupation and age have been observed to be more significant and have contributed the most to the RF model, respectively. Methods like LCA and RF appear to be suitable for applying DM and extracting useful results from SILC data, which are comprehensive and detailed data. Similar DM processes for different official data can be used to achieve useful results.

The opinions and comments expressed in the present study belong to the author of the thesis and do not constitute the opinions of the TurkStat. The same rule applies to all studies on the basis of this thesis.

Keywords: Data mining, SILC, latent class analysis, k-modes, random forests

1.1 Literatür Özeti

Bir ülkede toplam gelirin fert ve hanehalkları arasındaki dağılımını incelemek önem taşımakla birlikte, gelir grupları arasındaki farkların tespit edilmesi sosyal yapının iyileştirilmesine yönelik katkı sağlamaktadır [1]. Gelir dağılımı araştırmaları, istatistiksel bir çalışma olmakla birlikte özellikle makro ekonominin en önemli konularından birisi durumundadır. Ülkemizde özellikle 1960'lı yıllardan itibaren birçok kişi veya kuruluş tarafından gelir dağılımı araştırmaları yapılmıştır. Gelir dağılımına ilişkin kapsamlı araştırmalar gerçekleştiren kuruluşların başında Türkiye İstatistik Kurumu (TÜİK) gelmektedir. Özellikle 2000 yılından itibaren TÜİK her yıl ülkemizdeki gelir dağılımı araştırmalarını yapmakta ve yayınlamaktadır.

Kamu alanında, bilimsel çalışmalarda ve iş hayatında sürekli büyüyen veri yığınları oluşmaktadır. 1990'lı yıllar ile birlikte veri tabanları hızlı gelişim göstermiş ve bilgilerin keşfi ile analizi karışık bir işlem haline gelmiştir. Büyük veri yığınlarının oldukça küçük miktarı kullanılagelmekte ve kontrol edilebilirlik yönünden çok büyük olan veriler analiz edilebilirlik yönünden ise düzensiz ve anlamsız bir duruma bürünmüştür. [2].

Veri madenciliği (VM) ile ilgili birçok alanda yapılan çalışmalar, TÜİK tarafından derlenen gelir ve yaşam koşulları verisinin VM ile incelenmesi hususunda ilham kaynağı olmuştur. Türkiye'de resmi istatistikler konusunda VM ile ilgili yapılan başlıca çalışmalar şöyle özetlenebilir:

[3], 1976 yılından beri Atatürk Üniversitesinden mezun olmuş ve halen okumaya devam eden öğrencilerin mezun oldukları lisenin türü ve mezuniyet dereceleri ile kazanmış oldukları fakülteler arasındaki ilişkiyi VM yöntemleri kullanarak incelemişlerdir. [4], 2005-2008 yılları arasında İstanbul Menkul Kıymetler

Borsasında (İMKB) işlem gören 697 KOBİ'yi ele alarak VM tekniklerinden CHAID karar ağacı algoritması ile bu KOBİ'lerin güçlü ve zayıf yönlerini tespit etmiştir. [5], uçaklardaki parçaların herhangi bir arıza öncesi düzeltme ve önlemeye yönelik işlemlerinin yapılmasına yönelik ikaz seviyelerinin belirlenmesi için kural geliştirmek amacıyla Türkiye'de bir havayolu işletmesine ait parça söküm raporları üzerinde sınıflandırma algoritmaları kullanarak VM çalışması yapmıştır. [6] çalışmalarında, gıda, içki ve giyecek satışı yapan bir perakende işletmenin 6 aylık bir dönemde alışveriş yapmış müşterileri VM tekniklerinden Kohonen ağırları tekniği ile kümelemişlerdir. [7], bayanlar arasında sık rastlanan bir kanser türü olan meme kanserinin tanı ve teşhisi amacıyla Irvine California Üniversitesi VM havuzundan alınan 569 hastaya ait meme altı doku örnekleri VM yöntemlerinden C4.5 karar ağacı yöntemi ile analiz etmiştir. [8], Fırat Üniversitesi Teknik Eğitim Fakültesi Bilgisayar Öğretmenliği Bölümünde 2010-2011 öğretim yılında okuyan üçüncü ve dördüncü sınıf lisans öğrencilerinin sosyal ağırları kullanımları ve sosyal ağırlara ilişkin düşüncelerini tespit etmeye yönelik, bu bölümde okuyan 145 öğrenciye uyguladıkları anketten elde edilen verilere VM algoritmalarından birliktelik kuralı uygulayarak kurallar elde etmişlerdir. [9], yaptıkları çalışmada, farklı GSM şebekelerinin çekim gücü verilerinden bir veri bankası oluşturarak bu verileri VM yöntemlerinden iki-aşamalı kümeleme tekniği kullanarak analiz edip şebeke performanslarını karşılaştırmışlardır. [10], bir hastanenin veri tabanı içerisindeki tablolardan 96 aylık veriyi kullanarak 1 Ocak 2001 ve sonrasında hastaneye yapılan toplam aylık hasta başvuru sayılarını ele alıp gelecek aylarda oluşacak hasta başvuru yoğunluklarını üstel düzgünleştirme, Bileşik Otoregresif Hareketli Ortalamalar (ARIMA) ve yapay sinir ağırları modelleri ile tahmin etmişlerdir. [11], öğrencilere öğrenci başarısını etkileyebileceği düşünülen, öğrencilerin demografik özelliklerini kişisel, sosyal, ekonomik ve barınma detayında içeren anket uygulamış ve öğrencilerin akademik başarılarına etkisi olan faktörleri VM yöntemlerinden yararlanarak araştırmışlardır. [12], Cumhuriyet Üniversitesi Sosyal Bilimler Enstitüsü öğrencilerine ait verileri kullanarak verileri en iyi sınıflandıran algoritmanın bulunmasının yanında, öğrencinin notları, öğrencinin yüksek lisans / doktora öğrencisi olması, Sivas

ilinden olup olmaması, araştırma görevlisi olup olmaması, cinsiyeti, ders döneminin farklı olması gibi faktörlerin öğrenci başarısıyla ilintili olup olmadığını tespit etmeye çalışmıştır. [13], Bilecik Şeyh Edebali Üniversitesi Bilgisayar Mühendisliği Bölümü 2., 3. ve 4. sınıf öğrencilerine 8 farklı teknik seçmeli ders seçimlerinin nedenlerine ilişkin anket uygulamış olup elde edilen bilgiler arasında ne çeşit ilişkilerin yer aldığı, gösterdiği dağılım, ders seçimi ile seçilme nedeni arasında ne tür kuralların olduğunu VM ile tespit etmeye çalışmışlardır. [14], 2005-2008 yılları arasında İMKB’de işlem gören 140 sanayi işletmesini ele alarak finansal başarısızlıkları VM ve diskriminant analizi yöntemleri ile tahmin edip yöntemlerin sonuçlarını karşılaştırmışlardır. [15], mezun olamayacak öğrencilerin uyarılabileceği veya belirli bir değerin altında ortalamaya sahip olan öğrencilerin daha yoğun çalışmalarını önerebilmek için öğrencilerin mezuniyet notlarına ilişkin erken tahmin yapabilecek bir VM uygulaması gerçekleştirmek amacıyla, 2007 yılında okula başlayan ve 2011 yılında Fırat Üniversitesi, Eğitim Fakültesi, Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümünden mezun olan öğrencilerin mezuniyet notlarını VM yöntemlerinden yapay sinir ağı ve karar ağacı kullanarak tahmin etmişlerdir. [16], Ankara’da yer alan Yasemin Karakaya BİLSEM’de 2012-2013 eğitim öğretim yılının 2. döneminde öğrenim görmekte olan, 12 ve üstü yaştaki 113 üstün yetenekli öğrencinin verilerini kullanarak öğrencilerin ilgi alanlarını tahmin etmeye yönelik VM yöntemlerinden sınıflandırma analizini kullanmış ve bir arada ilgi duyulan alanları tespit etmek içinse birliktelik analizi kullanmışlardır. [17] yaptıkları çalışmada, İstanbul ilinde bulunan 42 farklı türdeki ortaöğretim kurumunu 2011 LYS sonuçlarına göre okul ve program türleri doğrultusunda başarı performanslarına göre eğitsel VM yöntemlerinden kümeleme analizi kullanarak kümelemiş ve karar ağacı modelleri ile de MF, TM ve TS test türü/türlerinin hangilerinin kümelere ayrışmada etkili olduklarını belirlemişlerdir. [18], karıştırma tipi, çimento dozajı, taze betondan numune alma zamanı ve numune boyutu değişkenlerine göre beton basınç dayanımının önceden belirlenmesine yönelik alternatif bir yöntem olarak VM’ni kullanarak modeller geliştirmişlerdir. [19] çalışmasında, resmi istatistiklerde veri madenciliği yöntemlerinin kullanımını geniş şekilde ele almıştır. Ayrıca 2012 yılı TÜİK

hanehalkı bütçe anketi mikro veri setindeki hane ve tüketim verilerini kümeleme ve birliktelik analizi yöntemleri iki aşamalı olarak analiz etmiştir. [20], 2012 yılı TÜİK Gelir ve Yaşam Koşulları Araştırması (GYKA) mikro kesit veri setini kullanarak hane ve fert verilerine ilişkin değişkenleri kohonen ve k-ortalamalar (k-means) yöntemleri ile kümelemiş ve yöntemlerin sonuçlarını karşılaştırmıştır. [21], 2010-2014 yılları arasındaki TÜİK Yaşam Memnuniyeti Araştırması verileri üzerinde etkin performansla sahip iki farklı yeni nesil metasezgisel algoritma kullanarak kümeleme analizi gerçekleştirmiştir. [22], 2013 yılı TÜİK GYKA verilerini kullanarak ferdin bir ay boyunca geçinebilmek için sahip olması gereken en düşük eşdeğer fert başına aylık net gelir değişkenini CHAID analizi ile bağımsız değişkenler ekseninde incelemiş ve bu geliri etkileyen etmenleri belirlemiştir. [23], 2014-2016 yılları arası TÜİK Hanehalkı İşgücü Araştırması (HİA) verilerine sınıflandırma yöntemlerinden C5.0, CHAID, lojistik regresyon ve Bayes ağlarını uygulamış ve model değerlendirme kriterlerine göre veri setine en uygun yöntem C5.0 karar ağacı yöntemi olmuştur. [24], 2016 yılı TÜİK GYKA verisinden 13 adet hane değişkeni seçerek çoklu uyum analizi (multiple correspondence analysis) uygulayarak borçluluk, harcama, yaşam koşulları vb. değişkenler ışığında bireyleri sınıflandırmayı amaçlamıştır. [25], 2016 yılı TÜİK Hanehalkı Bilişim Teknolojileri Kullanım Araştırması verilerine CHAID, C5.0, C&RT ve QUEST karar ağacı algoritmaları uygulayarak hanehalkının internet hizmeti sahipliğini hanehalkı karakteristiklerinin ne şekilde etkilediğini belirlemiştir. Buna göre en iyi başarıyı C5.0 algoritması göstermiştir. [26], 2014-2017 yılları arası TÜİK HİA mikro verilerini kullanarak hanehalklarının işgücü durumunu belirleyen etmenleri lojistik regresyon analizi ile araştırmıştır. Ayrıca karar ağacı algoritmaları ile iş arama süresine göre işsizler sınıflandırılmıştır.

[27] çalışmalarında, resmi istatistikleri ulusal istatistik enstitüleri (ÜİE) yanı sıra devlet daireleri ve yerel yönetimler tarafından toplanan idari ve tescil kayıtları ile sayım ve istatistiki araştırmalarda toplanan veriler olarak tanımlamaktadırlar ([28]). VM'nin resmi veriye uygulanması ise, verilerin farklı anketlerden veya idari kaynaklardan elde edilmesi ve gözlenen olgunun ölçümleri olarak uygun şekilde yorumlanması olarak tanımlanabilir ([29]). İstatistikçilerin adlandırdığı şekilde

“VM” veya bilgisayar bilimcilerin adlandırdığı şekilde “bilgi keşfi” son yirmi yılda hızla gelişmiş olup resmi istatistikler alanında giderek daha önemli hale gelmektedir ([30]). VM’nin; astronomiden kimyaya kadar birçok alanda kullanılması ve tanıtılmasına rağmen ([31], [32]), resmi istatistiklerin analizinde yeni desen ve modelleri belirlemek için VM’nden tam olarak istifade edildiğine dair çok az kanıt vardır ([33]). [34], yeni binyılın ilk beş yılında, resmi istatistiklerde VM’ne ilişkin rapor edilmiş çok az uygulamanın bulunduğunu belirtmişlerdir. Ayrıca, 2010 yılı itibariyle bile bu konuda çok az değişiklik olmakla birlikte, bildirilen tek çalışma resmi verilere VM tekniklerinin küçük bir uygulaması olan [27] tarafından yapılan çalışma olmuştur. [35], VM’nin resmi istatistiklerde daha zengin ve derin kavrayışlar elde etmek için tamamlayıcı olarak kullanılmasının elverişli olduğunu vurgulamıştır. ÜİE’nin veri toplama ile görevlendirilmiş olması ve yaygın uygulama olarak analizin dışarıya devredilmesiyle birlikte ([34], [27]); resmi istatistikçilerin beklenmedik desen veya modelleri bulmanın aksine kesin sorulara cevap bulmak ve tahminler yapmak şeklinde amaçlarının olmasından dolayı resmi istatistikler için VM uygulamalarının çok az sayıda olması şaşırtıcı olmamaktadır ([33], [27]). Yıllar geçtikçe, Ulusal İstatistik Enstitüleri (NSI), araştırmacılar ve akademisyenler resmi verilere VM teknikleri uygulamak için çok çaba sarfetmişlerdir. Resmi veriler için yayınlanmış en eski VM uygulaması 1996 yılında olmuştur ([28]).

Yabancı literatürde resmi istatistiklere VM tekniklerinin uygulandığı çalışmalar aşağıdaki gibi özetlenebilir:

[36], örnekleme anketlerinden topladığı verilere idari verileri ekleyerek bir nüfus sayımı yürütmek amacıyla 1990 Norveç nüfus sayım verilerine yapay sinir ağı impütasyonu uygulamıştır. ABD Kamu sağlığı Merkezi (The U.S. Public health Service), 1996 Tıbbi Harcama Panel Anketi (The Medical Expenditure Panel Survey-MEPS) çalışmasında anket sonrası cevapsızlık ağırlıklarını ayarlama sınıflandırma ağacı modelleri kullanmıştır ([37]). [38], yüksek ölüm oranlarını açıklamak için, mekânsal alt grup madenciliği (spatial subgroup mining) ve CBS haritalama (GIS mapping) arasındaki ilişkinin gösterilmesinde birliktelik analizi kullanarak 1991 İngiltere nüfus sayımı verilerini analiz etmişlerdir. [39],

ikidilliliğin geçen yüzyılın başındaki etkilerini ortaya çıkarmak için 1901 Kanada nüfus sayımı verilerine karar ağaçları tekniği uygulamışlardır. Resmi istatistikler için makine öğrenmesi tekniğinin yegâne uygulaması olan çalışmalarında [40], Portekiz İstatistik Enstitüsü'nün dış ticaret anketlerindeki hataları saptamak için, endüktif öğrenme algoritmaları ile sapan değer tespit metotlarını bir araya getirmişlerdir. [41], İngiltere'nin Stockport kentinde yapılan nüfus sayımının verilerinden mekânsal birliktelik kuralları elde etmek için çoklu-ilişkisel VM'ye dayanan bir yöntem önerip yeni bilgi olarak yorumlanabilecek bazı yeni kurallar keşfetmişlerdir. [42], coğrafi referanslanmış değişkenler arasındaki karmaşık ilişkileri analiz etmenin yanında, tahminlerin yapısal belirsizliğini kestirmeye olanak sağlaması için Bayes regresyon kullanmış olup İngiltere'nin Stockport kentinde yapılan nüfus sayımının verilerinden hareketle kentteki uzun süreli hastalık değerini tahmin etmek için Bayes model uygulamışlardır. [43], İspanya'nın nüfus birimleri (censal section) için hanehalkı başına ekonomik endeksleri hesaplamak için yapay sinir ağlarına dayalı bir prosedürle sembolik veri analizi uygulamışlardır. [44], bir kentsel alanın erişilebilirliği ile ilgili konuları ele almak amacıyla coğrafi referanslanmış 1991 İngiltere nüfus sayımı verilerine mekânsal birliktelik kuralları madenciliği uygulamışlardır. Böylece yapılan bu uygulama, değişik desenlerin tanımlanmasını sağlayarak kent plancıları için yeni bir bilgi haline gelmiştir. [45], nüfus verilerini mekânsal dağılımlarıyla ilişkilendiren veri analiz teknikleri için artan bir talep olduğunu belirtmişlerdir. [46], Romanya Ulusal İstatistik Enstitüsü sitesinden 1990 ve 2004 yılları arasındaki mezun öğrencilerin verilerini temin edip VM tekniklerinin bir uygulamasını yaparak, bilgi yönetim sistemlerinde VM tekniklerini kullanmanın üniversitelerin verimliliklerini arttırabileceğini belirtmişlerdir. [47] çalışmalarında, Amerika'daki Ulusal Tarım İstatistikleri Servisi (national agricultural statistics Service-NASS) tarafından güncel resmi istatistiklerin derlenmesinde VM'nin incelenmekte ve geliştirilmekte olduğunu göstermişlerdir. [48], NASS tarafından kullanılan cevapsızlık ağırlıklandırma metodolojisinin, 2007 tarım sayımı verilerine sınıflandırma ağaçlarının kullanılıp uygulanmasıyla artık değişmiş olduğunu belirtilmiştir. [49], NASS tarafından her birinde farklı

formatlarda anket uygulanan Amerika'daki 50 ayrı şehri hiyerarşik kümeleme analizi ile sırasıyla önce 20 ve sonra 5 bölgeye ayırmışlardır. Böylece anketin bölgeselleştirilmesi ile anket formatının sayısı azaltılmakla birlikte, cevaplayıcı yükünün azalabileceği ve anketin uygulama sürecinin daha kolay olabileceğini ifade etmişlerdir. [50], tarımsal uygulamaları temsil etmeyecek kayıtları tanımlayabilmek için nüfus sayımı başlangıç posta listesindeki kayıtlara sınıflandırma ağacı modelleri uygulamıştır, böylece gereksiz veri toplama maliyetlerinin düşürüleceğini ve tüm nüfus sayımı sürecinin etkinliğinin artacağını ifade etmiştir. [51], NASS tarafından 2007 tarım sayımı verilerine eksik veri impütasyonu amacıyla kümeleme analizi uygulandığını belirtmişlerdir. [52], tarımsal anket verilerindeki raporlama hatalarını analiz etmek için sınıflandırma ağacı modelleri kullanmış olup, verinin yapısı ile ilgili varsayımlarla sınırlanmadığı için sınıflandırma ağacı modellerini lojistik regresyon gibi yöntemlere göre daha faydalı bulmuşlardır. [53], Burkina Faso ülkesindeki 1960-2000 yılları arasındaki kadın nüfusunun yeniden inşa sürecindeki popülasyon dinamiklerini Bayes hiyerarşik model kullanarak modellemişlerdir. [54] çalışmasında genetik programlamanın resmi istatistiklerde nasıl başarıyla uygulanabileceğinden bahsetmiştir. [55], ABD Sayım Bürosu (The U.S. Census Bureau) tarafından 2010 Nüfus Sayımı Kapsama Ölçümü çalışmasında ırk sınıflamasının yinelemeli ayırması için CART (classification and regression tree) karar ağacı modeli yaklaşımında aşamalı regresyon uyguladığını belirtmiştir. [56], buğday ve pirinç hastalıkları için sınıflama kuralları keşfetmek amacıyla Hindistan tarımsal verilerine C4.5 karar ağaçları algoritmasını uygulamışlardır. [57], iş kazalarını etkileyen faktörleri analiz etmek amacıyla Finlandiya Kaza Sigorta Kurumları Federasyonu (federation of accident insurance institutions-FAII) veri tabanından 2006-2007 yılları arasındaki iş kazaları verisine karar ağaçları ve birliktelik kuralları yöntemlerini uygulamıştır. [58] yaptıkları çalışmada, çeliklerin yorulma mukavemetini tahmin etmek için Japonya Ulusal Malzeme Bilimi Enstitüsü (NIMS) MatNavi veri tabanındaki çeliklerin yorgunluk özelliklerini içeren verilere yapay sinir ağaçları, karar ağaçları ve çok değişkenli polinomiyal regresyon gibi ileri veri analiz yöntemleri uygulanmış olup bu yöntemlerin önceki

alıřmaların zerinde bir tahmin doęruluęu elde ederek nemli bir geliřme saęladıęı grlmřtr.

Yatay kesit ve zaman serisi verilerine nazaran panel veri ekonomik alıřmalarda birok avantaja sahip olmaktadır. Panel veri daha karmařık davranıřsal modellerin kurulumunu ve testini kolaylařtırmaktadır. Panel veri kullanımı, rneklem alıřmalarında sıklıkla karřılařılan ekonomik problemleri byk oranda azaltmaktadır. Bu nedenle tketicilerle ilgili alıřmalarda daha ok panel veri tercih edilmektedir. ([59]). Panel veri kullanımının bir dięer avantajı da dinamik model alıřmalarında daha uygun olmasıdır. Bu yzden iřsizlik, fakirlik gibi ekonomik durumları ieren alıřmalarda panel verinin kullanımı belli bir zaman aralıęında bireysel davranıřların irdelenmesine olanak saęlamaktadır ([60]). Panel veriye VM yntemlerinin uygulandıęı yabancı literatrdeki bařlıca alıřmalar ařaęıdaki gibi zetlenebilir:

[61] alıřmasında; [62] tarafından geliřtirilen regresyon aęaları yntemini ok deęiřkenli regresyon aęaları (multivariate regression trees-MRT) yntemi olarak geniřleterek srekli tipte olan panel veriyi modellemek amacıyla kullanmıřtır. [63], panel veri modellerine uyarlamak amacıyla ok deęiřkenli uyumlu regresyon uzanımları (multivariate adaptive regression splines / MARS) yntemini kullanmıřtır. Yine [64], oklu ikili baęımlı deęiřkenleri (multiple binary response variables) analiz etmek amacıyla sınıflandırma aęacı metodolojisini kullanarak parametrik olmayan bir yntem geliřtirmiřtir. [65], panel baęımlı deęiřkenlerin oluřturduęu veriyi kmelemede olabilirlik oran istatistięine (likelihood ratio statistic) dayalı bir ikili blmleme (binary partitioning) algoritması kullanmıřtır. [56] Avusturalya'daki 12 eřit avcı rmcek ile 6 eřit iklimsel karakteristik arasındaki iliřkiyi modellemek amacıyla veri setine ok deęiřkenli regresyon aęaları (MRT) yntemini uygulamıřtır. [67], CART karar aęacı yaklařımını genelleřtiren yinelemeli blnmeye (recursive partitioning) dayalı bir yntem kullanarak zamana baęlı aıklayıcı deęiřkenler (time-dependent variables) durumunda panel veriyi modellemeyi amalamıřlardır. [68], Missouri Ozark ormanlarındaki birlikte yetiřen eřitli bitki trlerinin bolluęunu tahmin etmek iin [61] tarafından panel veri iin kullanılan yaklařımın bir varyasyonu olarak MRT

yöntemini kullanmışlardır. [69], doğum, okula başlamak, meslek edinmek, evlenmek dizisi şeklinde ifade edilebilecek yaşam seyri (life course) verilerine sınıflandırma ağaçları yöntemini uygulamışlardır. [70], [71] ve [72], genelleştirilmiş tahmin denklemlerini (generalized estimating equations-GEE) kullanarak çoklu bağımlı değişkenleri analiz etmek için genelleştirilmiş çok değişkenli karar ağacı (generalized multivariate decision tree) olarak isimlendirdikleri bir yöntem geliştirmişlerdir. [73], MRT için bir yansız değişken seçim metodu geliştirmişlerdir. [74], bireysel panel veri şeklinde ifade edilebilecek kişisel yaşam seyri verilerine VM yöntemlerinin uygulanmasını ele almışlardır. [75] yaptıkları çalışmada, panel ya da hiyerarşik kümelenmiş verilere uygulanabilecek RE-EM ağacı isminde geliştirdikleri yöntemi ve uygulamalarını göstermişlerdir. [76], GUIDE (generalized, unbiased interaction detection and estimation) algoritmasını kullanarak panel veya çoklu bağımlı değişkenlerin oluşturduğu verilere regresyon ağaçları yöntemini uygulamışlardır. [77] panel veri için, ağaç tabanlı model ile karma etkiler (mixed-effects) modelinin faydalarını bir araya getirecek şekilde yeni bir ağaç algoritması önermişlerdir. [78], panel veri için regresyon ağaçları oluşturmak amacıyla LongCART algoritmasını önermişlerdir. [79], karesel çıkarsama fonksiyonunu (quadratic inference function-QIF) kullanarak CART yaklaşımını panel veriye uyarlamıştır. [80] çalışmasında, sınıflama ve regresyon ağaçları yöntemlerinin panel veriye uygulamalarını da ele alarak 50 yıllık literatür geçmişindeki gelişmeleri özetlemiştir.

Her ne kadar statik anlamda gelir ve yaşam koşullarıyla ilgili çalışmalar bulunsa da kişilerin yıllar itibariyle takip edilebileceği bir veri setinin olmaması sebebiyle, Türkiye’de gelir ve yaşam koşullarının dinamik olarak ele alındığı niceliksel çok az sayıda çalışma bulunmaktadır. Örneğin, [81], Türkiye’de yoksulluk dinamiklerine ilişkin 2006 ve 2007 yılları bazında GYKA panel verisini kullanarak araştırma yapmıştır. TÜİK GYKA anketi için kesit veri setlerinin kullanılmış olduğu statik anlamdaki çalışmalara ise [82], [83], [84] tarafından yapılan çalışmalar örnek olarak verilebilir.

Gelir, sosyoekonomik durum ve yaşam standartları için önemli bir ölçüm değişkenidir ve karar alma süreçlerinin oluşturulmasında da önemlidir. Bu çalışmada, gelirin bireyin sosyoekonomik veya refah durumu üzerindeki etkisini görmek için hanehalkı geliri yerine bireysel gelir ele alınmaktadır. Literatürde son yıllarda yapılan çalışmalara bakıldığında, [85] 2005 yılı EU-SILC verileri kullanarak İrlanda'daki göçmenlerin kazançlarını tanımlayıcı istatistikler ve regresyon analizi ile analiz etmişlerdir. [86], 2007 yılına ait EU-SILC Çek ulusal modülünü kullanarak Çek çiftçilerin gelir dağılımını ve yaşam koşullarını harcanabilir gelire göre kişi başı ortalama gelir hacmine göre analiz etmiştir. SILC verilerinde değişkenlerin çok azı sürekli, çoğu kategorik değişkendir. Kategorik değişkenli büyük veri dosyalarının analizi için [87], kategorik verilerin kümelenmesinde kullanılan yöntemleri incelemiştir. [88], 2011 yılı için Çek EU-SILC verilerini kullanarak, hiyerarşik küme analizi ve çeşitli k-kümeleme yöntemleri ile maddi yoksunluğu gösteren nominal göstergeleri analiz etmiştir. Yine aynı veri setiyle [89] çalışmasında yakın zamanda geliştirilen benzerlik ölçülerinin yanı sıra alternatif kategorik veri kümeleme yöntemleri olan iki-aşamalı küme analizi ve GSA kullanmışlardır. [90], 2005, 2010 ve 2014 yıllarına ait Çek EU-SILC verilerinde VM tekniklerini kullanarak hanehalkı harcanabilir gelirini etkileyen sosyoekonomik faktörleri araştırmışlardır. Buna ek olarak, [91] fırsat eşitsizliğini tahmin etmek amacıyla 2011 yılı EU-SILC verisi için ağaç tabanlı tahmin yöntemleri, RO ve GSA yaklaşımlarını kullanmışlardır.

1.2 Tezin Amacı

ÜİE'ler tarafından üretilen resmi veriler devletin ekonomik ve sosyal karar alma sürecinde önemli bir rol oynamaktadır. Resmi verilerin geleneksel istatistiksel yaklaşımlardan ziyade VM yöntemleri ile ele alınması, yeni bilgiler ve gizli kalıpların ortaya çıkarılması bakımından önem kazanmaktadır.

Çok boyutlu GYKA verisini VM yöntemleriyle kapsamlı bir şekilde değerlendirmeyi amaçladığımız bu çalışmada, hanehalkı fertleri arasında yıllık toplam geliri etkileyen değişkenler açısından ne tür genellemeler ya da kuralların ortaya konulabileceği araştırılmak istenmiştir.

Bu çalışmanın amacı, Türkiye'deki bireylerin gelir durumlarını etkileyen sosyoekonomik faktörleri belirlemek ve gelir kategorilerindeki yerlerine ve sosyoekonomik profillerine göre fertleri sınıflandırmak ve refah durumlarını araştırmaktır. Bu çalışmanın odak noktası, bireysel gelirin ölçümü ve bu geliri etkileyen değişkenlerdir. Bu amaçla TÜİK'den 2015 yılı kesit mikro veri setleri temin edilerek VM analizleri gerçekleştirilmiştir. Çalışmada öncelikle fert geliri değişkeni kategorize edilerek fert gelirini etkileyen sosyoekonomik değişkenler belirlenmiştir. Daha sonra belirlenen değişkenlerin oluşturduğu veri setine en uygun VM yöntemleri belirlenmiştir. Veri setinin kategorik değişkenlerden oluşmasından dolayı veri setine en uygun VM yöntemleri olarak GSA ve k-modlar yöntemlerinin kümeleme amaçlı, rastgele orman (RO) yönteminin ise sınıflandırma amaçlı seçilmesi kararlaştırılmıştır. GSA ve k-modlar analizleri sonucunda oluşan sınıf ve küme tanımları karşılaştırılmıştır. Yine RO analizi sonucunda bir sınıflandırma modeli elde edilmiş ve modelde hangi değişkenlerin önemli rol oynadığı tespit edilmiştir.

1.3 Orjinal Katkı

Türkiye'de hanelerden ziyade bireylerin gelirlerini, sosyoekonomik ve refah durumlarını VM yöntemleriyle kapsamlı bir şekilde inceleyen ampirik çalışmalar yetersizdir. Bu çalışma ile bireylerin refah durumlarına ilişkin sonuçların VM yöntemleriyle karşılaştırmalı olarak sunulması katkıda bulunulması amaçlanmıştır.

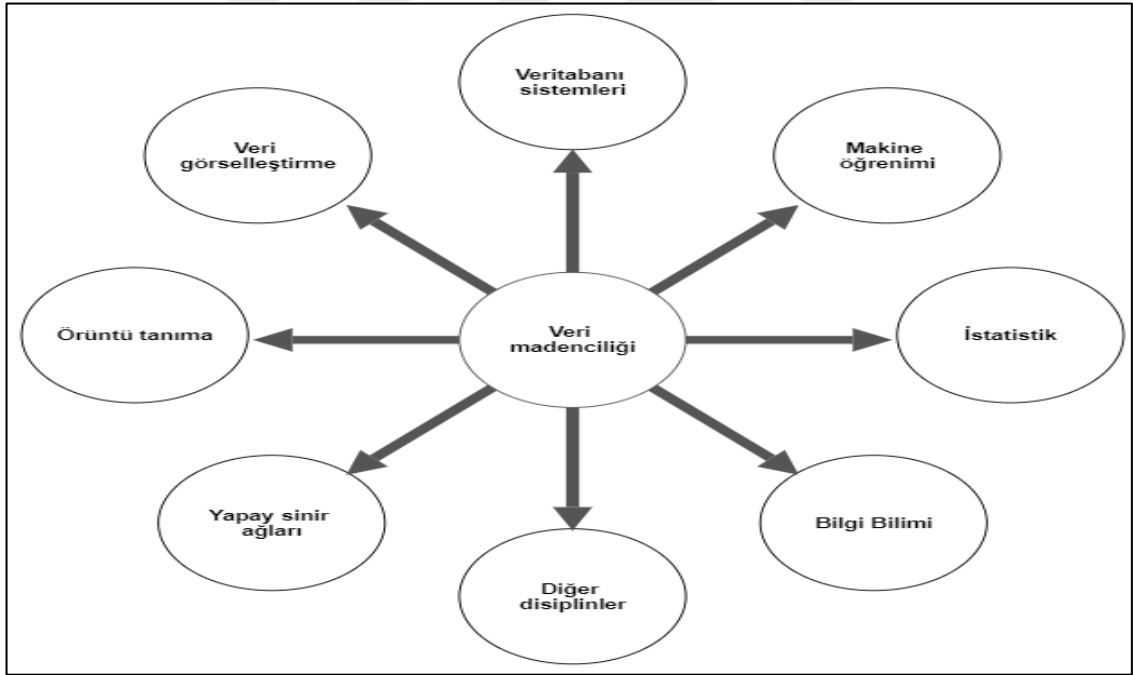
Gelir; sosyoekonomik durum ve yaşam standartları için önemli bir ölçüm değişkenidir ve karar verme süreçlerinin oluşturulmasında da önemlidir. Bu çalışmada, hanehalkı geliri yerine bireysel gelir ele alınarak gelirin ve geliri etkileyen değişkenlerin bireyin sosyoekonomik veya refah durumu üzerindeki etkisi görülmek istenmiştir. Literatürde gelirle ilgili yürütülmüş VM çalışmalarının çoğu hanehalkı harcanabilir geliriyle ilgilenmiş ve sonuçlar sınırlı yöntemlerle elde edilmiştir. Bu çalışmada gelişmiş VM yöntemleri kullanılarak GYKA verileri kapsamlı bir şekilde incelenmiş, bireysel gelir ile bu geliri etkileyen değişkenler

tespit edilerek sınıflar ve kümeler elde edilmiş ve sonuçlar karşılaştırmalı olarak yorumlanmıştır.



2.1 Veri Madenciliği Kavramı

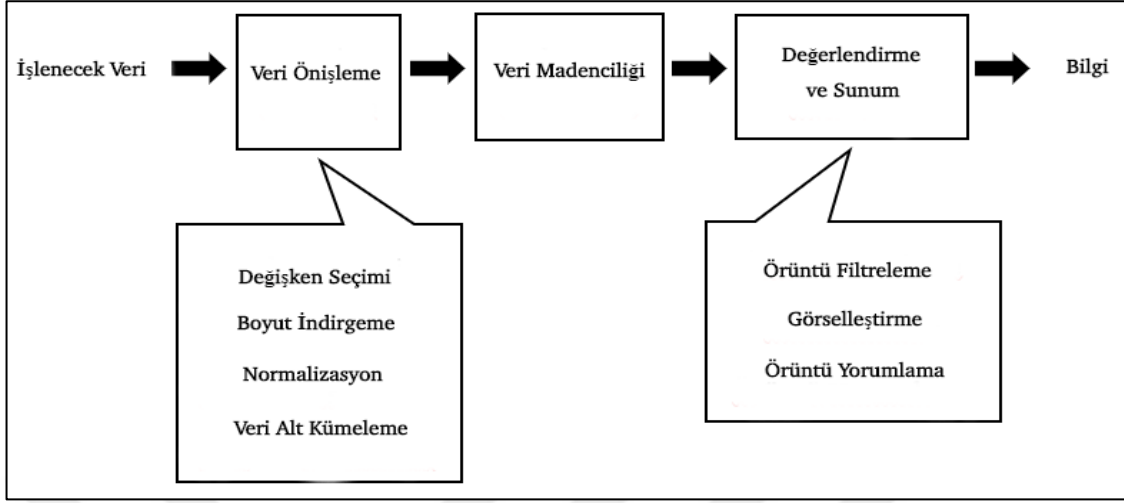
VM, bilgi teknolojilerinin gelişmesiyle beraber büyük hacimli verilerden önceden bilinmeyen ve gizli kalmış anlamlı ve faydalı olabilecek bilgi ve örüntülerin elde edilmesi sürecidir. Madencilik kelimesi burada verinin büyüklüğünü ifade etmektedir. Dolayısıyla VM bir veri yığımından değerli ve gizli kalmış bilgilerin elde edilmesidir. Veri madenciliğinde klasik istatistiksel yöntemlerin aksine daha çok algoritmaların yer aldığı teknikler kullanılmaktadır. Büyük veriden anlamlı bilgilerin elde edilmesi ihtiyacı giderek artmaktadır. Veri madenciliğinin ilişkili olduğu alan ve disiplinler Şekil 2.1’de sunulmuştur.



Şekil 2.1 Veri madenciliğinin diğer alan ve disiplinler ile ilişkisi

VM, veri tabanlarında bilgi keşfi (VTBK) sürecinin bir parçası olarak görülebilir. VM kavramı uygulamada genellikle VTBK olarak da kullanılmaktadır. Verinin

önişlemeden geçirilip son işleme kadar olan süreci (VTBK) aşağıdaki gibi özetlenebilir (Şekil 2.2):



Şekil 2.2 Veri tabanlarında bilgi keşfi süreci (VTBK) ([92])

2.2 Veri Madenciliği Tarihçesi

1950’lerde ilk bilgisayarların ortaya çıkması ile VM kavramı oluşmaya başlamıştır. 1960’lı yıllarda verilerin depolanması/saklanması kavramı ile birlikte veri ambarı sistemleri ortaya çıkmış 1970’li yıllarda ise ilişkisel veri tabanı yönetim sistemleri geliştirilmeye başlanmış ve 1980’lerde yaygınlaşmaya başlamıştır. 1990’lı yıllarda “VM” terimsel olarak kullanılmaya başlanmış ve veri hacimlerinin artmasıyla birlikte 1992 yılında ilk VM yazılımı hazırlanmıştır. 2000’li yıllarla birlikte VM yöntemleri hızla geliştirilmeye başlanmış ve birçok alanda kullanılmaya başlanmıştır.

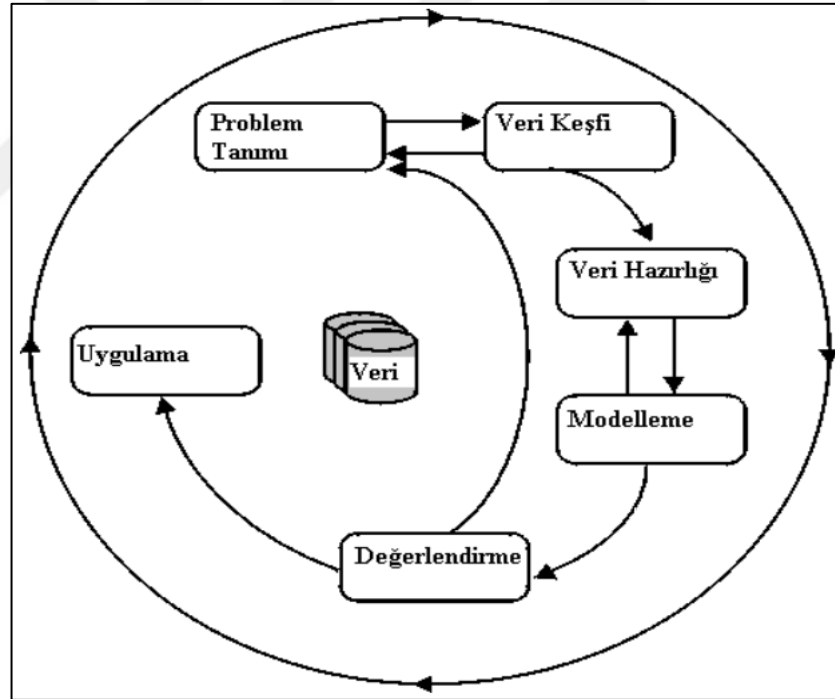
2.3 Veri Madenciliği Süreci

VM, büyük veri tabanlarından daha önceden bilinmeyen ve gizli kalmış anlamlı bilgi ve örüntüleri algoritmalar veya makine öğrenimi yöntemleri yoluyla ortaya çıkarma sürecidir. VM sürecinin adımları genel olarak aşağıdaki gibidir:

- Problemin anlaşılması ve tanımlanması (analizin amacı ve ne işe yarayacağı)
- Verilerin hazırlanması (veri setinin düzenlenmesi, temizlenmesi, dönüştürülmesi ve istenilen değişkenlerin elde edilmesi)

- Modelin kurulması (farklı modeller ve durumlar için elde edilecek sonuçlara göre problemin çözümü için en uygun modelin kurulması)
- Modelin değerlendirilmesi (model değerlendirme ölçütleri kullanarak modelin teknik açıdan değerlendirilmesi)
- Modelin kullanılması (kurulan modelin doğrudan bir uygulama şeklinde ya da başka bir uygulamanın alt parçası şeklinde kullanılması)
- Modelin izlenmesi (modelin anlaşılabilirliği ve kullanılabilirliği irdelenir ve zaman içerisinde meydana gelecek değişikliklerle birlikte model izlenerek gerekirse tekrar inşa edilir)

VM süreçlerinin karmaşık bir hale gelmesini önlemek ve standartlaştırmak amacıyla 1996 yılında Chrysler, SPSS ve NCR firmaları bir araya gelerek Çapraz Endüstri VM Standart Sürecini (CRISP-DM) oluşturmuşlardır. Bu süreç Şekil 2.3'te görselleştirilmiştir:



Şekil 2.3 CRISP-DM VM süreci ([93])

2.4 Veri Madenciliği Modelleri

VM modelleri tahmin edici ve tanımlayıcı olmak üzere iki gruba ayrılmaktadır. Tahmin edici modeller sonuçları bilinen verileri kullanarak sonuçları bilinmeyen verilere yönelik öngörüler yapılmasını sağlamaktadır. Tanımlayıcı modeller ise karar vermeye yardımcı olabilecek sonuçlar elde etmek amacıyla mevcut verilerden örüntüler elde edilmesini sağlamaktadır.

2.4.1 Tahmin Edici Modeller

Veri madenciliğinde tahmin edici modeller bağımlı değişken ile bağımsız değişkenler arasındaki ilişkileri belirlemek amacıyla kullanılmakta ve mevcut sonuçları esas alarak yeni değerlerin tahmin edilmesini sağlamaktadır. Tahmin edici modeller tahmin edilmek istenen niteliğin değerlerine göre farklılaşmaktadır. Tahmin edici modellerden sınıflandırma, bir birimi niteliklerine göre önceden tanımlanmış bir kategoriye atanmasını sağlamaktadır. Regresyon, süreklilik özelliği gösteren değerleri tahmin etmek amacıyla kullanılmaktadır. Zaman serisi analizi ise zamana göre sıralanmış ölçümlerin bulunmasını sağlamaktadır.

Sınıflandırma ve regresyon modellerinde yaygın olarak kullanılan yöntemler aşağıda yer almaktadır.

- Yapay Sinir Ağları (Artificial Neural Networks)
- Genetik Algoritmalar (Genetic Algorithms)
- k-En Yakın Komşuluk (k- Nearest Neighbour / kNN)
- Naïve-Bayes sınıflayıcısı
- Lojistik Regresyon (Logistic Regression)
- Karar Ağaçları (Decision Trees)
- Random Orman Sınıflandırıcısı (Random Forests)
- Destek Vektör Makineleri (Support Vector Machines)
- Bellek Temelli Nedenleme (Memory Based Reasoning)
- Diskriminant Analizi (Discriminant Analysis)
- Kaba Kümeler (Rough Sets)

2.4.2 Tanımlayıcı Modeller

Veri madenciliğinde tanımlayıcı modeller kümeleme ve birliktelik kuralları modelleridir.

Kümeleme, veri setinde yer alan benzer verileri ayırarak gruplandıran ve yaygın olarak birçok alanda kullanılan bir VM yöntemidir. Kümeleme yöntemleri aşağıdaki gibi kategorize edilebilir:

- Bölümlemeli yöntemler (partitioning based methods)
- Hiyerarşik yöntemler (hierarchical methods)
- Yoğunluk tabanlı yöntemler (density-based methods)
- Izgara tabanlı yöntemler (grid-based methods)
- Model tabanlı yöntemler (model-based methods)
- Merkeze dayalı yöntemler (centroid-based method)

Birliktelik kuralları, veriler arasındaki ilişkileri inceleyerek bir arada oluşabilecek durumları keşfetmeye çalışan VM yöntemidir. Özellikle pazarlama alanında sepet analizlerinde kullanılmaktadır. Birlikte satın alınan ürünleri belirleyerek müşterilerin alışveriş alışkanlıklarını tespiti ve böylece raf düzenlemesi ve ürünlere çabuk erişilebilmesini sağlamaktadır. Birliktelik analizi için yaygın kullanılan algoritmalara örnek olarak Apriori algoritması, FP Growth algoritması, SETM algoritması ve AIS algoritması gösterilebilir.

2.5 Model Değerlendirme Ölçütleri

VM analizleri sonrası elde edilen modellerin verdiği sonuçların doğruluğunun ölçülmesi ve modellerin performanslarının karşılaştırılarak en doğru ve güvenilir tahminler veren modelin seçilmesi önem teşkil etmektedir. Modellerin performanslarını değerlendirme işlemi model değerlendirme ölçütleri baz alınarak yapılmaktadır. Bir sınıflandırma modelinin doğruluğunu ölçmenin en yaygın yollarından biri Tablo 2.1'de gösterildiği gibi bir hata matrisi (karışıklık (confusion) matrisi) kullanmaktır. Hata matrisinde gerçek sonuçlar ile modelin tahminlemesi kıyaslanmaktadır.

Tablo 2.1 Hata matrisi örneği

		Tahmin Edilen Değerler		
		Pozitif	Negatif	Toplam
Doğru (Gerçek) Değerler	Pozitif	DP (Doğru Pozitif)	YN (Yanlış Negatif)	P
	Negatif	YP (Yanlış Pozitif)	DN (Doğru Negatif)	N
	Toplam	P'	Y'	P'+Y'=P+N

Tablo 2.1'e göre,

- DP: Doğru sınıflandırılan pozitif birimlerin sayısını
- YP: Gerçekte negatif iken pozitif olarak sınıflandırılan birim sayısını
- YN: Gerçekte pozitif iken negatif olarak sınıflandırılan birim sayısını
- DN: Gerçekte pozitif iken pozitif olarak sınıflandırılan birim sayısını

ifade etmektedir. YN ve YP yanlış sınıflandırılanları (misclassification) göstermektedir.

2.5.1 Doğruluk Oranı

Doğruluk (accuracy), ölçümün gerçekleşen değere olan yakınlığı olmakla birlikte doğru şekilde sınıflandırılan birimlerin yüzdesini göstermektedir. Doğruluk değeri bir modelin performansını ölçmek için kullanılan yaygın bir ölçüttür, fakat tek başına kullanımı yeterli görülmemektedir. Doğruluk oranını 1'e tamamlayan değer ise hata oranı (error rate) olmaktadır.

$$\text{Doğruluk} = \frac{DP+YP}{DP+YP+YN+DN} \quad (2.1)$$

$$\text{Hata Oranı} = 1 - \text{Doğruluk} = \frac{YN+YP}{DP+YP+YN+DN} \quad (2.2)$$

2.5.2 Duyarlılık

Duyarlılık (sensitivity), sınıflandırma modeli tarafından doğru tahmin edilen pozitif örneklerin oranını ifade etmektedir.

$$Duyarlılık = \frac{DP}{DP+YN} \quad (2.3)$$

2.5.3 Kesinlik

Kesinlik (precision) ölçütü, gerçek pozitif değerlere ilişkin, sınıflandırma modeli tarafından tahmin edilmiş pozitif değerlerin yüzdesini göstermektedir.

$$Kesinlik = \frac{DP}{DP+YP} \quad (2.4)$$

2.5.4 F-ölçütü

F-ölçütü, kesinlik ve duyarlılık ölçütlerinin harmonik ortalaması şeklinde tanımlanmaktadır. Her iki ölçütü birlikte ele alarak daha güvenilir sonuçlara ulaşılmasını sağlar. F-ölçütü değerleri 0 ile 1 arasında değişmekle beraber, bu aralıktaki büyük değerler daha iyi modellere karşılık gelmektedir.

$$F - \text{ölçütü} = 2 \times \frac{Duyarlılık \times Kesinlik}{Duyarlılık + Kesinlik} \quad (2.5)$$

2.5.5 Kappa Katsayısı

Kappa değeri sınıflandırma modelinin sınıflandırma başarısı hakkında bilgi sağlamaktadır. Kappa değeri -1 ile +1 arasında bir değer almaktadır. Kappa değeri aşağıdaki formül yardımıyla hesaplanmaktadır.

$$\kappa = \frac{P_0 - P_e}{1 - P_e} \quad (2.6)$$

Burada P_0 gözlenen sınıfların oranını, P_e ise beklenen sınıfların oranını göstermektedir.

[94] tarafından Kappa değer aralıkları aşağıdaki gibi yorumlanmaktadır:

- <0 : Hiç uyum yok
- $0 - 0,20$: Önemsiz uyum
- $0,21 - 0,40$: Orta derecede uyum
- $0,41 - 0,60$: Makul uyum
- $0,61 - 0,80$: Önemli uyum

- 0,81 - 1 : Neredeyse mükemmel uyum

2.5.6 Akaike Bilgi Kriteri (AIC)

Akaike bilgi kriteri (AIC) tahmin edilen modeller arasından en uygun modeli seçmek amacıyla [95] tarafından ortaya konulmuş bir uyum iyiliği ölçüsüdür. Modeller arasında en düşük AIC değerini sağlayan model en iyi uyum sağlayan modele yaklaşılmaktadır. AIC aşağıdaki formül ile hesaplanmaktadır.

$$AIC = 2k - 2\ln(L) \quad (2.7)$$

k : Modeldeki serbest parametre sayısı

L : Maksimum olabilirlik (yüksek değer modelin daha iyi olduğunu gösterir)

2.5.7 Bayes Bilgi Kriteri (BIC)

Bayes bilgi kriteri (BIC) [96] tarafından geliştirilmiştir. BIC olabilirlik fonksiyonuna dayalıdır ve AIC ile ilişkilidir. BIC kriteri AIC ile kıyaslandığında, AIC kriterine göre daha az sayıda parametreye dayalı modelleri seçmektedir. BIC formülü örneklem büyüklüğüne bağlı olması dolayısıyla AIC kriterinden farklılık göstermektedir. Modeller karşılaştırılırken en küçük BIC kriteri değerine sahip model tercih edilmektedir.

$$BIC = \ln(n)k - 2\ln(\widehat{L}) \quad (2.8)$$

k : Model tarafından tahmin edilen parametre sayısı

n : Örneklem büyüklüğü

$\widehat{L}$: Model tarafından elde edilen log-benzerlik değeri

2.5.8 Entropi

Genellikle bir sistemdeki belirsizliğin ölçüsü olarak tanımlanan entropi kavramı bilgi teorisinde bir istatistiksel deneye ilişkin belirsizlik ölçüsü olarak [97] tarafından aşağıdaki formül ile tanımlanmaktadır:

$$Entropi (S) = -\sum_{i=1}^n p_i \log p_i \quad (2.9)$$

Burada X kesikli rastgele değişken ve p_i olasılık değerleri olmak üzere, $\forall i = 1, 2, \dots, n$ değerleri için $\sum_{i=1}^n p_i = 1$ ve $p_i \geq 0$ olmaktadır. X değişkeninin sürekli

değerleri için (2.9) formülünde toplama yerine integral alma işlemi yapmak gerekmektedir. Entropi değeri 0 ile 1 arasında değişmektedir.



3.1 Çalışmada Kullanılan Veri Seti ve Özellikleri

Bu çalışmada, son yayınlanan GYKA 2015 yılı kesit mikro verisi TÜİK'den temin edilmiş ve bu verinin gelişmiş VM yöntemleriyle kapsamlı bir şekilde incelenmesi amaçlanmıştır. Bu amaçla fert, fert kayıt ve hane türündeki kesit veri setleri ele alınmış olup fert geliri ve onu etkileyen değişkenler VM yöntemleriyle araştırılarak fertlerin refah durumları irdelenmiştir.

TÜİK, ülke ekonomisinin durumunu tespit etmek ve fertlerin yaşam koşullarını tespit edebilmek amacıyla GYKA anketini uygulamaktadır. GYKA ile Türkiye'de hanehalkı ve fertlere ilişkin gelir dağılımını elde etmek ve yaşam şartları, sosyal dışlanmayı ve yoksulluğu gelir ile ölçerek fertlerin profillerinin belirlenmesi amaçlanmaktadır. Bu amaçla TÜİK 2006 yılından bu yana her yıl GYKA'yı uygulamaktadır. GYKA'da fertler 4 yıl boyunca takip edilmekte ve araştırma sonuçlarından her yıl kesit ve panel veri setleri elde edilmektedir ([98]). GYKA'ya ilişkin genel açıklama Ek A1'de yer almaktadır.

3.2 Veri İşleme

Büyük veri ile ilgili çalışmaların en çok zaman alan kısmı veri düzenleme aşaması olup bu aşama tamamlandıktan sonra analizler daha kısa sürede tamamlanabilmektedir. Çalışmada en güncel veriyi kullanabilmek amacıyla, GYKA 2015 yılı kesit türündeki mikro veri setleri TÜİK'den temin edilmiştir.

Veriyi tanıma aşamasında mikro veri setleri detaylı incelenmiştir. 59662 ferde ilişkin bilgi içeren 2015 yılı fert kesit mikro verisinden 15+ yaştaki fertler için geliri etkileyen değişkenlerden (fertlerin temel özellikleri, gelirleri ve gelir ile ilgili değişkenler) tüm fertler için geçerli olanlar (eğitim durumu, sağlık durumu vb.) ile birlikte bazı değişkenler analizler sırasında araştırmanın amacına ve merak konusuna göre dönüştürülüp birleştirilmiştir. Türkiye'de genellikle fonksiyonel,

kişisel, sektörel ve bölgesel olarak incelenen gelir dağılımı için bu çalışmada öncelikle, gelire etki eden değişkenler ve gelir bileşenleri ele alınmıştır.

Uygulamada, TÜİK tarafından mikro veri yayınlanmadan önce gerçekleştirildiği için, GYKA'na ilişkin örneklem seçimi ve büyüklüğü, anket sorularına ilişkin geçerlilik ve verilerin güvenilirliği ayrıca sınıanmamıştır. Veriyi düzenleme aşamasında, “csv” formatında temin edilen 2015 yılına ait kesit mikro veri setleri Microsoft Excel programı aracılığıyla veri türüne ayrı bir dosya olarak “csv” formatında düzenlenmiştir. Uygulama için 2015 yılına ait fert, fert kayıt ve hane türündeki kesit mikro veri setleri ele alınmıştır.

Çalışmada veri düzenleme aşaması en çok zaman harcanan aşama olmuştur. Fert, fert kayıt ve hane kesit veri setlerinden değişkenler elde edilip kodlanmıştır. GYKA için her fert Türkiye genelinde belli sayıda kişiyi temsil edecek şekilde TÜİK tarafından örnekleme yapıldığı için, fert veri setindeki ferdin yıl içinde elde ettiği gelirlerin toplamı olan FG140 değişkeni, fert için ağırlık katsayısı olan FB030 değişkeni ile ağırlıklandırılmıştır. Böylece gelir skalası kullanmak yerine, ilgili sene verilerini doğru şekilde yansıtmak ve doğru kıyaslama yapabilmek amacıyla fert geliri için ağırlıklandırılmış %20'lik gelir grubu aralıkları elde edilmiştir. Anlamlı gelir değerleri elde edebilmek için FG140 değişkenindeki sıfır ve eksi değerlerdeki gelirler hariç tutulup sadece pozitif gelir değerleri ele alınarak toplamda 36036 ferde ilişkin veri seti kullanılmıştır. FB030 ve FG140 değişkenleri kullanılarak ağırlıklandırılmış kantil değerleri (ağırlıklı kantiller / weighted quantiles) hesaplanmıştır. Ağırlıklı kantilleri kullanmamızın nedeni Türkiye genelini tahminleyecek şekilde gelir değişkeni için uygun kategorik aralıklar elde etmektir. Her gelir grubuna eşit sayıda veri düşmesi amaçlandığı için ağırlıklı yüzde 20'lik dilimler kullanılmıştır. Ağırlıklı yüzde 20'lik dilimler “Laeken” R paketi ile hesaplanmış olup aynı hesaplar Excel ortamında da yapıp çapraz kontrol sağlanmıştır. Ağırlıklı kantillere ilişkin [99] tarafından ortaya konulan yöntem kullanılmıştır. Ağırlıklı kantil yönteminde öncelikle FG140 değişkeni olan gelir değişkenindeki değerler küçükten büyüğe sıralanmaktadır. Daha sonra fert için ağırlık katsayısı olan FB030 değişkeni verideki sütun boyunca kümülatif olarak bir önceki satır değeriyle toplanmaktadır. En büyük FG140 fert gelirine karşılık gelen

FB030 kümülatif ağırlık katsayısı değeri beşe bölünerek karşılık geldiği FB030 kümülatif ağırlık katsayısı değerine göre ilgili FG140 değeri belirlenerek ağırlıklı kantil aralıkları belirlenmektedir. Fert gelirine ilişkin 2015 yılı itibariyle hesaplanan ağırlıklı kantil değerleri aşağıdaki gibidir (Tablo 3.1):

Tablo 3.1 Fert geliri için 2015 yılına ilişkin ağırlıklı kantiller

Ağırlıklı Kantil	Gelir Değeri	
	En Düşük	En Yüksek
1. kantil	0	6615
2. kantil	6616	12000
3. kantil	12001	16200
4. kantil	16201	26400
5. kantil	26401	+

Güncel 2015 yılı kesit GYKA verisi için düzenleme yapılırken temin edilen tüm mikro verilerde FB030 ve diğer FG ile başlayan fert gelir değişkenlerinin “108.417.357.491.725.00” gibi noktalı olduğu görülmüştü ve daha önce diğer yıllar verisi için yapıldığı gibi FG140 toplam fert geliri değişkeni için excel programında

“=EĞER(YADA(FG140>10000000000;FG140<10000000000);FG140/100000000000;BQ2)” şeklinde basamak düzeltme formülü kullanılmaktaydı. Fakat yeni temin edilen veride FG140 gelir değişkeninde bazı verilerin beklenenden büyük olduğu, ilk mikro verideki basamak sayısı aynı olduğu halde FG uzantılı gelir değişkenlerinin FG140 değişkenini verecek şekilde satır toplamalarına bakıldığında orijinal mikro veride basamaklardaki nokta sayılarının aynı olduğu halde FG uzantılı değişkenlerin farklı değerlere sahip olması gerektiği tespit edilmiştir. Dolayısıyla yukarıda kullanılan formül istenileni vermemiş ve yapılacak düzeltme tüm veri setinin düzenlenmesi ve analiz işlemlerinin tekrar ele alınarak baştan

yapılmasını gerektirmiştir. Bu noktada sorunu giderecek yöntemler araştırılmış olup sorunun Excel programında aşağıdaki ayarlamaların (Tablo 3.2) yapılarak giderilebildiği görülmüştür.

Tablo 3.2 Excel programında basamak düzeltme için yapılan ayarlar

Microsoft Excel Sürümü	Excel ondalık ayırıcı	Excel binlik ayırıcı	Excel sistem ayırıcıları ¹	Denetim Masası Bölge/Dil Ayarları		
				Ondalık simgesi	Ondalık basamak sayısı	Basamak gruplandırma simgesi
Excel 2007	Nokta olarak belirlenmelidir	Nokta silinerek yerine Space tuşu ile boşluk bırakılmalıdır ¹	Sistem ayırıcılarını kullan onay kutusu işaretli ise kaldırılmalıdır ¹	Virgül	9	Space tuşu ile boşluk bırakılmalıdır

Hedef değişken olan fert gelir değişkenine ilişkin değerlerin doğru elde edilerek doğru gelir aralıklarına ulaşılması yapılacak analizin sonuçlarını direkt etkileyecektir. Veri düzenlemesi aşamasında faydalanan RStudio programında veri setlerinde nokta (".") olarak bulunan hücrelere "NA" yazılması sağlanarak basamaklı değerlerle aynı işleme tabi tutulması engellenmek istenmiştir. Bu amaçla veri seti RStudio programına "readxl" R paketi kullanılarak eklenmeye çalışılmış, fakat veri setinde bazı sütunlarda rakamsal değer bulunduğu halde "TRUE" şeklinde hücre değerlerinin yazılarak programa eklendiği tespit edilmiştir. Dolayısıyla problemi gidermek için yapılan araştırmalar neticesinde RStudio programında "xlsx" uzantılı veri eklerken "readxl" paketinde `guess_max = 21474836` değeri belirlenip `> veri <- read_excel("veri.xlsx", sheet = 1, guess_max = 21474836, col_types = "numeric")` kodu yardımıyla excel verileri eklenmiş ve veri setinde tutarlılık sağlanmıştır. Bu noktadan itibaren başlangıç veri setlerinde tutarlılık sağlanıp ağırlıklı kantilleri oluşturarak gelir grupları tekrar elde edilmiş ve veri düzenlemenin bu uzun ve can alıcı aşaması tamamlanmıştır.

Veri düzenleme aşamasında RStudio programının yanında ağırlıklı olarak Microsoft Excel programı kullanılmıştır. Çalışmada ele alınan tüm değişkenler (eksik değerler / missing values dahil) kategorik olarak ilk değer 1'den başlayacak şekilde kodlanmıştır. Eksik değerler için bu değerleri ifade edecek yeni değişken değeri atamaları yapılmıştır.

3.3 Kullanılan Yöntemler

Çalışmada VM analizlerinde RStudio (versiyon 1.1.456 ve R versiyonu 3.5.1) yazılımı tercih edilerek kullanılmıştır.

Verinin VM yöntemleriyle analizi amacıyla öncelikle kümeleme yöntemi belirlenmiş, daha sonra ise kümeleme için değişken seçimi yapılmıştır. Ele alınan veri setindeki tüm değerler kategorik olduğundan, kategorik veri setlerine ilişkin kümeleme yöntemleri araştırılmış ve veri setinin niteliklerine en uygun olması bakımından GSA ve k-modlar kümeleme yöntemleri verilerden küme sınıfları elde edilmesi amacıyla belirlenmiştir. Daha sonra sınıflama modeli elde etmek amacıyla veri seti üzerinde az varsayım gerektiren RO yöntemi kullanılmıştır. GSA için R paketlerinden poLCA 1.4.1 versiyonu, k-modlar analizi için klaR 0.6-14 versiyonu ve RO analizi için randomForest 4.6-14 versiyonu kullanılmıştır. GYKA veri setindeki değişkenlerin çoğu kategorik olduğundan, kategorik verilere uygun VM yöntemleri araştırılmış ve EU-SILC verileri ve benzeri araştırmalar için [89] tarafından önerildiği gibi bu çalışmada da GSA yöntemi veri setine uygun görülerek tercih edilmiştir.

3.4 Değişken Seçimi

Analizler için başlangıçta fert, fert kayıt ve hane veri setlerinden fert gelirini etkilediği düşünülen 20 adet değişken (bunlardan bazıları mevcut değişkenlerden kodlanmıştır) belirlenip kodlanmıştır. Belirlenen bu başlangıç değişkenleri Tablo 3.3'te yer almaktadır.

Tablo 3.3 Fert gelirine yönelik seçilen başlangıç değişkenleri

Veri Türü	Değişken	Tanım
Fert Kayıt	AGE_11	Yaş grubu
Fert Kayıt	AGE_3	Yaş grubu
Fert Kayıt	AGE_6	Yaş grubu
Fert	EDU	Ferdin en son bitirdiği okul (Eğitim Durumu)
Fert	EMP	Ferdin Anketin Yapıldığı Andaki İstihdam (Çalışma) Durumu
Fert Kayıt	GEN	Cinsiyet
Hane	HIT	Hanenin Gelir Tipi (Faaliyet dışı gelirler: emeklilik, kira-menkul, sosyal transfer gelirleri, diğer gelirler)
Fert	HTH	Ferdin genel sağlık durumu (Sağlık Durumu)
Fert	IG	Ferdin % 20'lik gelir grubu
Hane	INT	Hanehalkının İNTERNET'e sahip olma durumu
Fert	MST	Medeni Durum
Fert	OCC	Ferdin esas işindeki meslek kodu (Meslek)
Hane	RGN	Geographical region where the household is located
Fert	SEC_10	Çalışılan Sektör (10 kategorili)
Fert	SEC_4	Çalışılan Sektör (4 kategorili)
Fert	SEC_7	Çalışılan Sektör (7 kategorili)
Fert	SSI	Ferdin esas işinden dolayı sosyal güvenlik kuruluşuna kayıtlılık durumu (SGKya kayıtlılık)
Fert	STA	Ferdin esas işindeki işteki durumu (İstihdam şekli)

Tablo 3.3 Fert gelirine yönelik seçilen başlangıç değişkenleri (devamı)

Hane	TEL	Hanehalkının TELEFON HATTI'na (sabit) sahip olma durumu
Hane	WSY	Oturulan konutta BORULU SU SİSTEMİ'nin mevcut olma durumu

Başlangıç değişkenlerinden “çalışılan sektör (SEC)” değişkeni 4, 7 ve 10 kategori olmak üzere 3 farklı değişken olarak, yine “yaş (AGE)” değişkeni 3, 6 ve 11 kategori içerecek şekilde 3 farklı değişken olarak kodlanmıştır. Belirlenen başlangıç değişken kümesi şu değişkenlerden oluşmaktadır: Medeni durum (MST), Ferdin en son bitirdiği okul (EDU), Ferdin genel sağlık durumu (HTH), Ferdin anketin yapıldığı andaki istihdam (çalışma) durumu (EMP), Ferdin esas işindeki işteki durumu (STA), Ferdin esas işindeki meslek kodu (Meslek) (OCC), Ferdin esas işinden dolayı sosyal güvenlik kuruluşuna kayıtlılık durumu (SGK'ya kayıtlılık) (SSI), Cinsiyet (GEN), Hanenin bulunduğu coğrafi bölge (RGN), Oturulan konutta borulu su sisteminin mevcut olma durumu (WSY), Hanehalkının telefon hattına (sabit) sahip olma durumu (TEL), Hanehalkının internete sahip olma durumu (INT), Hanenin gelir tipi (faaliyet dışı gelirler: emeklilik, kira-menkul, sosyal transfer gelirleri arasında) (HIT), Ferdin %20'lik gelir grubu (IG) ve SEC4, SEC7, SEC10, AGE3, AGE6, AGE11 değişkenleri. Veri setinde birden fazla çeşit kategoride yer alan yaş ve sektör değişkenlerini azaltmak için Boruta R paketinden faydalanılarak değişkenlerin önem düzeyleri belirlenmiş ve değişken seçimi (IG gelir değişkeni hariç 19 adet değişken arasında) yapılmıştır. Boruta ile değişken seçim işlemine ait uygulama ve çıktılar Ek B-5'te yer almaktadır. Bu sonuçlara göre, yaş değişkenlerinden AGE3 ve AGE11, sektör değişkeninden ise SEC7 ve SEC4 değişkenleri veri setinden çıkarılmıştır. Çalışmada ilk sırada uygulanacak GSA analizi öncesi karmaşıklığı azaltmak için uzman görüşüne de başvurularak 15 değişkenin hepsinin fert gelirini etkileyip etkilemediği doğrultusunda analizler için gerekli 14 değişken belirlenmiştir (SSI, HTH, TEL

değişkenleri veri setinden çıkarılıp EHSIZE¹ değişkeni eklenerek). GSA analizi öncesi karmaşıklığı azaltmak için LCAvarsel R paketi kullanılarak kalan bu 14 değişken arasında değişken seçim işlemi yapılmıştır. Değişken seçim prosedürü için LCAvarsel paketinde genetik algoritma, swap ve stepwise algoritmaları kullanılmaktadır. LCAvarsel ile değişken seçim işlemine ait uygulama ve çıktılar Ek B-6'da sunulmuştur. LCAvarsel paketinde genetik algoritma kullanıldığında STA, OCC, RGN, AGE_6 ve GEN değişkenleri önemli bulunmuştur. Swap ve stepwise algoritmaları kullanıldığında ise WSY, INT, HIT, EHSIZE_4 değişkenleri her iki algoritmada da önemli bulunmuştur. Bu nedenle, LCAvarsel algoritması kullanılarak yapılan değişken seçim işlemi farklı denemelere karşın, çok az değişkeni önemli bularak istenilen sonucu vermemiştir. Birbiriyle ilişkili olduğu düşünülen ve poLCA sonuçlarına göre sınıflara göre çok değişim göstermeyen değişkenler çıkarılmıştır. HIT, WSY, EHSIZE_4 değişkenlerinin poLCA ile yapılan GSA denemelerinde sınıfları fazla etkilemediği görüldüğünden bu değişkenler yerine AGE, GEN, MST değişkenleri eklenmiştir. Nihayetinde uzman görüşleri de baz alınarak sınıf sayısını azaltmak için en uygun değişken kümesi kararlaştırılmıştır: STA, OCC, RGN, INT, SEC, EDU, IG, AGE, GEN, MST. Karşılaştırılabilirlik açısından k-modlar ve RO analizleri için de bu değişken kümesi baz alınmıştır. Çalışmada analizler için kullanılan nihai veri setine ilişkin değişkenler Tablo 3.4'te yer almaktadır.

¹ EHSIZE değişkeni, OECD modifiye edilmiş eşdeğerlik ölçeğine göre hesaplanmıştır ve bu ölçek:

- hane reisi için 1.0 ağırlık;
- 14 yaşından büyük herhangi bir hane üyesi için 0,5 ağırlık;
- 15 yaşından küçük her çocuk için 0.3 ağırlık

vermektedir, Eşdeğer hane halkı büyüklüğü bu ağırlıklar kullanılarak elde edilen sonuçların toplanmasıyla bulunmuştur. Bu değişken EHSIZE_4 isminde 1-2, 3-4, 5-6, 7+ aralıkları için kodlanmıştır.

Tablo 3.4 GSA, k-modlar ve RO uygulamalarında kullanılan nihai veri setine ilişkin değişkenler

Değişken	Tanım	Kategori
INT	Hanehalkının internete sahip olma durumu	1: Evet
		2: Hayır
RGN	Hanenin bulunduğu coğrafi bölge	1: Marmara
		2: Ege
		3: İç Anadolu
		4: Akdeniz
		5: Karadeniz
		6: Doğu Anadolu
		7: Güney Doğu Anadolu
SEC	Çalışılan sektör	1: Tarım
		2: Sanayi
		3: İnşaat
		4: Ticaret
		5: Ulaştırma
		6: Hizmet
		7: Mali
		8: Kamu/Yönetim
		9: Sağlık
		10: Sektörü Yok

Tablo 3.4 GSA, k-modlar ve RO uygulamalarında kullanılan nihai veri setine ilişkin değişkenler (devamı)

EDU	Ferdin en son bitirdiği okul (eğitim durumu)	1: Okur yazar olmayan / bir okul bitirmeyen
		2: İlkokul
		3: Ortaokul
		4: Genel lise
		5: Meslek lisesi
		6: Yüksek Öğrenim
MST	Medeni durum	1: Evli
		2: Diğer
OCC	Ferdin esas işindeki meslek kodu (meslek)	1: Yöneticiler, Profesyonel Meslekler
		2: Teknisyen-Tekniker, Yardımcı Profesyonel, Büro Hizmetleri
		3: Hizmet ve Satış Elemanları
		4: Nitelikli Tarım, Hayvancılık, Su Ürünlerinde Çalışanlar, Tesis ve Makine Operatörleri ve Montajcılar
		5: Sanatkârlar ve ilgili işlerde çalışanlar
		6: Nitelik Gerektirmeyen İşlerde
		Çalışanlar
		7: Mesleği Yok

Tablo 3.4 GSA, k-modlar ve RO uygulamalarında kullanılan nihai veri setine ilişkin değişkenler (devamı)

AGE	Yaş grubu	1: 15-24
		2: 25-34
		3: 35-44
		4: 45-54
		5: 55-64
		6: 65+
STA	Ferdin esas işindeki işteki durumu	1: Ücretli
		2: Yevmiyeli
		3: İşveren
		4: Kendi hesabına
		5: Ücretsiz aile işçisi
		6: İstihdamda (faal) değil
GEN	Cinsiyet	1: Erkek
		2: Kadın
IG	Ferdin %20'lik gelir grubu	1: 1. ağırlıklı %20'lik dilim
		2: 2. ağırlıklı %20'lik dilim
		3: 3. ağırlıklı %20'lik dilim
		4: 4. ağırlıklı %20'lik dilim
		5: 5. ağırlıklı %20'lik dilim

3.5 Modelleme

Çalışmada analizler Intel ® Core™ i7-8700K CPU @ 3.70GHz işlemcili 16 GB RAM kapasiteli bilgisayar üzerinde gerçekleştirilmiş ve işlemcinin tüm çekirdekleri kullanılarak paralel işlem ile hesaplamalar hızlandırılmaya çalışılmıştır.

STA, OCC, RGN, INT, SEC, EDU, IG, AGE, GEN, MST değişkenleri için başlangıçta sınıf sayısı 30 olarak belirlenmesine ve iterasyon sayısı artırılmasına rağmen yerel minimum (local minima) değeri (poLCA sonuçlarında en düşük Bayes Bilgi Kriteri (BIC) değeri) bulunamamıştır. Dolayısıyla eş-ortak değişken (covariate / concomitant variable) kullanılmadan gerçekleştirilen GSA analizinde stabil ve yorumlanabilir bir sonuç elde etmek için sınıf sayısı 10 baz alınarak sonuçlar değerlendirilmiştir. Yine aynı değişkenler kullanılarak aynı veri setinde k-modlar algoritması ile kümeleme yapılarak sonuçlar GSA sonuçlarıyla karşılaştırılmak istenmiştir. k-modlar algoritması, ele alınan veri seti kategorik olduğundan k-ortalamlar algoritması yerine tercih edilmiştir. “klaR” R paketinin kullanıldığı k-modlar analizinde iterasyon değeri 3000 alınarak gerçekleştirilmiştir.

Belirlenmiş olan değişkenler üzerinden RO sınıflama modeli kullanılarak, gelecekteki hanehalkı fertleri için kurallar oluşturularak bir sınıflama modeli elde edilmesi amaçlanmıştır. Veriler “randomForest” ve “caret” R paketleri kullanılarak RStudio ortamında analiz edilmiştir. 36036 gözlemi içeren verinin %70’i eğitim verisi %30’u ise test verisi olarak kullanılmıştır. RO analizinde sonuçlar öncelikle eğitim verisi için oluşturulmuş, daha sonra test verisi için de oluşturularak karşılaştırılmıştır. RO analizinde uygulanan aşamalar aşağıdaki gibidir:

- Öncelikle eğitim verisi için RO modeli uygulanmış ve hata matrisi (error matrix) elde edilmiştir ve yine aynı RO modeli kullanılarak tüm eğitim verisi için tahmin (prediction) modeli oluşturulmuştur.
- Daha sonra aynı RO modeli üzerinden test verisi için tahmin sonuçları elde edilmiş ve hata matrisi yoluyla doğruluk değerindeki iyileşme incelenerek daha önce eğitim verisi için elde edilmiş sonuçlarla karşılaştırılmıştır.

- Bu ilk RO modeli için hata oranı (error ratio) grafiği elde edilerek yapılan denemelerle torba-dışı (out-of-bag (OOB)) hata oranı kestirimi değerinin kaç ağaç değerinden sonra stabil olduğu belirlenmiştir.
- Ayrıca OOB hatasını m_{try} değerinin fonksiyonu olarak grafikleştirerek hatanın en düşük gerçekleştiği m_{try} değeri belirlenmiştir.
- Elde edilen en uygun ağaç değeri ve m_{try} parametrelerini kullanarak eğitim ile test verilerine RO modeli uygulayıp modelin doğruluk değerindeki değişim gözlenmek istenmiştir.
- Histogram grafiği ile RO modelindeki ağaçlara ilişkin düğüm sayıları görselleştirilmiş olup RO modeli için değişkenlerin önem düzeyleri ifade edilmiş ve değişkenlerin modelde bulunma sıklıkları gösterilmiştir.
- Kısmi bağımlılık grafikleri ile değişkenlerin sınıflara olan marjinal etkisi ölçülmüştür.
- RO modeline ilişkin tek bir ağacın çıktısı elde edilmiş, ayrıca modele ilişkin ilk 4 ağaç görselleştirilmiştir.

3.5.1 Gizli Sınıf Analizi

1950 yılında Lazarsfeld çok değişkenli veri setleri için model tabanlı bir kümeleme yöntemi olan GSA yöntemini geliştirmiştir. GSA resmi istatistiklere çeşitli uygulamaları olan ve son yıllarda popülerleşmeye başlayan, yüksek kalitede anlamlı istatistikler sunan olasılık temelli oldukça güçlü bir kümeleme tekniğidir. GSA model-bazlı olup önsel olasılıkları kullanarak cevaplayıcıları gizli sınıflara sınıflandırır ([100], [101]). GSA ile iki veya daha fazla gözlenen kategorik değişken kullanılarak kategorik gizli değişkenler elde edilebilmekte ve faktör analizinin kategorik karşılığı olarak görülmektedir. Gizli ve gözlenen değişkenlerin kategorik ya da sürekli özellikte olma durumlarına göre sınıflama yöntemleri Tablo 3.5'te yer almaktadır.

Tablo 3.5 Değişken tipine göre gizli değişken modelleri

	Gözlenen Değişkenler	
Gizli Değişken	Sürekli	Kategorik
Sürekli	Faktör Analizi	Gizli Özellik Analizi
Kategorik	Gizli Profil Analizi	Gizli Sınıf Analizi

GSA kategorik veriler için kümeleme analizi olarak da görülebilir. GSA analizinin amacı gözlenen değişkenlerden gizli sınıf adı verilen tek bir kriter elde etmektedir. GSA analizinin en temel varsayımı yerel bağımsızlık varsayıdır. Bu varsayıma göre gözlenen değişkenler arasındaki ilişki bir gizli sınıfın varlığı ile ifade edilebiliyorsa, bu gizli sınıfın sabit tutulması gözlenen değişkenlerin istatistiksel olarak bağımsız olmasını sağlamalıdır. Dolayısıyla GSA bu koşulu sağlayacak gizli sınıfları elde etmeyi amaçlamaktadır. GSA analizinde değişkenler kategorik olduğundan çok değişkenli normal dağılım varsayımı aranmamaktadır.

[101] tarafından önerilen standart GSA modeli aşağıdaki gibi ifade edilmektedir:

$A, B, \dots, E$ gözlenen değişkenler ve X bir gizli sınıf değişkeni olmak üzere; A 'nın I sayıda, B 'nin J sayıda, ..., E 'nin M sayıda kategorisi, X gizli sınıf değişkeninin T sayıda gizli sınıfı olsun. Buna göre GSA modeli:

$$\pi_{ij\dots mt}^{AB\dots EX} = \pi_t^X \pi_{it}^{AX} \pi_{jt}^{BX} \dots \pi_{mt}^{EX} \quad (3.1)$$

$$i = 1, \dots, I; j = 1, \dots, J; m = 1, \dots, M; t = 1, \dots, T \quad ([102])$$

[103] yukarıdaki GSA modelini loglineer model şeklinde aşağıdaki gibi tanımlamıştır:

$$\pi_{ij\dots mt}^{AB\dots EX} = \eta \tau_i^A \tau_j^B \dots \tau_m^E \tau_t^X \tau_{it}^{AX} \tau_{jt}^{BX} \dots \tau_{mt}^{EX} \quad (3.2)$$

$$\ln \pi_{ij\dots mt}^{AB\dots EX} = \lambda + \lambda_i^A + \lambda_j^B + \dots + \lambda_m^E \lambda_t^X \lambda_{it}^{AX} \lambda_{jt}^{BX} + \dots + \lambda_{mt}^{EX} \quad (3.3)$$

Burada π_t^X gizli sınıf olasılığı gözlemin t 'inci gizli sınıfta bulunma olasılığını ifade eder. Tüm bu olasılıkların toplamı T sayıda sınıf için 1'e eşit olmaktadır. π_{it}^{AX}

şeklindeki olasılıklar ise gözlenen değişkenin gizli sınıftaki ağırlığını ifade eden koşullu olasılıklardır ([102]).

3.5.2 k-modlar Kümeleme Algoritması

k-modlar kümeleme algoritması ilk defa [104] tarafından yayınlanan ve kategorik veriyi kümelemek için geliştirilmiş ilk algoritmalarından biridir. Öklid uzaklığı yerine basit eşleşme uzaklığı (simple matching dissimilarity measure) kullanan k-modlar algoritması standart k-ortalamlar algoritmasının kategorik verilere uyarlanmış bir türevidir. k-modlar algoritması küme merkezlerini göstermek için modları kullanmakta ve her iterasyonda modları güncellemektedir. Bir veri setini k adet kümeye bölmek için k-modlar algoritması aşağıdaki adımları uygular.

- Adım 1: Başlangıç küme modları (merkez) olarak rastgele k nesne seçilir
- Adım 2: Her bir nesne ile küme modu arasındaki uzaklık hesaplanır ve her nesne kendisine en yakın olan kümeye atanır. Tüm nesneler kümelere atanan kadar bu adım tekrarlanır.
- Adım 3: Her bir küme için yeni bir mod seçilir ve önceki ile karşılaştırılır. Eğer ikisi de farklı ise Adım 2'ye dönlür, aksi takdirde ise algoritma durdurulur.

Yukarıdaki kümeleme işlemi aşağıda yer alan k-modlar amaç fonksiyonunun en düşük değerini almasını sağlamaya çalışmaktadır.

$$F(U, Z) = \sum_{l=1}^k \sum_{i=1}^n \sum_{j=1}^m u_{i,l} d(x_{i,j}, z_{l,j}) \quad (3.4)$$

(3.4) formülünde, $U = [u_{i,j}]$ bir $n \times k$ boyutunda örüntü (partition) matrisi, $Z = \{Z_1, Z_2, \dots, Z_k\}$ bir mod vektörü kümesi ve $d(\cdot, \cdot)$ uzaklık fonksiyonudur. k-modlar algoritması k-ortalamlar algoritmasına benzemekle birlikte kategorik verileri kümelemede etkin bir kümeleme yöntemidir ve yerel minimum sonuçlar üretebilmektedir ([105]).

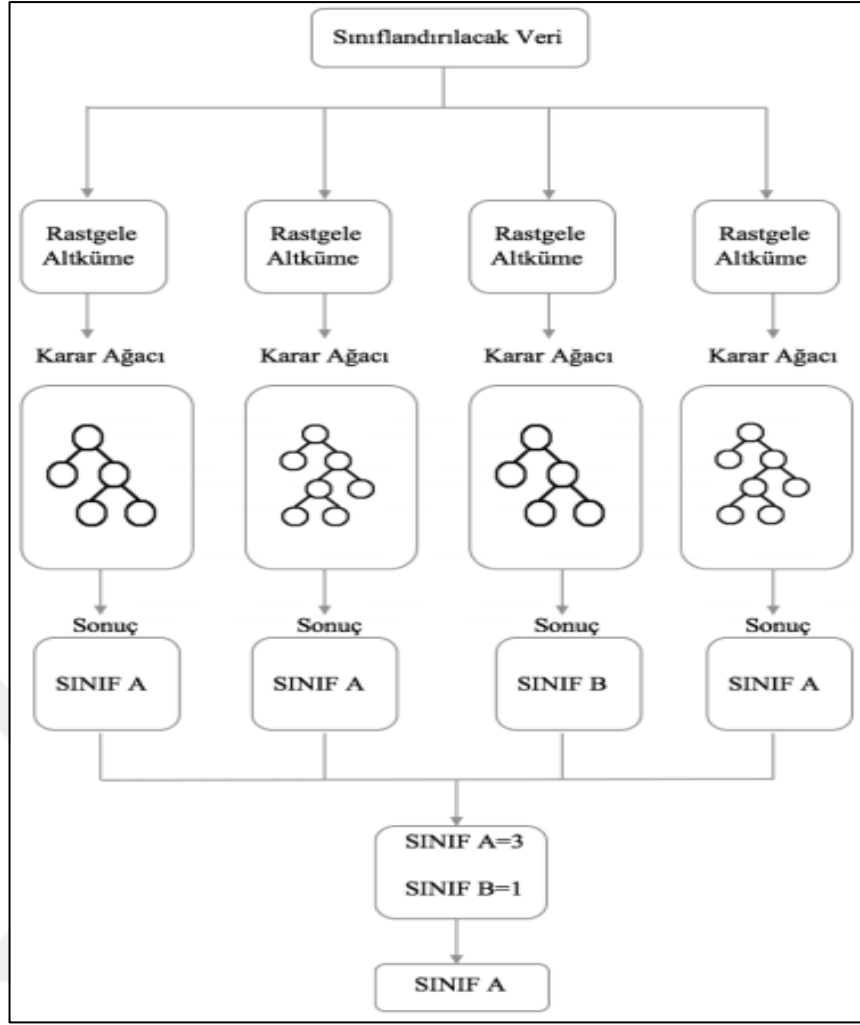
3.5.3 Rastgele Orman (Random Forest)

RO, sınıflandırma ve regresyon problemlerini çözmek için [106] tarafından geliştirilen karar ağaçları topluluğudur. RO yüksek boyutlu verilerde makine öğrenimi ve örüntü tanıma konusunda oldukça popüler, güçlü ve başarılı bir

topluluk öğrenme tekniğidir. Faktör tipi değişkenleri sınıflandırmak için RO kullanılabilir. RO, birden fazla karar ağacı oluşturur ve bunları daha doğru ve istikrarlı bir tahmin oluşturmak için birleştirir.

RO birçok karar ağacı oluşturur. RO modeline ilişkin çıktı tüm tekli ağaçların verdiği oylarla belirlenir. Her bir karar ağacı bir ağaç algoritması kullanarak girdi verisine ilişkin bootstrap örneklemeleri sınıflandırarak oluşturulur. Daha sonra her ağaç test verisini sınıflandırmak için kullanılır. Her ağaç, test verilerini etiketlemek için bir karar oluşturur ve bu karar oy olarak adlandırılır. Ağaçlar arasında en fazla olan oy toplandıktan sonra orman sınıflama sonucu için karar verir ([107]).

Düğümler ağaçlarda dallara ayrılırken mevcut değişkenlerin hepsi yerine rastgele seçilen bir alt kümesi içindeki en iyi sonucu verecek olan değişken kullanılarak işlem yapılır ([108]). RO sınıflandırıcısına ilişkin ağaç yapısı aşağıda yer almaktadır (Şekil 3.1):



Şekil 3.1 RO ağaç yapısı ve işleyişi ([109])

RO algoritması ise aşağıdaki adımlardan oluşmaktadır.

$b = 1$ den B 'ye olmak üzere;

- Eğitim verisinden N büyüklüğünde bootstrap örnekleme oluşturulur.
- Bootstrap örneklemden T_b RO ağacı oluşturulur, minimum düğüm sayısı olan n_{min} büyüklüğüne erişene kadar ağacın her bir uç düğümü için aşağıdaki adımlar yinelemeli olarak devam ettirilir,
 - I. p değişkenden rastgele olarak m değişken seçilir
 - II. m değişken içinden en iyi ayırma noktası belirlenir
 - III. Düğüm iki ayrı alt düğüme ayrılır

Daha sonra ağaçlar topluluğuna ilişkin $\{T_b\}_1^B$ çıktı oluşturulur. Yeni bir x noktası için tahmin oluşturulmak istenirse:

- Regresyon için: $\hat{f}_{rf}^B(x) = \frac{1}{B} \sum_{b=1}^B T_b(x)$,
- Sınıflama için: $\hat{C}_b(x)$ b 'inci RO ağacının sınıf tahmini olmak üzere, $\hat{C}_{rf}^B(x) = \text{oy çokluğu (majority vote)} \{\hat{C}_b(x)\}_1^B$ 'dir ([110]).



4.1 Gizli Sınıf Analizi ile Sonuçlar

“STA, OCC, RGN, INT, SEC, EDU, IG, AGE, GEN, MST” değişkenleri baz alınarak aşağıdaki R kodu (Tablo 4.1) yardımıyla GSA analizi (eş-ortak değişken (covariate/concomitant variable) kullanılmadan) gerçekleştirilmiş ve 10 ayrı gizli sınıf elde edilmiştir.

Tablo 4.1 Seçilmiş 10 değişken ile GSA’ya ilişkin komutlar

> set.seed(100000)
> library(poLCA)
> library(foreach)
> library(doParallel)
> registerDoParallel(cores=6)
> f_Data_sel <- cbind(STA, OCC, RGN, INT, SEC, EDU, IG, AGE_6, GEN, MST)~1
> sink("output.txt")
> print(lc_sel <- foreach(i=2:30, .packages="poLCA") %dopar% poLCA(f_Data_sel, Data, nclass=i, maxiter=3000, tol=1e-5, na.rm=FALSE, nrep=10, verbose=TRUE,
calc.se=TRUE))
> sink()

Komut çıktıları aşağıdaki gibidir (Şekil 4.1):

```

Estimated class population shares
0.1439 0.0992 0.1142 0.1427 0.0862 0.1517 0.0552 0.064 0.0553 0.0877

Predicted class memberships (by modal posterior prob.)
0.1445 0.1021 0.1136 0.1447 0.0867 0.1514 0.0529 0.0636 0.053 0.0874

=====
Fit for 10 latent classes:
=====
number of observations: 36036
number of estimated parameters: 439
residual degrees of freedom: 35597
maximum log-likelihood: -393680.6

AIC(10): 788239.1
BIC(10): 791967.2
G^2(10): 121462.7 (Likelihood ratio/deviance statistic)
X^2(10): 3770834 (Chi-square goodness of fit)

```

Şekil 4.1 Seçilmiş 10 değişken ile GSA çıktıları

Yukarıdaki 10 sınıflı GSA modeli çıktısında “Estimated class population shares” ve “Predicted class memberships” değerlerinin birbirine yakın çıkması modelin iyi uyum sağladığı yönünde bir gösterge olmaktadır. “Estimated class population shares” değerleri her gizli sınıfa ne kadar gözlem düştüğüne ilişkin tahmini oranlar olmaktadır. Örneğin 10 sınıflı modelde gözlemlerin payı sırasıyla, gizli sınıf 1’de %14.39, gizli sınıf 2’de %9.92, gizli sınıf 3’de %11.42 olarak tahmin edilmiştir. “Predicted class memberships” değerleri yeni seçilen bir ferdin hangi olasılıkla hangi sınıfa dahil olacağını tahmin etmektedir. Örneğin yeni seçilen bir ferdin Sınıf 6’ya ait olma olasılığı %15,14 şeklindedir. GSA modellerine ilişkin uyum istatistikleri Tablo 4.2.’de yer almaktadır.

Tablo 4.2 GSA’nın sınıflara göre model uyum istatistikleri

Gizli Sınıf Sayısı	Log-Olabilirlik	AIC	BIC	Olabilirlik Oranı (G2)	Ki-kare (χ^2)
2	-440063.3	880300.5	881039.3	214228.1	9389058
3	-423522.8	847307.7	848420.2	181147.3	6592438

Tablo 4.2 GSA'nın sınıflara göre model uyum istatistikleri (devamı)

4	-414272	828894.1	830380.2	162645.7	7256971
5	-409407.7	819253.4	821113.2	152917.1	7227581
6	-404562.9	809651.9	811885.4	143227.5	6572636
7	-401188.2	802990.4	805597.6	136478.1	6555825
8	-397973.2	796648.4	799629.2	130048	6925361
9	-395495.6	791781.2	795135.6	125092.8	4542961
10	-393680.6	788239.1	791967.2	121462.7	3770834

Tablo 4.2 BIC bilgi kriterinin en düşük değeri yönünden incelendiğinde en iyi modelin 10 gizli sınıflı (BIC = 791967.2) model olduğu görülmüştür. GSA çıktılarındaki değişkenlere ilişkin gizli sınıf olasılıkları tablo şeklinde Ek B-1'de yer almaktadır. Bu tabloyu, örneğin gizli sınıf "Sınıf 1" için yorumlamak gerekirse, "EDU (Ferdin en son bitirdiği okul)" değişkeni için "2 (Primary school)" yanıtı veren bir ferdin gizli sınıf 1'de yer alma olasılığının %51.92 olduğu görülmektedir. Böylece, gizli sınıf 1'deki fertlerin Primary school mezunu olma olasılığının (Pr(2)) daha yüksek olduğu ifade edilebilir. Yine bu tablodan elde edilen gizli sınıflarda değişkenlerin en yüksek olasılıkla aldıkları değişken değerleri gizli sınıfları karakterize etmesi bakımından aşağıdaki gibi özetlenmiştir (Tablo 4.3):

Tablo 4.3 GSA sonucu elde edilen gizli sınıflar

	Gözlenebilen Değişken									
Gizli Sınıf	STA	OCC	RGN	INT	MST	AGE	GEN	SEC	EDU	IG
1	6	7	1	2	1	5	1	10	2	3
2	1	3	1	2	1	3	1	4	2	4
3	1	4	1	1	1	3	1	2	2	4

Tablo 4.3 GSA sonucu elde edilen gizli sınıflar (devamı)

4	1	1	1	1	1	2	1	6	6	5
5	4	4	5	2	1	4	1	1	2	4
6	6	7	1	2	2	2	2	10	1	1
7	2	6	6	2	1	1	1	3	2	2
8	6	7	1	1	2	1	2	10	3	1
9	1	3	1	2	1	3	2	6	2	2
10	1	3	1	1	2	1	1	6	3	3

Yine, elde edilen gizli sınıflar, değişkenlerin en yüksek olasılıkta elde ettiği değişken değerlerine göre Ek B-2’de tanımlanmaktadır.

4.2 k-modlar Algoritması ile Sonuçlar

STA, OCC, RGN, INT, MST, AGE, GEN, SEC, EDU, IG değişkenleri kullanılarak aynı veri setinde k-modlar algoritması ile kümeleme yapılarak sonuçlar GSA sonuçlarıyla karşılaştırılmak istenmiştir. k-modlar algoritması, ele alınan veri seti kategorik olduğundan k-ortalamar algoritması yerine tercih edilmiştir. k-modlar algoritması aşağıdaki kod (Tablo 4.4) yardımıyla veri setine uygulanmıştır:

Tablo 4.4 Seçilmiş 10 değişken ile k-modlar analizine ilişkin komutlar

<code>> set.seed(100000)</code>
<code>> library(klaR)</code>
<code>> c1 <- kmodes(Data_kmodes, 10, iter.max = 3000, weighted = FALSE, fast = TRUE)</code>

GYKA verisine uygulanan k-mod analizinde, her kümeye ilişkin küme büyüklükleri, mod değerleri ve basit eşleşme uzaklıkları (simple matching distances) çıktıları elde edilmiştir (Tablo 4.5):

Tablo 4.5 Küme büyüklükleri

Küme	1	2	3	4	5	6	7	8	9	10
Büyüklik	3405	7840	2893	4723	3587	1653	1899	1434	5390	3212

Tablo 4.5'e bakıldığında, ikinci küme 7840 örnek büyüklüğüne sahip olup en büyük fert sayısına sahiptir. Sekizinci kümedeki fert sayısı ise sadece 1434'tür. Tablodan hareketle verilerin kümelere dengeli dağıldığı söylenebilir. Çalışmada başlangıç olarak, kaç tane kümeye ihtiyacımız olduğuna karar vermek için, üçten fazla değişken belirlemeye çalışıldı, ancak aynı zamanda karmaşıklığı azaltıp yorumlanabilirliği artırmak için 10'dan fazla küme elde edilmemesi tercih edildi. Bu sebeple küme sayısı 10 olarak belirlenerek, GSA için kullanılan değişkenler vasıtasıyla 10 küme elde edilerek sonuçların GSA sonuçlarıyla birebir karşılaştırılması sağlandı. k-modlar ile küme modları aşağıdaki gibi elde edilmiştir (Tablo 4.6):

Tablo 4.6 k-modlar kümelemesinin küme modları

	Değişkenler									
Kümelere	STA	OCC	RGN	INT	MST	AGE	GEN	SEC	EDU	IG
1	1	1	5	1	1	2	2	6	6	5
2	6	7	1	2	2	6	2	10	1	1
3	1	5	4	2	1	1	1	2	3	5
4	4	4	6	2	1	3	1	1	2	2
5	1	4	1	2	1	2	1	2	2	4
6	1	3	3	2	1	4	2	6	2	1
7	1	2	1	1	2	2	1	6	6	4
8	1	3	3	2	2	2	1	4	3	2

Tablo 4.6 k-modlar kümelemesinin küme modları (devamı)

9	6	7	1	2	1	5	1	10	2	3
10	1	1	1	1	1	3	1	6	6	5

k-modlar analizinden elde edilen kümeler, karşılık geldikleri küme modlarına göre Ek B-3'te tanımlanmaktadır. Her kümeye ilişkin elde edilen basit-eşleşme uzaklıkları ise Tablo 4.7'de yer almaktadır:

Tablo 4.7 Kümeler için basit-eşleşme uzaklıkları

Küme	1	2	3	4	5	6	7	8	9	10
Basit eşleşme uzaklığı	12730	25458	13124	17787	13124	6791	7309	5601	15990	9815

İyi bir kümeleme metodu, küme içi uzaklıkların minimum olmasıyla yüksek kaliteli kümeler üretir. Uzaklığı ölçmek için basit eşleşme uzaklığı (simple-matching distance) kullanılmıştır. Tablo 4.5'te ikinci kümenin toplam uzaklığı 25458'dir. Sekizinci küme, en düşük basit eşleşme uzaklığına sahiptir. Uzaklığın düşük olması küme içi verilerin birbirine yüksek benzerlik gösterdiğini ifade etmektedir. Bu, sekizinci kümedeki en düşük fert sayısına bağlıdır. Her kümedeki gelir grupları ise Tablo 4.8'de elde edilmiştir:

Tablo 4.8 Her kümede gelir grubuna göre fert sayıları

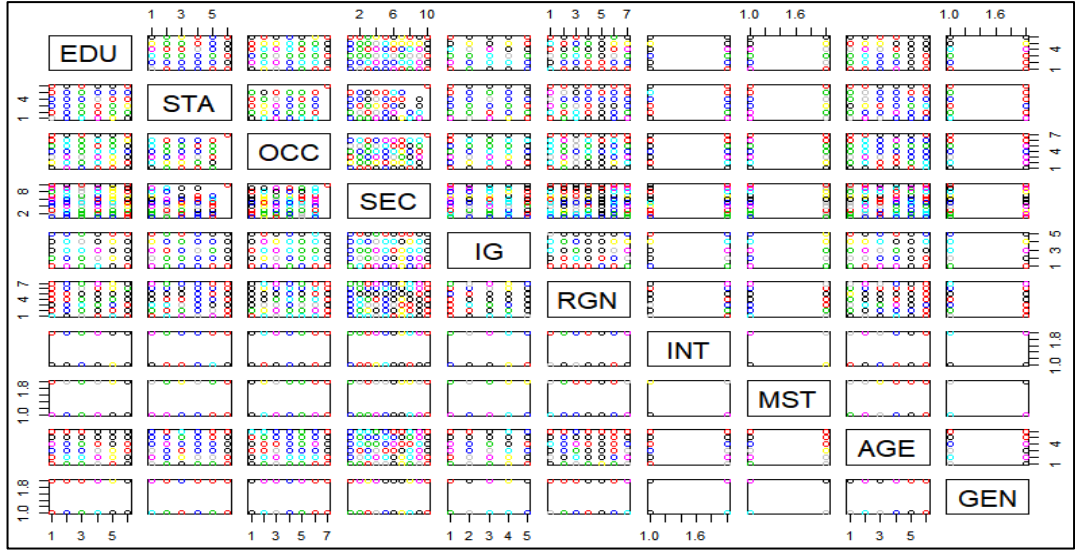
Gelir Grubu					
Küme	1. 20%	2. 20%	3. 20%	4. 20%	5. 20%
1	242	377	466	381	1939
2	4508	2343	506	386	97
3	422	458	766	426	821

Tablo 4.8 Her kümede gelir grubuna göre fert sayıları (devamı)

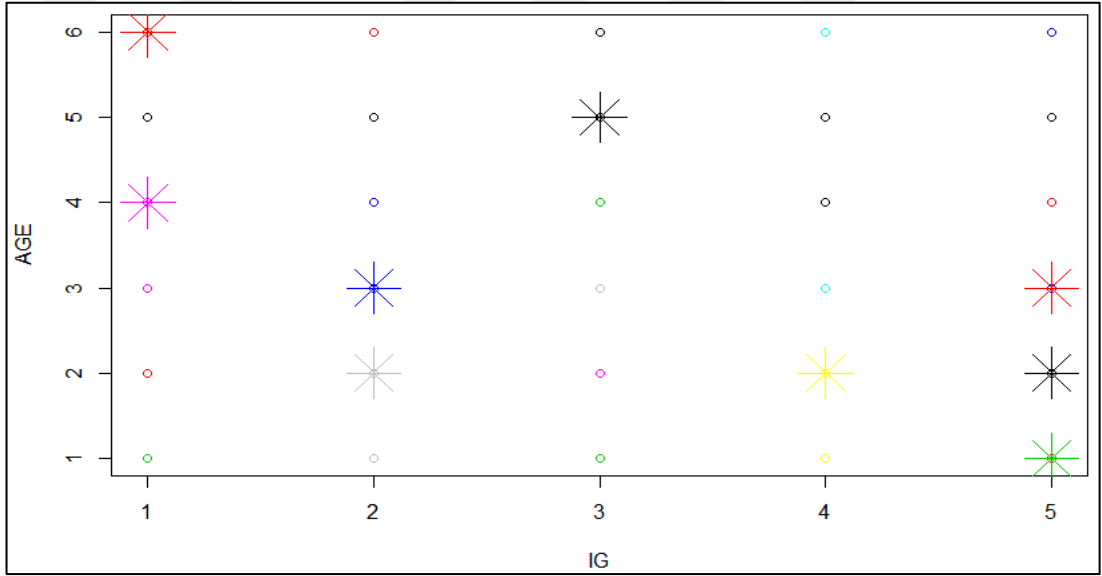
4	746	1476	780	957	764
5	141	253	799	2191	203
6	628	487	289	160	89
7	218	227	425	799	230
8	180	694	345	119	96
9	410	1343	2120	1162	355
10	49	117	246	256	2544
Toplam	7544	7775	6742	6837	7138
Toplam (%)	%20.93	%21.58	%18.71	%18.97	%19.81

Tablo 4.8’de en düşük gelir grubundakilerin çoğunun (4508) 2. kümede, en yüksek gelir grubundakilerin çoğunun (2544) ise 10. kümede yer aldığı görülmektedir. 1. ve 10. kümelerde yer alan fertlerin çoğunluğu en yüksek gelir grubunda yer alırken, 2. ve 6. kümelerdeki fertler için ise en düşük gelir grubu söz konusudur. Toplam sütunundaki fert sayılarının birbirine yakın çıkması ağırlıklı kantil yönteminde her gelir grubuna eşit fert yüzdesinin düşmesi amaçlandığı içindir. İktisat literatürüne bakıldığında, [111] kişisel gelir dağılımını da ele aldığı çalışmasında ortalama fert gelirini baz alarak %20’lik dilimlere göre gelir dağılımı sonuçlarını elde etmiştir.

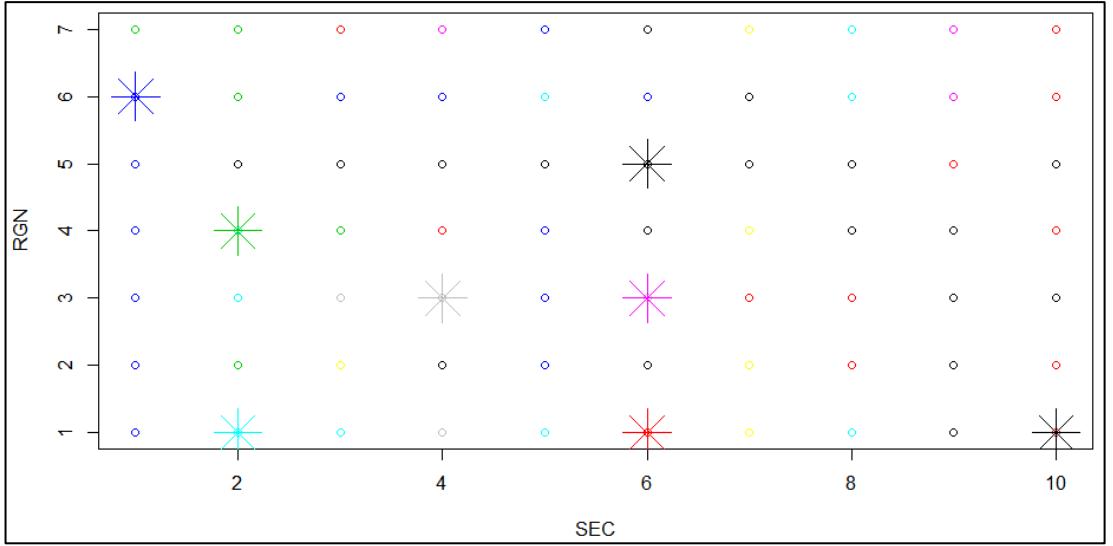
Aşağıda ise, ilk grafik (Şekil 4.2) tüm değişkenler arasındaki verilerin kümelere göre dağılımını gösterirken, diğer ikili küme grafikleri (Şekil 4.3, 4.4 ve 4.5), seçilen değişkenler arasındaki verilerin kümeler arasında nasıl dağıldığını göstermektedir:



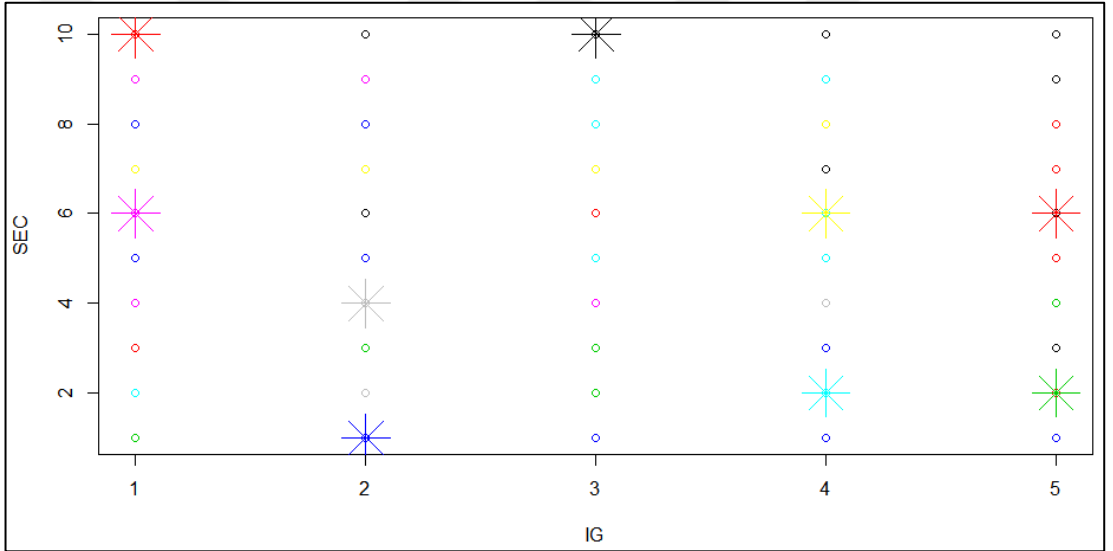
Şekil 4.2 k-modlar kümeleme analizi sonrası GYKA verisi için küme grafikleri



Şekil 4.3 AGE ve IG değişkenleri arasındaki verinin kümelerle göre dağılımı



Şekil 4.4 RGN ve SEC değişkenleri arasındaki verinin kümelerle göre dağılımı



Şekil 4.5 SEC ve IG değişkenleri arasındaki verinin kümelerle göre dağılımı

4.3 RO Algoritması ile Sonuçlar

Belirlenmiş olan değişkenler üzerinden RO sınıflama modeli kullanılarak gelecekteki hanehalkı fertleri için kural çıkarımı yoluyla kriterler oluşturularak bir sınıflama modeli elde edilmesi amacıyla “randomForest” paketi kullanılarak veriler RStudio ortamında analiz edilmiştir. RO analizinde IG (cevap değişkeni) diğer tüm dokuz değişkenin (girdi değişkenlerinin) fonksiyonu olarak seçilmiştir. IG faktör değişken olduğundan bu RO sınıflandırma (classification) şeklinde sonuçlanmıştır.

4.3.1 Eğitim Verisi için RO Modeli

Ağaçların sayısı (ntree değeri) varsayılan olarak 500'dür. Her bölünmede kullanılan değişken sayısı ise (No. of variables tried at each split, yani m_{try} değeri) varsayılan olarak toplam değişken sayısının karekökü (9 değişkenin karekökü 3) şeklindedir. Eğitim verisi için IG gelir grubu değişkeni baz alınarak RO algoritması uygulandığında elde edilen hata matrisi aşağıda (Tablo 4.9) yer almaktadır:

Tablo 4.9 Eğitim verisinin RO modelinde hata matrisi

Gelir Grubu	1. kantil	2. kantil	3. kantil	4. kantil	5. kantil	Sınıflama Hatası
1. kantil	3332	1207	401	230	133	%37.2
2. kantil	1225	2283	1057	653	288	%58.5
3. kantil	499	1150	1593	1165	409	%66.9
4. kantil	238	647	986	1951	965	%59.2
5. kantil	90	174	340	963	3406	%31.5

Tablo 4.9'da eğitim verisi için RO modeli uygulandığında, OOB değeri 50.5 çıkmış ve modelin doğruluk oranı da bu dolayda elde edilmiştir. Hata matrisine bakıldığında, tahminlerin 5. gelir grubunu tahmin etmede iyi olduğu görülmektedir. En yüksek hata %66.9 olarak 3. gelir grubunu tahmin etmede görülmektedir.

4.3.2 Eğitim Verisi için Hata Matrisi ve Tahmin

Tablo 4.10 Eğitim verisinin tahmini için hata matrisi

Gelir Grubu	1. kantil	2. kantil	3. kantil	4. kantil	5. kantil	Sınıflama Hatası
1. kantil	3973	879	342	161	66	%26.71

Tablo 4.10 Eğitim verisinin tahmini için hata matrisi (devamı)

2. kantil	834	3365	746	395	107	%38.22
3. kantil	250	688	2824	584	238	%38.39
4. kantil	150	397	662	3140	578	%36.27
5. kantil	96	177	242	507	3984	%20.42

Tablo 4.10'da, çaprazdaki değerler verinin çoğunun doğru sınıflandırıldığını göstermektedir. Örneğin verilerin 3973 tanesinin 1. gelir düzeyi olarak sınıflandırıldığı, yine 2. gelir düzeyi için 3365 tane doğru sınıflandırma yapıldığı görülmektedir. Çapraz olmayanlar ise yanlış sınıflandırılanlar şeklinde yer almaktadır. Çapraz diyagonal değerleri toplanıp eğitim verisinin sayısına (25385) bölüldüğünde doğruluk değerini (%68.1) vermektedir. %50 üzeri bir doğruluk değeri olduğu görülmektedir. Buradaki doğruluk değerinin (%68.1) yüksek çıkmasının sebebi tüm eğitim verisinin modelde kullanılmış olmasından kaynaklıdır (yani bir önceki 1-OOB (1-50.5%) değerinden daha yüksek çıkmıştır). Ayrıca doğruluk için elde edilen (0.6752, 0.6867) güven aralığının da dar bir güven aralığı olduğu, böylece sonuçların iyi olduğu söylenebilir. Bunun yanında Kappa değeri de 0.60 olarak önemli bir değerde sonuçlanmıştır.

4.3.3 Test Verisi için Hata Matrisi ve Tahmin

Tablo 4.11 Test verisi tahmini için hata matrisi

Gelir Grubu	1. kantil	2. kantil	3. kantil	4. kantil	5. kantil	Sınıflama Hatası
1. kantil	1407	480	208	107	38	%37.19
2. kantil	527	985	424	271	73	%56.80
3. kantil	141	433	676	454	132	%63.18
4. kantil	101	250	458	832	399	%59.22

Tablo 4.11 Test verisi tahmini için hata matrisi (devamı)

5. kantil	65	121	160	386	1523	%32.46
------------------	----	-----	-----	-----	-------------	---------------

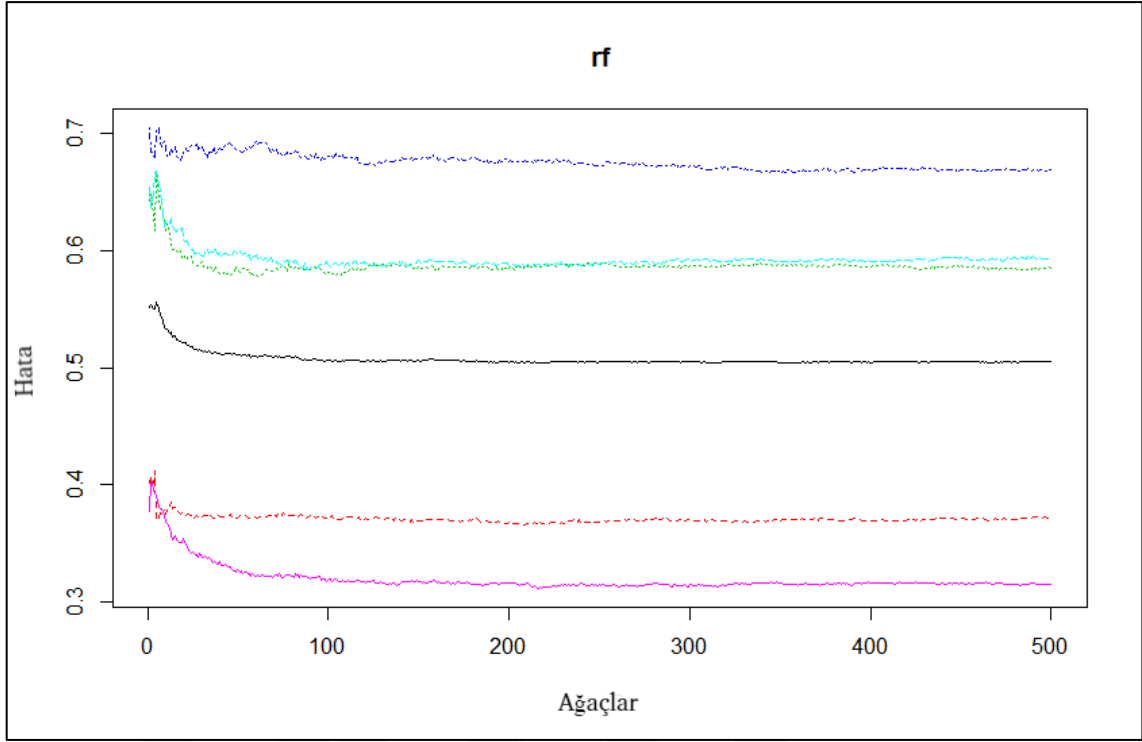
Tablo 4.12 Test verisi sınıfları için duyarlılık değerleri

Sınıf	1. kantil	2. kantil	3. kantil	4. kantil	5. kantil
Duyarlılık	0.6278	0.43411	0.35099	0.40585	0.7035

Test verisi ile tahminde doğruluk oranı eğitim verisine göre düşük (%50.92) gerçekleşmiştir. Test verisi RO modeli tarafından eğitilmemiş olduğundan, bunun, model doğruluğu için daha doğru bir değerlendirme olduğu söylenebilir. Güven aralığı ise (0.4996, 0.5187) olarak elde edilmiş olup hala dar ve iyi olduğu söylenebilir. Bunun yanında, öncekine göre hata matrisinde daha az doğru sınıflandırma olduğu (Tablo 4.11) görülmektedir. Duyarlılık değerleri hata matrisindeki ilgili sınıf değerinin ilgili sütun toplamına bölünmesiyle oluşmaktadır. Tablo 4.12'deki duyarlılık değerlerine bakıldığında Sınıf 5'i tahmin etmede en iyi duyarlılık değerinin %70.4 olduğu görülmektedir ve diğer sınıflara göre daha yüksektir.

4.3.4 Rastgele Ormanın Hata Oranı

RO modelindeki hata oranı RO modelini kullanarak (eğitim verisi üzerinden) grafik yoluyla öğrenilebilmektedir:

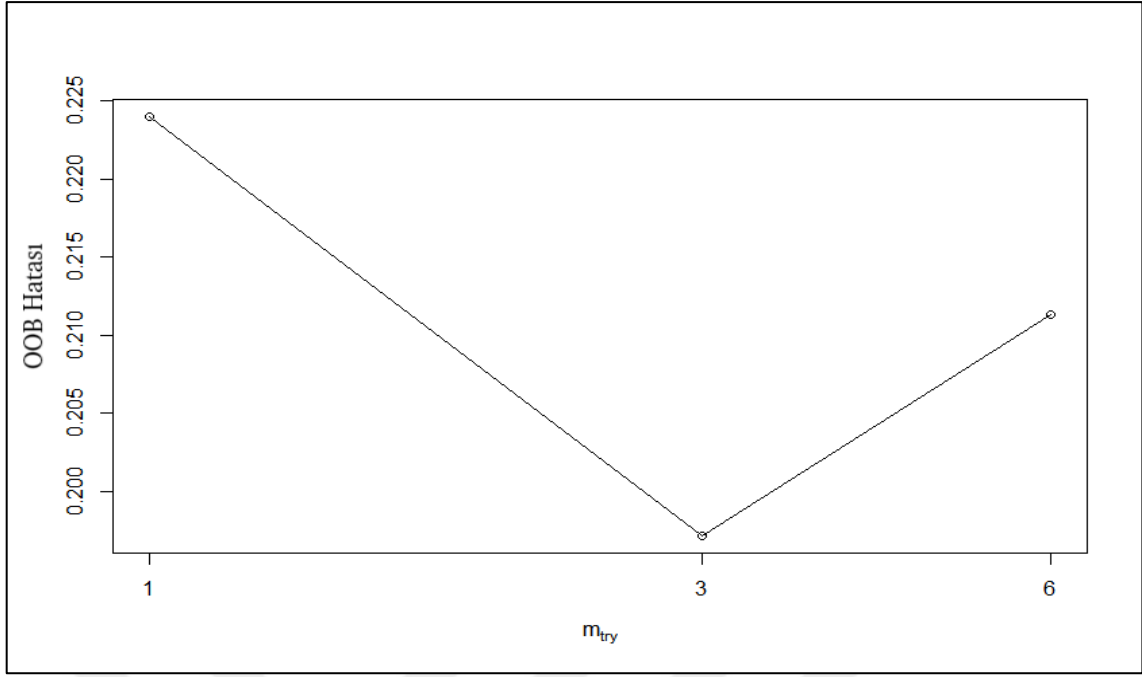


Şekil 4.6 RO hata oranı

Grafikte (Şekil 4.6) ağaç sayısı arttıkça OOB hata oranının giderek düştüğü ve sabitleştiği görülmektedir. Bu hatanın yaklaşık 300 ağaçtan sonra (farklı değerler denenerek OOB hata oranının en düşük olarak 310 ağaçta gerçekleştiği görülmüştür) sabilleştiği ve bu değerden sonrasının modeli iyileştiremediği görülmektedir.

4.3.5 RO Modelinin İyileştirilmesi (Tune İşlemi)

m_{try} değerinin iyileştirilmesi amacıyla OOB hata oranı m_{try} değerinin fonksiyonu olarak grafikleştirilmiştir (Şekil 4.7):



Şekil 4.7 En uygun m_{try} değerinin belirlenmesi

OOB hatasının başlangıç olarak $m_{try} = 1$ 'de oldukça yüksek, $m_{try} = 3$ 'te ise en düşük değerinde ve $m_{try} = 6$ olduğunda ise tekrar arttığı görülmektedir. Dolayısıyla bu grafikten m_{try} değerinin 3 olarak seçilmesi gerektiği görülmektedir. Başlangıçtaki RO modeline dönerek $n_{tree}=310$ ve $m_{try} = 3$ parametrelerini de ekleyerek eğitim verisinde RO modeli uygulandığında aşağıdaki hata matrisi (Tablo 4.13) elde edilmiştir:

Tablo 4.13 Eğitim verisi için iyileştirilmiş modelin hata matrisi

Gelir Grubu	1. kantil	2. kantil	3. kantil	4. kantil	5. kantil	Sınıflama Hatası
1. kantil	3348	1184	402	235	134	%36.87
2. kantil	1234	2276	1072	645	279	%58.66
3. kantil	508	1137	1592	1178	401	%66.94
4. kantil	239	645	993	1950	960	%59.26
5. kantil	93	174	341	952	3413	%31.37

İlk baştaki modelde OOB hatası %50.5'ti, şimdi ise %50.45'e düştü. 1. ve 5. sınıfı tahmin etmede sınıflama hatasının iyileştiği (ilk baştaki modele göre düşmüş) görülmektedir. Eğitim verisinin hatası (training error) için bakılmak istendiğinde aşağıdaki hata matrisine (Tablo 4.14) ulaşılmıştır:

Tablo 4.14 Eğitim verisinin tahmini için iyileştirilmiş modelin hata matrisi

Gelir Grubu	1. kantil	2. kantil	3. kantil	4. kantil	5. kantil	Sınıflama Hatası
1. kantil	3973	879	342	161	66	26.71%
2. kantil	835	3365	746	393	106	38.20%
3. kantil	250	688	2826	584	239	38.39%
4. kantil	149	398	661	3140	577	36.24%
5. kantil	96	176	241	509	3985	20.41%

310 ağaç için doğruluk değeri (%68.11) olarak daha yüksek (önceki %68.10) elde edilmiştir. Asıl test, test datasına dayalı olacak, buna göre hata matrisi aşağıdaki gibi elde edilmiştir (Tablo 4.15):

Tablo 4.15 Test verisi için iyileştirilmiş modelin hata matrisi

Gelir Grubu	1. kantil	2. kantil	3. kantil	4. kantil	5. kantil	Sınıflama Hatası
1. kantil	1410	479	210	104	38	37.08%
2. kantil	517	980	430	270	76	56.89%
3. kantil	146	431	664	449	131	63.54%
4. kantil	105	258	462	845	401	59.20%
5. kantil	63	121	160	382	1519	32.34%

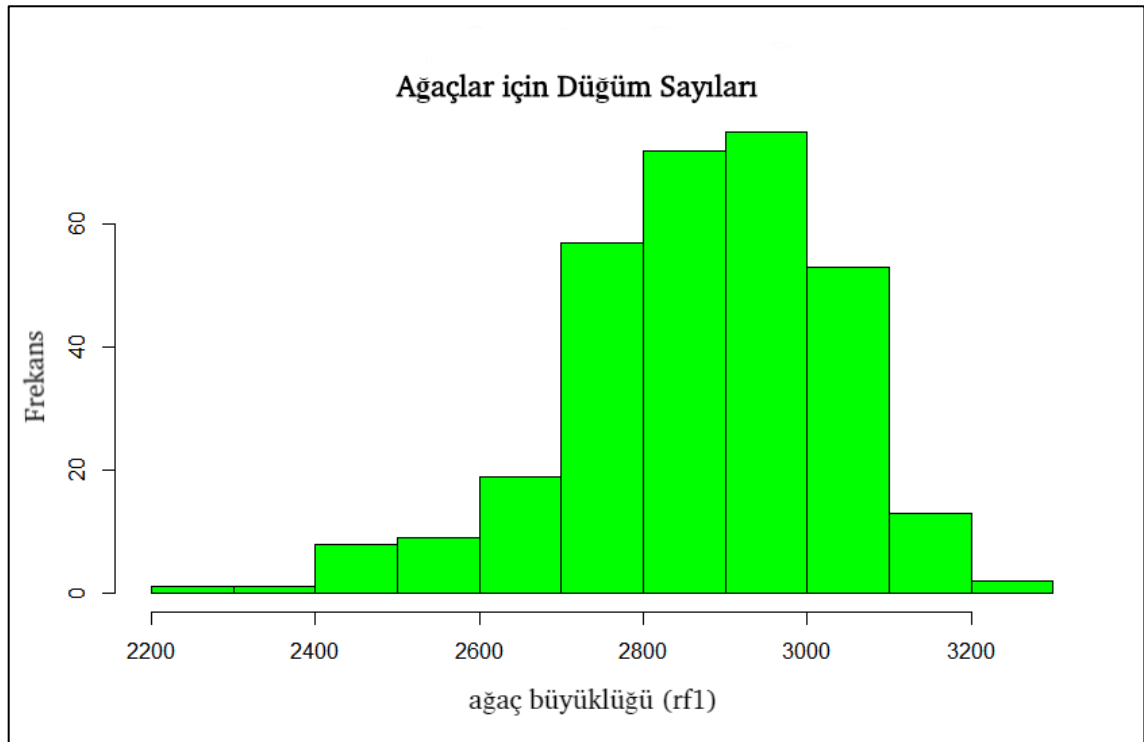
Tablo 4.16 Test verisi sınıfları için iyileştirilmiş modelin duyarlılık değerleri

Sınıf	1. kantil	2. kantil	3. kantil	4. kantil	5. kantil
Duyarlılık	0.6292	0.43191	0.34476	0.41220	0.7016

Tablo 4.16’da önceki sonuçlara göre Sınıf 1 ve 4’te duyarlılık (sensitivity) değerlerinde artış (iyileşme) bulunduğu görülmektedir.

4.3.6 Ağaçlar için Düğüm (Node) Sayısı

Ağaçların sayısal büyüklüğü bir histogram kullanılarak elde edilmiştir:

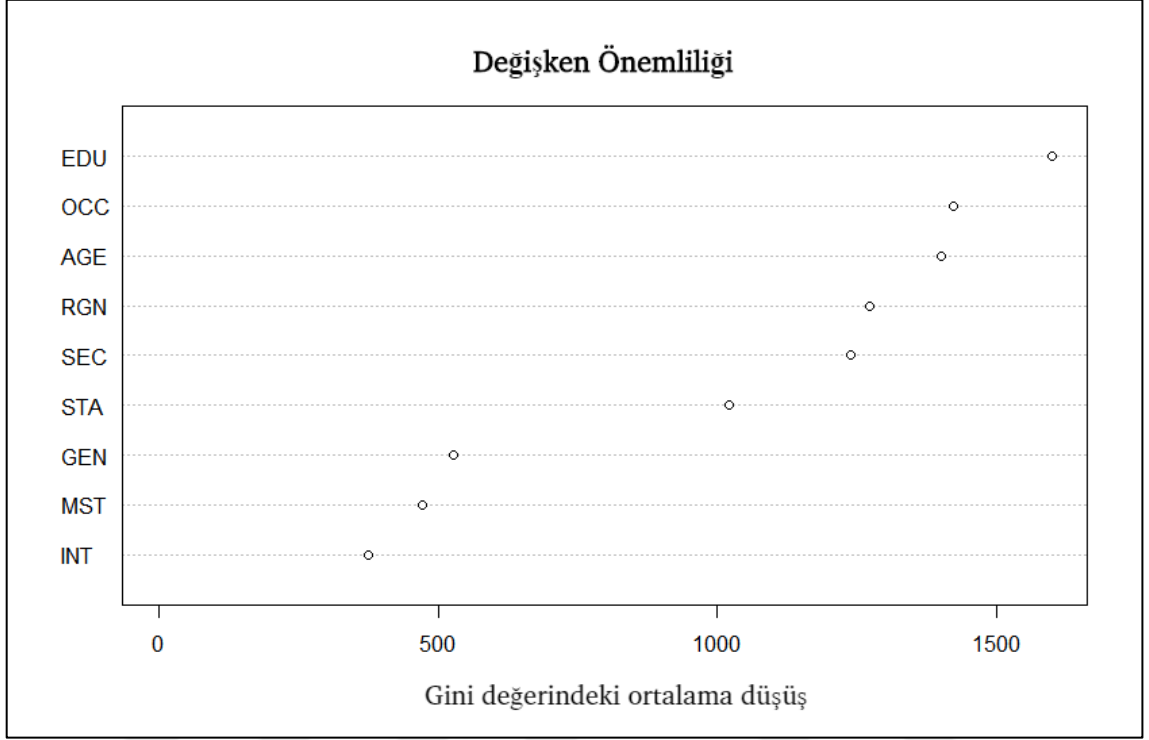


Şekil 4.8 Ağaçlar için düğüm sayıları

Şekil 4.8’de 310 ağacın her birinde bulunan düğüm sayıları görülmektedir. En büyük histogram barının frekansının 60’a yakın olduğu görülmektedir. 2900’den fazla düğüm içeren yaklaşık 60 ağaç bulunmaktadır. Ayrıca 2200’den az veya 3200’den fazla düğüm içeren çok az ağaç vardır. Ağaçların çoğunluğunun 2900 düğüme sahip olduğu söylenebilir.

4.3.7 Değişken Önemliliği

Modelde hangi değişkenlerin önemli rol oynadığını görebilmek için değişken önemliliği grafiği (variable importance plot) elde edilmiştir:



Şekil 4.9 Değişken önemliliği

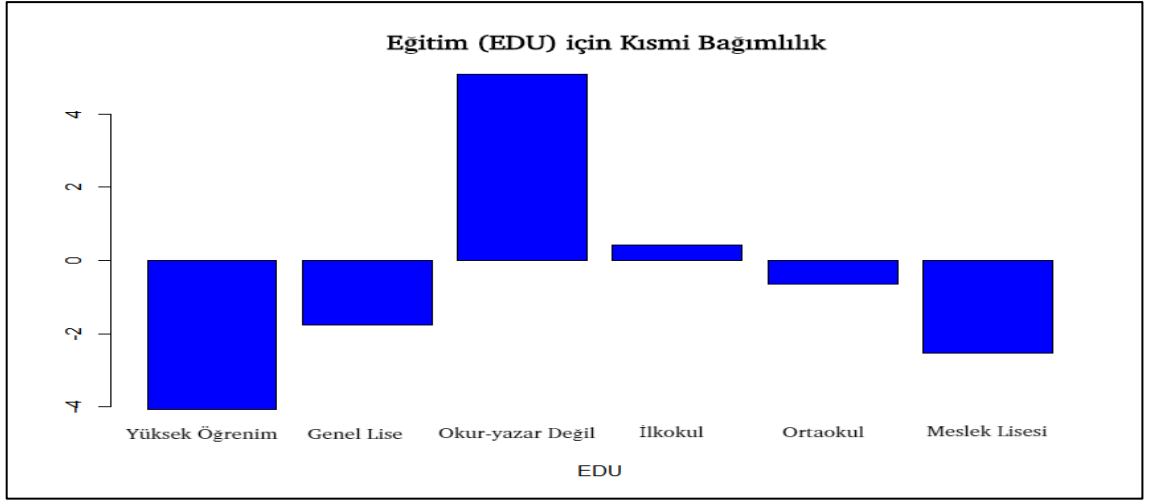
Şekil 4.9'daki grafik, her değişken çıkarıldığında düğümlerin ağacın sonunda ne kadar saf olduğunu göstermektedir. Örneğin EDU değişkeni kaldırıldığında ortalama Gini değeri 1500 kadar düşmektedir. EDU, OCC ve AGE değişkenlerinin diğer değişkenlere göre Gini parametresine en çok katkıyı yaptığı görülmektedir. Tablo 4.17'de, 9 girdi değişkeni için grafiği oluşturan veri noktalarını ifade eden gerçek değerler "Gini İndeksindeki Ortalama Düşüş" satırında yer almaktadır. "Bulunma Sıklığı" değerleri ise her değişkenin RO modelinde kaç kez bulunduğunu/ne sıklıkta görüldüğünü (ağaçlardaki frekansları) göstermektedir. Örneğin EDU değişkeni modelde 140887 kez bulunmuştur.

Tablo 4.17 Gini indeksindeki ortalama düşüş ve bulunma sıklığı

Değişken	Gini İndeksindeki Ortalama Düşüş	Bulunma Sıklığı
EDU	1598.37	140887
STA	1020.4	61423
OCC	1421.57	104829
SEC	1237.46	142353
RGN	1273.45	198165
INT	374.19	52228
MST	471.47	33863
AGE	1401.21	129448
GEN	526.98	26709

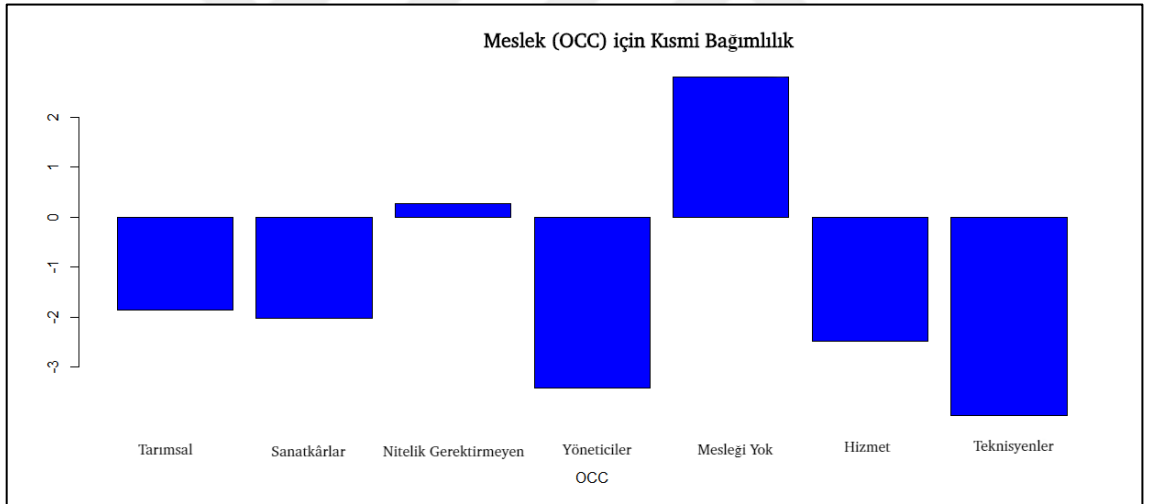
4.3.8 Kısmi Bağımlılık Grafikleri

Kısmi bağımlılık grafikleri sınıflama olasılığına (sınıflamaya) bir değişkenin marjinal etkisinin grafiksel tasvirini verir. Örneğin EDU ve OCC değişkenlerine Sınıf 1 bazında bakıldığında Şekil 4.10 ve 4.11, Sınıf 5 bazında ise Şekil 4.12 ve 4.13 elde edilmektedir.



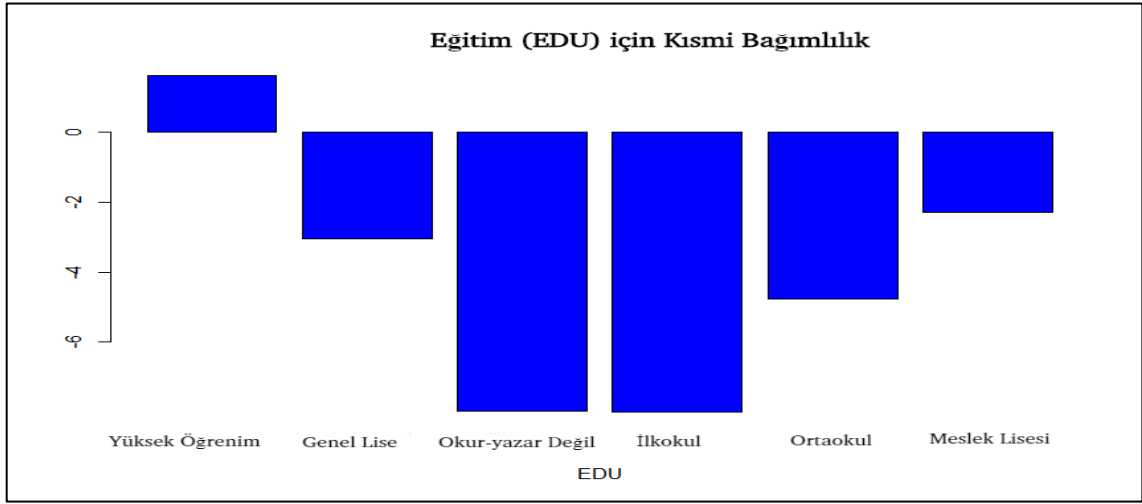
Şekil 4.10 “EDU” değişkeni için sınıf 1 bazında kısmi bağımlılık

Şekil 4.10’da EDU değişkeni “Illiterate” (Okur yazar olmayan / bir okul bitirmeyen) değerini aldığı anda, diğer aldığı değerlere göre Sınıf 1’i (en düşük gelir düzeyi) daha güçlü tahmin etme eğilimindedir.



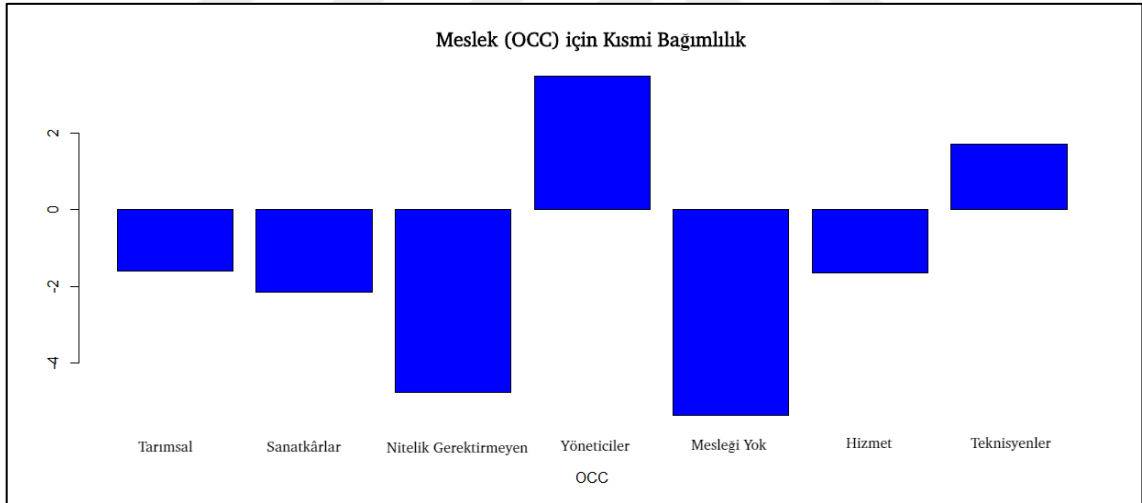
Şekil 4.11 “OCC” değişkeni için sınıf 1 bazında kısmi bağımlılık

Şekil 4.10’da OCC değişkeni “Mesleği Yok” değerini aldığı anda, diğer aldığı değerlere göre Sınıf 1’i (en düşük gelir düzeyi) daha güçlü tahmin etme eğilimindedir.



Şekil 4.12 “EDU” değişkeni için sınıf 5 bazında kısmi bağımlılık

Şekil 4.11’de EDU değişkeni “Yüksek Öğrenim” değerini aldığı anda, diğer aldığı değerlere göre Sınıf 5’i (en yüksek gelir düzeyi) daha güçlü tahmin etme eğilimindedir.



Şekil 4.13 “OCC” değişkeni için sınıf 5 bazında kısmi bağımlılık

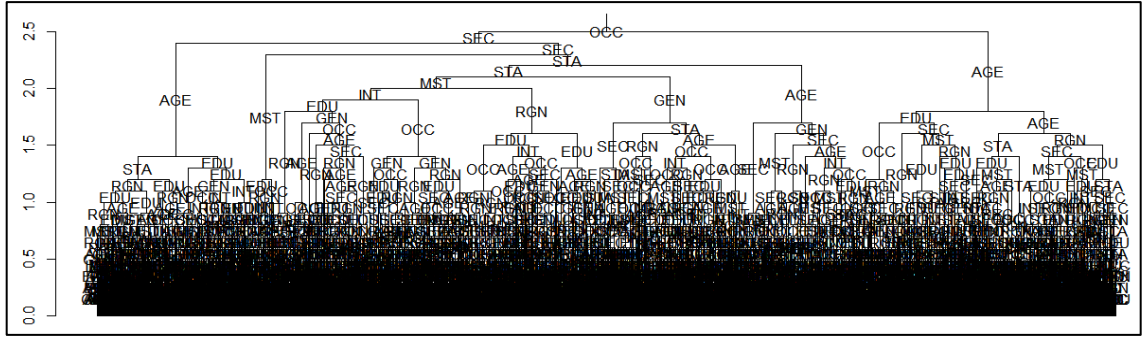
Şekil 4.12’de OCC değişkeni “Yöneticiler, Profesyonel Meslekler” değerini aldığı anda, diğer aldığı değerlere göre Sınıf 5’i (en yüksek gelir düzeyi) daha güçlü tahmin etme eğilimindedir.

4.3.9 RO Modelinden Tek Bir Ağacın Seçilerek Elde Edilmesi

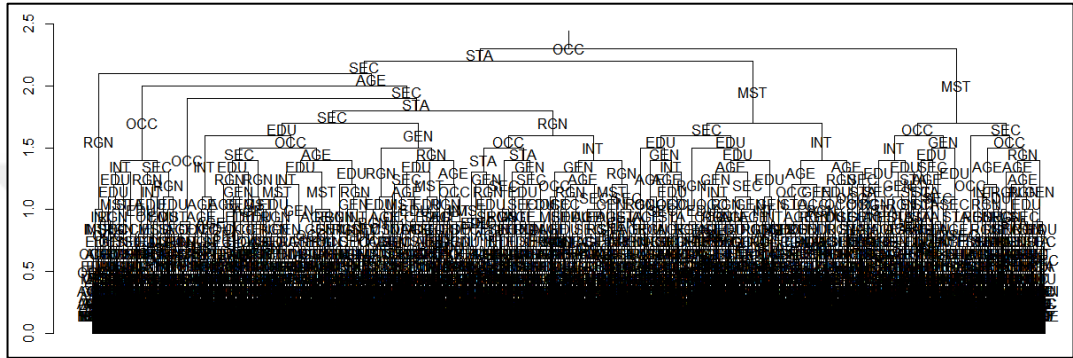
	left	daughter	right	daughter	split	var	split	point	status	prediction
1	2	3	OCC	55	1	<NA>				
2	4	5	SEC	40	1	<NA>				
3	6	7	AGE	3	1	<NA>				
4	8	9	AGE	7	1	<NA>				
5	10	11	SEC	64	1	<NA>				
6	12	13	EDU	1	1	<NA>				
7	14	15	AGE	24	1	<NA>				
8	16	17	STA	5	1	<NA>				
9	18	19	EDU	4	1	<NA>				
10	20	21	MST	1	1	<NA>				
11	22	23	STA	26	1	<NA>				
12	24	25	OCC	8	1	<NA>				
13	26	27	SEC	68	1	<NA>				
14	28	29	STA	10	1	<NA>				
15	30	31	RGN	62	1	<NA>				
16	32	33	RGN	30	1	<NA>				
17	34	35	EDU	12	1	<NA>				
18	36	37	AGE	16	1	<NA>				
19	38	39	EDU	26	1	<NA>				
20	40	41	EDU	3	1	<NA>				
21	42	43	OCC	5	1	<NA>				
22	44	45	STA	8	1	<NA>				
23	46	47	AGE	1	1	<NA>				
24	48	49	MST	1	1	<NA>				
25	50	51	RGN	47	1	<NA>				
26	52	53	RGN	23	1	<NA>				
27	54	55	MST	1	1	<NA>				
28	56	57	EDU	1	1	<NA>				
29	58	59	STA	16	1	<NA>				
30	60	61	SEC	402	1	<NA>				
31	62	63	EDU	17	1	<NA>				
32	64	65	EDU	20	1	<NA>				
33	66	67	EDU	11	1	<NA>				
34	68	69	AGE	2	1	<NA>				
35	70	71	RGN	47	1	<NA>				
36	72	73	SEC	8	1	<NA>				
37	74	75	SEC	8	1	<NA>				
38	76	77	GEN	1	1	<NA>				
39	78	79	INT	1	1	<NA>				
40	80	81	OCC	5	1	<NA>				
41	82	83	RGN	55	1	<NA>				
42	84	85	EDU	2	1	<NA>				
43	86	87	INT	1	1	<NA>				
44	88	89	MST	1	1	<NA>				
45	90	91	GEN	1	1	<NA>				
46	92	93	SEC	385	1	<NA>				
47	94	95	GEN	1	1	<NA>				
48	96	97	SEC	402	1	<NA>				
49	0	0	<NA>	0	-1	5_Quantile				
50	98	99	AGE	1	1	<NA>				

Şekil 4.14 Tek bir ağaca ilişkin RO ağaç çıktısı

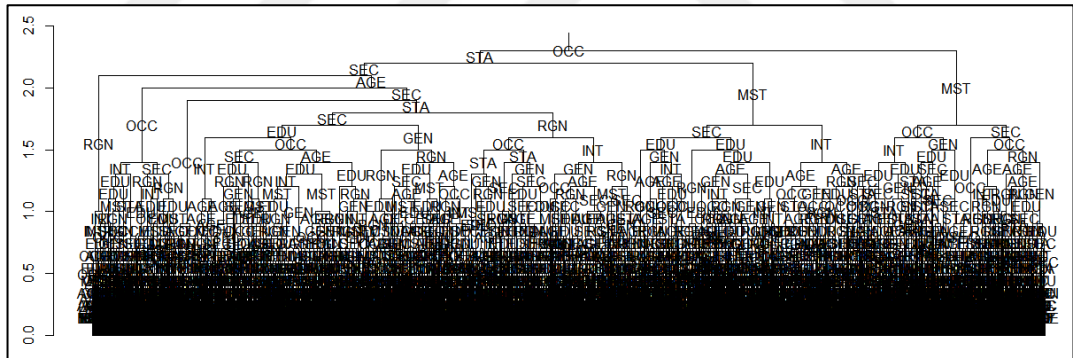
Şekil 4.14’te, tek bir ağaca ilişkin elde edilen ağaç çıktısının bir kesiti yer almaktadır. Burada “status” sütunundaki “-1” değeri uç düğümü (terminal node) ifade etmektedir. RO modelinin ilk dört ağacı ise aşağıdaki şekillerde (Şekil 4.15, 4.16, 4.17 ve 4.18) görselleştirilmiştir:



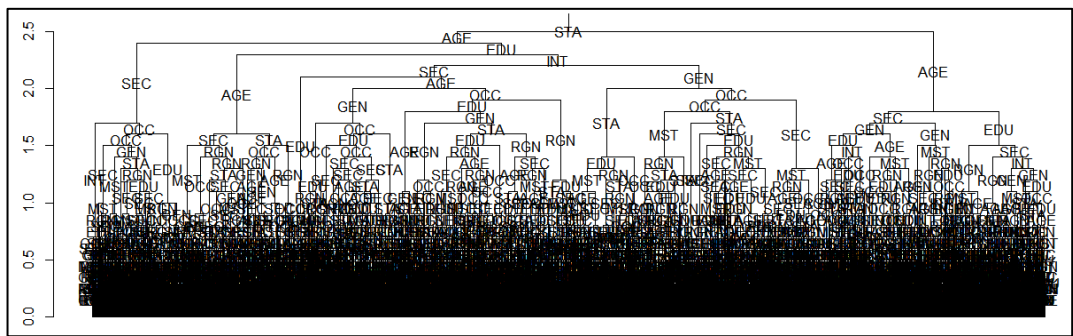
Şekil 4.15 RO modelinin ilk ağacı



Şekil 4.16 RO modelinin ikinci ağacı



Şekil 4.17 RO modelinin üçüncü ağacı



Şekil 4.18 RO modelinin dördüncü ağacı

Fert ve hanehalkları için sosyoekonomik durum ve yaşam standartları açısından önemli bir gösterge olan gelir, bireyin refah durumunu yansıtması ve karar alıcılar açısından etkin sosyoekonomik politikaların geliştirilmesi yönünden büyük öneme sahiptir. Gelir dağılımı araştırmaları, fert ve hanehalkları arasındaki gelir dağılımını incelemek ve karar alıcıların sosyal yapının iyileştirilmesi yönünde kararlar alması için önem teşkil etmekle birlikte en önemli makro ekonomik konulardan biri konumundadır.

Türkiye’de son 70 yılda birçok kişi veya kuruluş tarafından gelir dağılımı araştırmaları gerçekleştirilmiştir. Dünyada ÜİE’ler tarafından üretilen gelir dağılımına ilişkin veriler devletin ekonomik ve sosyal karar alma sürecinde önemli bir yer teşkil etmekte, ülkemizde ise TÜİK, gelir dağılımı araştırmalarını yapan kuruluşların başında gelmektedir.

Günümüz koşullarında büyüyen ve genişleyen veri hacmiyle birlikte, resmi istatistiklerden yeni bilgi ve gizli kalıpların elde edilmesi ve ileriye yönelik tahminlerin yapılması ekonomik ve sosyal karar alıcılar açısından önem teşkil etmektedir. Halen mevcut geleneksel istatistiksel yöntemler resmi istatistiklerden anlamlı bilgiler çıkarmak için yetersiz kalmaktadır. Büyük ve kapsamlı bir resmi istatistik verisi olan GYKA verisini VM yöntemleri ile incelemek, resmi istatistikler konusunda yeni VM yöntemlerinin geliştirilmesini kolaylaştıracak ve ele alınan VM yöntemlerinin mevcut resmi istatistikler için kullanılabilirliğini artıracaktır. Bu amaçla bu çalışmada TÜİK’den temin edilen 2015 yılı GYKA kesit verisi uygun VM yöntemleri kullanılarak analiz edilmiştir. Çalışmada, Türkiye’deki fertlerin gelir durumlarını sosyoekonomik değişkenler ışığında irdelemek ve fertleri sosyoekonomik profillerine ve refah durumlarına göre VM yöntemleriyle sınıflandırmak amaçlanmıştır. Oldukça kapsamlı bir veri olan GYKA verisini VM yöntemleriyle ele almadan önce hedef değişken olarak fert geliri ve fert gelirini

etkileyen deęişkenlerin belirlenmiřtir. Fertlerin yıllık gelirine etki eden deęişkenlerin fert gelirini ne düzeyde etkiledikleri, bu deęişkenlerin kendi aralarındaki iliřki durumları ve bu deęişkenler için ne tür kural ve genellemelerin ortaya konulabileceęi arařtırılmak istenmiřtir. Bireysel geliri etkileyen deęişkenler belirlenerek bireylerin refah durumu VM yöntemleriyle arařtırılmıřtır. Veri seti istenen deęişkenlere göre düzenlendikten sonra, veri setine en uygun VM yöntemleri arařtırılmıřtır. Veri setinde deęişken tipleri kategorik olduęundan kategorik verilere uygun VM yöntemleri ele alınmıřtır. Veri seti üzerinde kümeleme ve sınıflandırma yöntemlerinin uygulanması amaçlanmıř ve kategorik veri setlerine en uygun olan VM yöntemlerinden GSA ve k-modlar kümeleme yöntemleri kullanılarak bireylerin sosyoekonomik profilleri belirlenmiřtir. Ayrıca RO yöntemi kullanılarak, bireysel gelirin baęımlı deęişken olarak ele alındıęı modelde geliri etkileyen deęişkenlerin modeldeki önem düzeyleri belirlenmiř ve bireylerin refah durumu için bir sınıflama modeli elde edilmiřtir.

GSA, resmi istatistiklere birçok uygulaması olan olasılık temelli oldukça güçlü bir tekniktir. GSA yöntemi karmařık olmakla birlikte yaygın olarak kullanılan bir teknik olduęundan, kolay uygulanabildięinden ve yüksek kaliteli anlamlı istatistikler üretmesinden dolayı ele alınan kategorik veriyi kümelemek ve deęişkenlerden gizli sınıfların elde edilebilmesi amacıyla tercih edilmiřtir. Dięer yandan kategorik verileri kümelemede alternatif bir yöntem olarak k-ortalamlar yönteminin bir uzantısı olan k-modlar yöntemi de tercih edilmiř ve GSA yönteminin sonuçları ile karřılařtırılarak her iki yöntemle iliřkin kümeleme sonuçlarının tutarlılıęı görölmek istenilmiřtir. Aynı veri setine uygulanan farklı kümeleme yöntemleri çok farklı sonuçlar ortaya çıkarabilmektedir. Dolayısıyla farklı veri tipleri farklı yöntemler gerektireceęinden, uygulamada hangisinin iře yarar olduęunu önermek oldukça zordur ve hangisinin etkin bir yöntem olduęu belirsizdir.

Yapılan analizler öncesi veri seti düzenlendikten sonra deęişken seçim iřlemi yapılmıřtır. Çalışmada ferdin yıllık geliri baęımlı deęişken olarak belirlendięinden fert gelirini etkileyen dięer veri seti deęişkenleri belirlenmeye çalışılmıřtır. Bu amaçla öncelikle 20 adet baęımsız deęişken belirlenmiř, daha sonra ise analiz

öncesi karmaşıklığı azaltmak için sırasıyla Boruta ve LCAvarsel R paketleri kullanılarak değişken sayısı azaltılmaya çalışılmıştır. Yine poLCA R paketi ile yapılan GSA denemelerinde gizli sınıflara göre değişim göstermeyen değişkenler veri setinden çıkarılmış ve uzman görüşüyle beraber 10 değişkenden oluşan nihai veri seti elde edilmiştir. Analizlerin karşılaştırılabilir olması açısından GSA, k-modlar ve RO yöntemleri için bu nihai veri seti kullanılmıştır. Değişken seçim işlemlerinde paket programlar her zaman anlamlı ya da istenilen değişkenleri veremeyebilmektedir. Bu çalışmada da değişken sayısını indirmek için kullanılan bu yöntemler sadece bir kısım değişkeni bağımlı değişkenle anlamlı olacak şekilde ortaya koyduğundan dolayı uzman görüşüne de başvurularak daha tutarlı bir değişken kümesi elde edilmiştir.

GSA analizinin uygulamasında başlangıç sınıf sayısı 30 olarak belirlenmesine rağmen yerel minimum değerine ulaşamamış olup yorumlanabilir sonuçlar elde edebilmek için sınıf sayısının 10 olarak belirlenmesine karar verilmiştir. GSA analizinde “Estimated class population shares” sonuçlarına bakıldığında, 10 sınıflı model çözümünde en çok gözlemin sırasıyla %15,17, %14,39 ve %14,27 oranlarıyla gizli sınıf 6 (en düşük gelirli ve en eğitimsiz), gizli sınıf 1 (orta gelirli, düşük eğitilmiş yaşlı erkekler) ve gizli sınıf 4’te (en yüksek gelirli ve en eğitilmiş) yer aldığı görülmüştür. Yine en az gözlemin sırasıyla %5,52 ve %5,53 oranlarıyla gizli sınıf 7 (düşük gelirli, Doğu Anadolu bölgesinde inşaat sektöründe yevmiyeli çalışan) ve gizli sınıf 9 ‘da (düşük gelirli, orta yaşta evli olan bayanlar) yer aldığı görülmüştür. Yine “Estimated class population shares” değerlerine oldukça yakın çıkan “Predicted class memberships” değerleri modelin veriye iyi uyum sağladığını göstermekle birlikte, aynı zamanda yeni seçilen bir ferdin hangi olasılıkla hangi sınıfa dahil olacağını ifade etmektedir. Burada, yeni seçilen bir ferdin en yüksek ve en düşük olasılıkla dahil olacağı gizli sınıflar sırasıyla gizli sınıf 6 ve gizli sınıf 9 olacaktır. GSA analizinde 10’a kadar olan gizli sınıf modelleri için elde edilen model uyum istatistiklerinde en düşük BIC bilgi kriteri değeri 10 gizli sınıflı modelde görülmüş olup bu model tercih edilip sonuçları elde edilmiştir. Yine GSA analizinde değişkenlere ilişkin gizli sınıf olasılıkları (Ek B1) elde edilmiş ve her gizli sınıfta en yüksek olasılıkla yer alan ilgili değişken değerleri belirlenmiştir. Bu

değişken değerleri vasıtasıyla gizli sınıflar Ek B2’de tanımlanmıştır. Bu gizli sınıflar incelendiğinde gizli sınıfların tanımlarının oldukça tutarlı olduğu görülmekte, bunun da GSA yönteminin güçlü, stabil ve olasılıksal olmasından kaynaklandığı belirtilebilir.

Karşılaştırılabilirlik açısından GSA analizi için kullanılan veri seti k-modlar algoritması için de kullanılmıştır. Yine k-modlar analizinde de yorumlanabilirliği artırmak için küme sayısı 10 ile sınırlandırılmıştır, böylece GSA analizindeki sınıf sayısı ile de uyum sağlanmıştır. Elde edilen kümeler küme modlarına göre Ek B3’te tanımlanmıştır. k-modlar analizi sonuçları incelendiğinde en çok gözlemin sırasıyla 2. küme (en düşük gelirli ve en eğitimsiz) ve 9. kümede (orta gelirli, eğitim seviyesi düşük yaşlı erkekler) yer aldığı görülmüştür. Yine en az gözlemin sırasıyla 8. kümede (düşük gelirli, ticaret sektöründe ücretli çalışan evli olmayan genç erkekler) ve 6. kümede (düşük gelirli orta yaşlı bayanlar) yer aldığı görülmektedir. Bu sonuçlara göre k-modlar analizinin GSA analiziyle benzer sonuçlar ürettiği görülmektedir. Her iki analizde de en çok gözlemin “en düşük gelirli ve en eğitimsiz” kategorisine düştüğü, yine “düşük gelirli orta yaşlı bayanlar” olarak tanımlanabilecek kategoride ise az sayıda gözlemin mevcut olduğu ifade edilebilir.

Ek B-2 ve Ek B-3’te yer alan sınıf/küme tanımları sonuçlarına bakıldığında, GSA ve k-modlar sınıf/küme çiftlerinden sırasıyla 6-2, 4-1, 1-9, 3-5, 9-6 çiftlerinin birbirlerine yakın benzerlik gösterdiği, yine 2-8, 5-4 çiftlerinin ise düşük benzerlik gösterdiği görülmektedir. k-modlar analizi sonuçları için 3., 7. ve 10. küme tanımlarının GSA sınıf tanımlarından herhangi biriyle benzerlik göstermediği ve bu küme tanımlarının tutarlı olmadığı görülmektedir. Bu sonuçlardan, kategorik GYKA verisinde GSA yönteminin k-modlar yöntemine göre daha tutarlı sonuçlar verdiği ve bu tür resmi veride başarıyla uygulanarak anlamlı sonuçlar üretebileceği görülmüştür.

Çalışmada ayrıca, 10 adet kategorik değişkenden oluşan aynı veri seti için, fert gelirinin (IG) cevap değişkeni ve diğer değişkenlerin ise girdi değişkenleri olarak belirlenerek RO analizi ile bir sınıflandırma modeli elde edilmesi ve değişkenlerin

önem düzeylerinin de görülmesi amaçlanmıştır. Bu amaçla veri seti 70/30 oranında bölünerek eğitim ile test verisi için RO modeli elde edilmiştir. RO modelinin hata oranının en düşük gerçekleştiği parametreler saptanarak RO modeli iyileştirilmiştir. Buna göre RO modelinde kullanılan girdi değişkenlerinin önem sırası belirlenerek modele en çok katkı yapan değişkenler tespit edilmiştir. Eğitim (EDU), meslek (OCC) ve yaş (AGE) değişkenlerinin modele en çok katkısı yapan en önemli girdi değişkenler olduğu görülmüştür. Dolayısıyla elde edilen anlamlı bulgular ile, ele alınan RO yönteminin de GYKA verisi için anlamlı sonuçlar üreten uygulanabilir bir VM yöntemi olduğu görülmüştür.

[90] tarafından gerçekleştirilen çalışmada fakir ve zenginlerin gelir seviyelerini bu çalışmadakine benzer şekilde yaş, eğitim, ekonomik aktivite ve hanehalkı reisinin çalıştığı sektör gibi değişkenlerin ağırlıklı olarak belirlediği görülmüştür. [89] tarafından yapılan çalışmada ise GSA'nın zengin ve fakir hane oranlarını en iyi şekilde tahmin ettiği görülmüştür. Bu çalışma, GSA yönteminin daha tutarlı bir ekonomik yorumlama sağladığına dair bulgumuzla uyumludur.

Bu çalışma GYKA verisindeki fertlerin gelir ve refah durumlarıyla sınırlandırılmış olsa da, bu çalışmada kullanılan yöntemler gelecek araştırmalar için resmi veri ile ilgili çalışmalarda başka yöntemler geliştirmek amacıyla kullanılabilir. GSA, k-modlar ve RO yöntemlerinin kullanıldığı bu çalışmada, kullanılan verinin belli bir yıla ait kategorik kesit veri olması, ele alınan değişken sayısı ve algoritmalara ilişkin parametreler bir sınırlılık oluşturmaktadır. GSA analizinde belirlenen sınıf sayısı ile k-modlar analizinde belirlenen küme sayısı da bir sınırlılık oluşturmaktadır. Ayrıca RO analizi öncesi veri seti için kullanılan 70/30 bölümlene oranının da bir kısıt oluşturduğu söylenebilir. Bu sınırlılıklara rağmen elde edilen sonuçların bireylerin gelir eksenli değişkenler etrafında tutarlı bir şekilde kümelendiği ve ülke genelinde tutarlı sosyoekonomik profillerin elde edildiği söylenebilir. Resmi veriye ait panel verilerin ya da farklı yıllara ait kategorik ya da sürekli kesit verilerin kullanıldığı çalışmalar veriye uygun farklı algoritmalar yoluyla ele alınarak gelecek araştırmalar gerçekleştirilebilir. Ayrıca bireylerin SES (socio-economic status) indeks aralıkları belirlenerek sonuçları mevcut sosyoekonomik profillerle karşılaştırılabilir. Gerçekleştirilmiş olan bu

alıřma ile VM metotlarının GYKA gibi kapsamlı bir resmi veriye nasıl uygulanabileceęi bir VM süreci olarak ayrıntılı olarak gösterilmiřtir. Elde edilen bu sonuçlar karar alıcılara fertlerin sosyoekonomik profilleri aracılığıyla yol gösterebilir.



-
- [1] TÜİK, Gelir ve Yaşam Koşulları Araştırması 2012, http://www.turkstat.gov.tr/IcerikGetir.do%3Fistab_id%3D232, Şubat 2014.
- [2] M. Kantardzic, "Data mining concepts, models, methods and algorithms," [Elektronik Sürüm], IEEE Press, Piscataway- NJ, s. 7., 2001.
- [3] Y.Z. Ayık, A. Özdemir ve U. Yavuz, "Lise türü ve lise mezuniyet başarısının, kazanılan fakülte ile ilişkisinin veri madenciliği tekniği ile analizi," Atatürk Üniversitesi, Sosyal Bilimler Enstitüsü Dergisi, cilt. 10, no. 2, s. 441-454, 2007.
- [4] A. Koyuncugil ve N. Özgülbaş, "İMKB'de işlem gören KOBİ'lerin güçlü ve zayıf yönleri: CHAID karar ağacı uygulaması," Dokuz Eylül Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, cilt. 23, no. 1, s. 1-21, 2008.
- [5] F. Gürbüz, L. Özbakır ve H. Yapıcı, "Türkiye'de bir havayolu işletmesine ait parça söküm raporlarına ilişkin veri madenciliği uygulaması," Gazi Üniv. Müh. Mim. Fak. Dergisi, cilt. 24, no. 1, 73-78, 2009.
- [6] Ç. Taşkın ve G.G. Emel, "Veri madenciliğinde kümeleme yaklaşımları ve kohonen ağları ile perakendecilik sektöründe bir uygulama," Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, vol. 15, no. 3 s. 395-409, 2010.
- [7] M. Danacı, M. Çelik ve A.E. Akkaya, "Veri madenciliği yöntemleri kullanılarak meme kanseri hücrelerinin tahmin ve teşhisi," <http://iys.inonu.edu.tr/>, 2010.
- [8] M. Koç ve M. Karabatak, "Sosyal ağların öğrenciler üzerindeki etkisinin veri madenciliği kullanılarak incelenmesi," 5th International Computer & Instructional Technologies Symposium, Fırat University, Elazığ-Turkey, 2011.
- [9] S. Savaş ve N. Topaloğlu, "Veri madenciliği yöntemi ile GSM şebekelerinin performans analizi," Gazi Üniv. Müh. Mim. Fak. Dergisi, cilt. 26, no. 4, s. 741-751, 2011.
- [10] S. Irmak, C.D. Köksal ve Ö. Asilkan, "Hastanelerin gelecekteki hasta yoğunluklarının veri madenciliği yöntemleriyle tahmin edilmesi," International Journal of Alanya Faculty of Business, cilt. 4, no. 1, s. 101-114, 2012.
- [11] K. Çağdaş ve O.A. Erdem, "Öğrenci başarısını etkileyen faktörlerin veri madenciliği yöntemleriyle incelenmesi," Gazi Üniversitesi Politeknik Dergisi, cilt. 15, no. 2, s. 111-116, 2012.

- [12] M.A. Alan, "Veri Madenciliği ile Lisansüstü Öğrenci Verileri Üzerine Bir Uygulama," Dumlupınar Üniversitesi Sosyal Bilimler Dergisi, cilt. 33, no. 1, s. 165-174, 2012.
- [13] E. Güngör, N. Yalçın ve N. Yurtay, "Apriori algoritması ile teknik seçmeli ders seçim analizi," Akademik Bilişim 2013 Konferansı Bildirileri, 2013.
- [14] E. Yakut ve B. Elmas, "İşletmelerin finansal başarısızlığının veri madenciliği ve diskriminant analizi modelleri ile tahmin edilmesi," Afyon Kocatepe Üniversitesi, İİBF Dergisi, cilt. XV, no. 1, s. 261-280, 2013.
- [15] D. Şengür ve A. Tekin, "Öğrencilerin mezuniyet notlarının veri madenciliği metotları ile tahmini," Bilişim Teknolojileri Dergisi, cilt. 6, no. 3, s. 7-16, 2013.
- [16] Ö.Ç. Gülen ve S. Özdemir, "Veri madenciliği teknikleri ile üstün yetenekli öğrencilerin ilgi alanlarının analizi," Üstün Yetenekli Eğitimi Araştırmaları Dergisi, cilt. 1, no. 1, s. 213-226, 2013.
- [17] Ö. Bilen, D. Hotaman, Ö.E. Aşkın ve A.H. Büyüklü, "LYS başarılarına göre okul performanslarının eğitsel veri madenciliği teknikleriyle incelenmesi: 2011 İstanbul örneği," Eğitim ve Bilim, cilt. 39, no. 172, s. 78-94, 2014.
- [18] C. Özel ve A. Topsakal, "Veri madenciliği kullanılarak beton basınç dayanımının belirlenmesi," Cumhuriyet Üniversitesi Fen Fakültesi Fen Bilimleri Dergisi (CFD), cilt. 35, no. 1, s. 43-57, 2014.
- [19] T. Değirmenci, "Resmi istatistiklerde veri madenciliği yaklaşımı," Yüksek Lisans Tezi, Kayseri, 2014.
- [20] A. Babaoğlu, "Veri madenciliği yöntemleri ve bir uygulama," Yüksek Lisans Tezi, Konya, 2015.
- [21] F. Turna, "Yeni nesil metasezgisel yaklaşımların kümeleme performanslarının karşılaştırılması: Resmi istatistiklerde uygulama," Doktora Tezi, Kayseri, 2016.
- [22] D. Tuncer, "Veri madenciliği yöntemi ile aylık kullanılabilir gelir tahmini ve göstergeleri," Yüksek Lisans Tezi, Edirne, 2018.
- [23] M.B. Kılıçalan, "Hanehalkı işgücü araştırma verileri ile veri madenciliği yöntemlerinin uygulanması ve modellerin karşılaştırılması," Yüksek Lisans Tezi, Ankara, 2018.
- [24] G. Kumru, "An empirical class analysis based on indebtedness, expenditure and living conditions," Yüksek Lisans Tezi, İstanbul, 2019.
- [25] M. Coşkun, "Hanehalkı bilişim teknolojileri kullanımının veri madenciliği teknikleri ile analizi," Yüksek Lisans Tezi, Ankara, 2019.
- [26] H.İ. Ekmekci, "Türkiye işgücü yapısının özelliklerinin incelenmesi: lojistik regresyon ve karar ağacı uygulaması," Yüksek Lisans Tezi, Ankara, 2019.

- [27] H. Hassani, S. Gheitanchi ve M.R. Yeganegi, "On the application of data mining to official data," *Journal of Data Science*; vol. 8, no. 1, pp. 75-89, 2010.
- [28] H. Hassani, G. Saporta ve E.S. Silva, "Data mining and official statistics: The past, the present and the future," *Big Data*, vol. 2, no. 1, pp. 34-43, 2014.
- [29] G. D'Angiolini, "Developing a metadata infrastructure for official data: the ISTAT experience," <http://www.di.uniba.it/~malerba/activities/mod02/pdf/dangiolini.pdf>.
- [30] B. Buelens, P. Daas ve J. Brakel, "Data mining for official statistics: challenges and opportunities," In: *The Proceedings of the 12th IEEE International Conference of Data Mining Workshops*, Brussels, Belgium, 2012, pp. 915.
- [31] J.H. Friedman, "Data mining and statistics: what's the connection?," Presented at the 29th Symposium on the Interface: Computing Science and Statistics, May 14-17, Houston, TX, 1997, pp. 3-9.
- [32] D.J. Hand, "Data mining: statistics and more?," *The American Statistician*, vol. 52, no. 2, pp. 112-118, 1998.
- [33] G. Saporta, "Data mining and official statistics," *Atti della Quinta Conferenza Nazionale di Statistica*, Rome, 2000, pp. 15-17.
- [34] S. Sumathi ve S.N. Sivanandam, "Introduction to data mining and its applications," New York: Springer, 2006.
- [35] E. Letouze, "Big data for Development," *Challenges & Opportunities*, New York: UN Global Pulse, 2012.
- [36] S. Nordbotten, "Neural network imputation applied to the Norwegian 1990 population census data," *Journal of Official Statistics*, vol. 12, no. 4, pp. 385-401, 1996.
- [37] S.B. Cohen, R. DiGaetano ve H. Goksel, "Estimation procedures in the 1996 medical expenditure panel survey household component," *MEPS Methodology Report No. 5*, AHCPR Publication No. 99-0027. Rockville, MD: Agency for Health Care Policy and Research, 1999.
- [38] W. Klosgen ve M. May, "Census data mining-an application," In: *Proceedings of the 5th European Conference on Principles of Data Mining*, Germany. Berlin: Springer, 2002, pp. 65-79.
- [39] C. Drummond, S. Matwin ve C. Gaffield, "Inferring and revising theories with confidence: data mining the 1901 Canadian census", In: *Proceedings of the ECML/PKDD'02 Workshop on Mining Official Data (MOD'02)*, 2002, August 19-23, Helsinki, Finland. Italy: University of Bari.
- [40] C. Soares, P. Brazdil ve C. Pinto, "Machine learning and statistics to detect errors in forms: competition or cooperation?," In: *Proceedings*

- of the ECML/PKDD'02 Workshop on Mining Official Data, Helsinki, Finland. Italy: University of Bari, 2002.
- [41] D. Malerba, F.A. Lisi ve F. Sblendorio, "Mining spatial association rules in census data: a relational approach," In: Proceedings of the ECML/PKDD'02 Workshop on Mining Official Data, Helsinki, Finland. Italy: University of Bari, 2002, pp. 1-14.
 - [42] G. Paaß ve J. Kindermann, "Bayesian regression mixtures of experts for geo-referenced data," *Intelligent Data Analysis*; vol. 7, no. 6, pp. 567-582, 2013.
 - [43] S. Frutos, E. Menasalva, C. Montes ve J. Segovia, "Calculating economic indexes per household and censal section from official Spanish databases," *Intelligent Data Analysis*, vol. 7, no. 6, pp. 603-613, 2003.
 - [44] A. Appice, M. Ceci, A. Lanza, F.A. Lisi ve D. Malerba, "Discovery of spatial association rules in geo-referenced census data: a relational mining approach," *Intelligent Data Analysis*, vol. 7, no. 6, pp. 541-566, 2003.
 - [45] P. Brito ve D. Malerba, "Mining official data," *Intelligent Data Analysis*, vol. 7, no. 6, pp. 497-500, 2003.
 - [46] I. Tudor ve M. Cărbureanu, "Data mining techniques in knowledge management in academic environment," *Buletinul*, vol. LIX, no. 1, pp. 29-34, 2007.
 - [47] D. Miller, J.S. McCarthy ve A.A. Zakzeski, "A fresh approach to agricultural statistics: data mining and remote sensing," In: Proceedings of the Joint Statistical Meetings, Washington, DC: American Statistical Association, 2009, pp. 3144-3155.
 - [48] W. Cecere, "2007 Census of agriculture non-response methodology" In: 2009 Joint Statistical Meetings, August 1-6, Washington, DC: American Statistical Association, 2009, pp. 2762-2769.
 - [49] M. Earp, S. Cox, J. McDaniel ve C. Crouse, "Exploring quarterly agricultural survey questionnaire version reduction scenarios", In: RRD Research Report, Washington, DC, pp. 1-36, 2009.
 - [50] S.C. Garber, "Census mail list trimming using SAS data mining," In: RRD Research Report, May 9, Washington, DC., 2009.
 - [51] J.S. McCarthy ve M.S. Earp, "Who makes mistakes? Using data mining techniques to analyze reporting errors in total acres operated," In: RRD Research Report. Washington, DC, 2009, pp. 1-20.
 - [52] J.S. McCarthy, T. Jacob ve D. Atkinson, "Innovative uses of data mining techniques in the production of official statistics," In: UN Statistical Commission Session on Innovations in Official Statistics, New York: UN Statistical Commission, 2009.
 - [53] M.C. Wheldon, A.E. Raftery, S.J. Clark ve P. Gerland, "Estimating demographic parameters with uncertainty from fragmentary data,"

Center for Statistics and the Social Sciences, University of Washington, Seattle, Washington, 2011, Working Paper 108.

- [54] M. Klucik, "Introducing new tool for official statistics: genetic programming," In: Eurostat International Conference (NTTS 2011), Brussels, Belgium. Belgium: Eurostat, 2011, pp. 22-24.
- [55] A. Gilar, "Recursive partitioning for racial classification cells," In: Proceedings of the Survey Research Methods Section, American Statistical Association-Session 628: Survey Analysis and Issues with Data Quality", Miami Beach, Washington, DC: American Statistical Association, 2011, pp. 2706-2720.
- [56] A. Nithya ve V. Sundaram, "Wheat disease identification using classification rules", International Journal of Scientific & Engineering Research, vol. 2, no. 9, pp. 244-248, 2011.
- [57] N. Nenonen, "Analysing factors related to slipping, stumbling, and falling accidents at work: Application of data mining methods to Finnish occupational accidents and diseases statistics database," Applied Ergonomics, vol. 44, no. 2, pp. 215-224, 2013.
- [58] A. Agrawal, P.D., Deshpande, A. Cecen, G.P. Basavarsu, A.N., Choudhary ve S.K. Kalidindi, "Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters," Integrating Materials and Manufacturing Innovation, vol. 3, no. 8, pp. 90-108, <2014.
- [59] C. Hsiao, "Analysis of Panel Data," Cambridge University Press: New York, 2003.
- [60] B.H. Baltagi, "Econometric analysis of panel data," John Wiley & Sons: England, 2005.
- [61] M.R. Segal, "Tree-structured methods for longitudinal data," Journal of the American Statistical Association, 87, 407-418, 1992.
- [62] L. Breiman, J.H. Friedman, R. Olshen ve A.C.G. Stone, "Classification and regression trees," Wadsworth International Group, Belmont, California, USA, 1984.
- [63] H.P. Zhang, "Multivariate adaptive splines for analysis of longitudinal data," Journal of Computational and Graphical Statistics, vol. 6, no. 1, pp. 74-91, 1997.
- [64] H.P. Zhang, "Classification tree for multiple binary responses," Journal of the American Statistical Association, vol. 93, no. 441, pp. 180-193, 1998.
- [65] Abdoell M., LeBlanc M., Stephens D., ve Harrison R.V. (2002), "Binary partitioning for continuous longitudinal data: categorizing a prognostic variable," Statistics in Medicine, vol. 21, no. 22, pp. 3395-3409, 2002.

- [66] G. De'Ath, "Multivariate regression trees: a new technique for modeling species-environment relationships," *Ecology*, vol. 83, no. 4, pp. 1105-1117, 2002.
- [67] G. Galimberti ve A. Montanari, "Regression trees for longitudinal data with time-dependent covariates," In K. Jajuga, A. Sokolowski, & H.-H. Bock (Eds.), *Classification, clustering and data analysis*, pp. 391-398, New York: Springer, 2002.
- [68] D.R. Larsen ve P.L. Speckman, "Multivariate regression trees for analysis of abundance data," *Biometrics*, vol. 60, no. 2, pp. 543-549, 2004.
- [69] G. Ritschard ve M. Oris, "Life course data in demography and social sciences: statistical and data mining approaches," In R. Levy, P. Ghisletta, J.-M. Le Goff, D. Spini, & E. Widmer (Eds.), *Towards an interdisciplinary perspective on the life course, advances in life course research*, pp. 289-320, Amsterdam: Elsevier, 2005.
- [70] S.K. Lee, "On generalized multivariate decision tree by using GEE," *Computational Statistics & Data Analysis*, vol. 49, no. 4, pp. 1105-1119, 2005.
- [71] S.K. Lee, "On classification and regression trees for multiple responses and its application," *Journal of Classification*, vol. 23, pp. 123-141, 2006.
- [72] S.K. Lee, H.C. Kang, S.T. Han ve K.H. Kim, "Using generalized estimating equations to learn decision trees with multivariate responses," *Data Mining and Knowledge Discovery*, vol. 11, pp. 273-293, 2005.
- [73] W.C. Hsiao ve Y.S. Shih, "Splitting variable selection for multivariate regression trees," *Statistics & Probability Letters*, vol. 77, no. 3, pp. 265-271, 2007.
- [74] G. Ritschard, A. Gabadinho, N.S. Müller ve M. Studer, "Mining event histories: a social science perspective," *International Journal of Data Mining, Modelling and Management*, vol. 1, no. 1, pp. 68-90, 2008.
- [75] R.J. Sela ve J.S. Simonoff, "RE-EM trees: A data mining approach for longitudinal and clustered data," *Machine Learning*, vol. 86, pp. 169-207, 2012.
- [76] W.Y. Loh ve W. Zheng, "Regression trees for longitudinal and multiresponse data," *The Analysis of Applied Statistics*, vol. 7, no. 1, pp. 495-522, 2013.
- [77] S.H. Eo ve H.J. Cho, "Tree-structured mixed-effects regression modeling for longitudinal data," *Journal of Computational and Graphical Statistics*, vol. 23, no. 3, pp. 740-760, 2013.
- [78] M.G. Kundu ve J. Harezlak, "Regression trees for longitudinal data," *arXiv:1309.7733 [stat.ME]*, 2014.
- [79] D. Liu, "Tree-based models for longitudinal data," Master's thesis, Retrieved from

https://etd.ohiolink.edu/!etd.send_file?accession=bgsu1399972118&disposition=inline, 2014.

- [80] W.Y. Loh, "Fifty years of classification and regression trees," *International Statistical Review*, vol. 82, no. 3, pp. 329-348, 2014.
- [81] S.D. Şeker, "Türkiyenin yoksulluk dinamikleri," *Doktora Tezi, Ortadoğu Teknik Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara*, 2011.
- [82] C. Okman, "Türkiye'de istatistiki Bölge-Düzey I ölçeğinde yoksulluğun ayrıştırılması," *Yüksek Lisans Tezi, İstanbul*, 2011.
- [83] H. Kırer, "Türkiye ve Almanya'nın kişisel gelir dağılımına ekonofizik yaklaşım," *Doktora Tezi, Yıldız Teknik Üniversitesi, Sosyal Bilimler Enstitüsü*, 2011.
- [84] E. Doğan, "Türkiye'de yoksulluğun ölçülmesi," *Uzmanlık Tezi, Kalkınma Bakanlığı*, 2014.
- [85] A. Barrett ve Y. McCarthy, "The earnings of immigrants in Ireland: Results from the 2005 EU survey of income and living conditions," *IZA Discussion Paper No. 2990. Available at SSRN: <https://ssrn.com/abstract=1012371>, (11, 2019), 2007.*
- [86] L. Stejskal and J. Stávková, "Living conditions of Czech farmers according to the EU statistics on income," *Agricultural Economics: Zemědělská ekonomika*, vol. 56, no. 7, pp. 310-316, 2010.
- [87] H. Řezanková, "Cluster analysis and categorical data," *Statistika*, vol. 3, pp. 216-232, 2009.
- [88] H. Řezanková, "Cluster analysis of economic data," *Statistika: Statistics and Economy Journal*, vol. 94, no. 1, pp. 73-86, 2014.
- [89] Z. Šulc and H. Řezanková, "Evaluation of recent similarity measures for categorical data," in *Proc. 17th AMSE*, pp. 249-258, 2014.
- [90] N. Birčiaková, I. Antořová, and F. Balák. "Determinants of Czech disposable household income and related housing quality," *Acta Universitatis*, vol. 65, no. 2, pp. 601-610, 2017.
- [91] P. Brunori, P. Hufe, and D.G. Mahler, "The roots of inequality: Estimating inequality of opportunity from regression trees forests," *Working Paper*, http://www.unicaldine.it/research/BHM_2017.pdf, (Oct. 22, 2019), 2019.
- [92] P.N. Tan, M. Steinbach ve V. Kumar, "Introduction to data mining," 1st ed., *Pearson International Edition, USA*, 2006.
- [93] U. Akküçük, "Veri Madenciliği: Kümeleme ve Sınıflama Algoritmaları," *İstanbul, Yalın Yayıncılık*, 2011.
- [94] J.R. Landis and G.G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159-174, 1977.

- [95] H. Akaike, "Information theory as an extension of the maximum likelihood principle," In: Petrov, B.N., Csaki, F. (Eds.), Second International Symposium on Information Theory. Akademiai Kiado, Budapest, 1973, pp. 267-281.
- [96] H.G. Schwarz, "Estimating the Dimension of a Model," The Annals of Statistics, vol. 6, no. 2, pp. 461-464, 1978.
- [97] C.E. Shannon, "A Mathematical Theory of Communication," The Bell System Technical Journal, vol. 27, pp. 379-423, 1948.
- [98] TÜİK, "Gelir ve Yaşam Koşulları Araştırması Metodolojisi Hakkında Genel Açıklama," http://www.tuik.gov.tr/HbGetir.do?id=21584&tb_id=10, 2016.
- [99] A. Alfons ve M. Templ, "Estimation of social exclusion indicators from complex surveys: The R package laeken," Journal of Statistical Software, vol. 54, no. 15, pp. 1-25, 2013.
- [100] P.F. Lazarsfeld ve N.W. Henry, "Latent structure analysis," Boston: Houghton Mifflin, 1968.
- [101] L.A. Goodman, "Exploratory latent structure analysis using both identifiable and unidentifiable models," Biometrika, vol. 61, no. 2, pp. 215-231, 1974.
- [102] E. Yıldırım, "Türkiye'nin Irak Savaşı'ndaki pozisyonuna dair görüşler: Savaş öncesi ve sonrası için karşılaştırmalı bir analiz," Balkan ve Yakın Doğu Sosyal Bilimler Dergisi, cilt. 1, no. 1, s. 51-60. 2015.
- [103] S.J. Haberman, "Log-linear models for frequency tables derived by indirect observation: Maximum likelihood equations," The Annals of Statistics, vol. 2, no. 5, pp. 911-924, 1974.
- [104] J.Z. Huang, "A fast clustering algorithm to cluster very large categorical data sets in data mining," In R. Ng (Ed). 1997 SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 1997, pp. 1-8, Tech. Rep 97-07). Vancouver, B.C., Canada: The University of British Columbia, Department of Computer Science.
- [105] Z.J. Huang, "Clustering categorical data with k-modes," In Encyclopedia of Data Warehousing and Mining, pp. 246-250, 2009.
- [106] L. Breiman. "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
- [107] S. Dua ve X. Du, "Data mining and machine learning in cyber security," CRC Press, 2011.
- [108] C. Akarsu, "Twitter verileri ile Türk televizyonları izlenme oranı sıralamaları tahmini," Yüksek Lisans Tezi, İstanbul, 2016.
- [109] B. Cerit, "Görüntü çözünürlüğünün otomatik cinsiyet ve yaş sınıflandırmasındaki etkisinin analizi," Yüksek Lisans Tezi, Ankara, 2016.

- [110] T. Hastie, R. Tibshirani, ve J. Friedman, “The elements of statistical learning,” Data Mining, Inference, and Prediction (2 ed.). New York, NY: Springer, 2009.
- [111] C. Doğan ve M. Tek, “Türkiye’de gelir dağılımının toplanma oranı yöntemiyle analizi,” Ekonomik ve Sosyal Araştırmalar Dergisi, vol. 3, no. 2, pp. 93-119, 2007.
- [112] TÜİK, “Gelir ve Yaşam Koşulları Araştırması Mikro Veri Seti (Kesit), 2015, http://www.tuik.gov.tr/MicroVeri/GYKA_2015/turkce/kilavuz/veri-seti-yapii/sii/index.html, 2016.



Bu bölümde GYKA anketi ile ilgili tanım, kavram ve genel açıklamalar ile temin edilen mikro veri setlerinde yer alan fert, fert kayıt ve hane veri setlerinin yapılarına ilişkin kılavuzlar sunulmuştur. GYKA hakkında genel açıklama bölümü için [98] nolu kaynak kullanılmış, veri seti kılavuzlarının yer aldığı bölümler için ise GYKA 2015 yılı mikro veri seti (kesit) dökümanından faydalanılmıştır ([112]).

A.1 GYKA Hakkında Genel Açıklama

Coğrafi kapsam: Türkiye Cumhuriyeti sınırları içinde bulunan tüm yerleşim yerleri kapsama dahil edilmiştir.

Kapsanan kitle: Türkiye Cumhuriyeti sınırları içindeki hanelerde yaşayan tüm hanehalkı fertleri kapsama alınmıştır. Ancak çalışmada kurumsal nüfus olarak tanımlanan üniversite yurtları, misafirhane, çocuk yuvası, yetiştirme yurdu, huzurevi, özel nitelikli hastane, hapishane, kışla ve ordu evlerinde yaşayanlar kapsam dışı tutulmuştur.

Kapsanan konular: Araştırmada gelir, yoksulluk, sosyal dışlanma ve diğer yaşam koşullarına ilişkin göstergeleri hesaplamak üzere aşağıdaki kategorilerde bilgi derlenmektedir:

- Konut
- Ekonomik durum
- Sosyal dışlanma
- Gayrimenkul sahipliği
- Eğitim
- Demografi
- Sağlık durumu
- İşgücü durumu

- Gelir durumu

ÖRNEKLEME TASARIMI:

Örneklem yöntemi: Tabakalı, iki aşamalı, küme örneklemesidir.

Örneklem birimi: Nihai örneklem birimi olarak **hanehalkı** tanımlanmıştır.

Tahmin boyutu: GYKA'nın yıllık sonuçlarından 2014 yılına kadar Türkiye, Kent, Kır ve İBBS 1. Düzey, panel araştırma sonuçlarından ise Türkiye genelinde gelir dağılımı, yoksulluk ve yaşam koşullarına ilişkin tahminler üretilmiştir. 2014 yılından itibaren yıllık kesit araştırma sonuçlarından Türkiye, İBBS 1. Düzey ve İBBS 2. Düzey bazında tahminler üretilmektedir.

Örneklem büyüklüğü ve cevapsızlık oranı: Yıllık örneklem büyüklüğü, araştırmanın tahmin boyutu, amaçları ve hedeflenen değişkenleri itibariyle örnekteki yıpranmaları da dikkate alarak başlangıçta yaklaşık 12800 hanehalkı civarında belirlenmiştir. Ancak, 2014 yılında İstatistiki Bölge Birimleri Sınıflaması 2. Düzey bazında tahmin üretilmesi hedeflenmiş ve 2011 yılından itibaren örnek hacmi örneğe yeni eklenen her alt örnek ile birlikte arttırılmaya başlanmıştır. Dolayısı ile 2015 yılı Gelir ve Yaşam Koşulları Araştırması'nda toplam 24461 hanehalkı ziyaret edilmiştir. Bu hanehalklarından; 23676'sı ile anket yapılmış, geri kalan 785 hanehalkı ile çeşitli nedenlerle anket yapılamamış ve cevaplamama formu düzenlenmiştir. Buna göre, 2015 yılında cevaplamama oranı Türkiye genelinde %3,2 olarak gerçekleşmiştir. Araştırmada, örnek hanehalklarında yaşayan 15 ve daha yukarı yaştaki fert sayısı toplam 59840 kişi olup, bu fertlerden 178'i ile çeşitli nedenlerle görüşülememiş veya ankete cevap vermemiştir. Dolayısı ile araştırmada fert cevapsızlık oranı %0,3'tür.

Nüfus ağırlıkları: Anket verilerinden kitle tahminlerinin üretilmesinde kullanılan ağırlık katsayıları, Adrese Dayalı Nüfus Kayıt Sistemi'ne göre revize edilen 2015 nüfus projeksiyonları esas alınarak hesaplanmıştır.

Panel anket uygulaması: Rotasyonel tasarımın kullanıldığı araştırmada, metodoloji gereği bir kısım hanehalkı bir yıldan diğer yıla örnekte kalırken, diğer

kısım örnekten çıkmakta ve yerine yeni örnek hanhalkları girmektedir. Bu araştırmada, yıllık örnek hacminin %75'inin panel kapsamda örnekte bırakılması öngörölmüştür. Bu doğrultuda panel uygulama, hedef kitleyi temsil edebilecek temel örnek seçimi ile başlamaktadır. Seçilen bu temel örnek hanhalkındaki 13 ve daha yukarı yaştaki tüm fertler belirlenmiş izleme kuralları doğrultusunda 4 yıl takip edilmektedir. Takip kurallarının amacı, hedef nüfusta olabilecek değışiklikleri temel örnekle yansıtmak ve fertleri belirli bir zaman takip etmektir.

REFERANS DÖNEMLERİ:

Referans dönemi; bilginin ait olduđu zaman süreci olarak tanımlanmakta olup, bu araştırmada farklı bilgiler için farklı referans dönemlerinin kullanımı söz konusudur.

- Gelir bilgilerinin referans dönemi “bir önceki takvim yılı”dır. Yani, 2015 uygulamasındaki gelir bilgileri 2014 yılına aittir.
- İstihdam bilgileri anketin uygulandıđı tarihten bir önceki hafta ile anketin yapıldıđı tarih dikkate alınarak sorgulanmıştır.
- Yaşam koşulları göstergeleri anketin yapıldıđı andaki durum itibariyle sorgulanmıştır.

VERİ DERLEME YÖNTEMİ:

Gelir ve Yaşam Koşulları Araştırması her yıl düzenli olarak uygulanmaktadır. Veri derleme çalışması Mart–Temmuz ayları arasında gerçekleştirilmektedir. Alan uygulaması TÜİK bölge müdürlükleri tarafından TÜİK tarafından geliştirilmiş olan HARZEMLİ veri giriş programı ile gerçekleştirilmekte olup çalışmada anketör, kontrolör, organizatör ve teknik yetkili olarak adlandırılan elemanlar görev almaktadır. Alan çalışması süresince alan organizasyon birimleri ile merkez organizasyon birimleri koordineli olarak çalışmakta ve araştırmanın veri derleme ve veri analiz aşamalarının başarıyla tamamlanması sağlanmaktadır.

Veri derleme çalışması öncesinde hanehalklarına anketi tanıtan, amacını anlatan, uygulamanın nasıl ve ne zaman yapılacağına ilişkin bilgi veren mektuplar gönderilmektedir.

Veri derleme çalışmasını takiben fertlerin anket çalışmasından elde edilen T.C. kimlik numaraları kullanılarak kurumların idari kayıtlarından; ücret/müteşebbis geliri, emekli/dul-yetim aylığı, işsizlik ödenekleri, sakatlık/yaşlılık/gazilik vb. aylıkları, aldıkları sosyal yardımlar, gelir/varlık vergileri ile ödedikleri

sigorta primlerine ilişkin bilgiler talep edilmekte ve anket verileri ile kontrol edilmektedir. Anketten gelen eksik/hatalı bilgiler idari kayıtlar yardımıyla edit/kontrol edilmektedir.

TANIM VE KAVRAMLAR

Hanehalkı ile ilgili tanımlar

Hanehalkı: Aralarında akrabalık bağı bulunsun ya da bulunmasın aynı konutta veya aynı konutun bir bölümünde yaşayan, temel ihtiyaçlarını birlikte karşılayan, hanehalkı hizmet ve yönetimine iştirak eden bir veya birden fazla kişiden oluşan topluluktur.

Kurumsal nüfus: Günlük yaşam gereksinimleri, yasal bir düzenlemeye dayalı olarak kurulan özel ya da tüzel kurum/kuruluşlarda kısmen ya da tamamen karşılanan, bireysel karar ve davranışlarında yetkili otoritenin kurallarına kısmen ya da tamamen bağımlı olarak hareket eden, ancak bireysel harcamalarına

karar verebilen kişilerin yaşamlarını sürdürdükleri alanlarda yaşayan nüfusa kurumsal nüfus denir. Bu tanıma göre; üniversite yurtları, misafirhane, çocuk yuvası, yetiştirme yurdu, huzurevi, özel nitelikteki hastane, hapishane, kışla ve ordu evlerinde yaşayanlar kurumsal nüfusa dahildir.

Hanehalkı ferdi: Hanehalkını meydana getiren topluluğun bir üyesidir. Daimi ikameti örnek hanehalkı adresi olan fertler, görüşme anında geçici olarak hanede bulunmasalar da hanehalkı ferdi olarak kabul edilirler. Bununla birlikte, kurumsal nüfus kapsamında yer alan fertler (zorunlu askerlik görevini yapmakta

olan er ve erbařlar, hapiste olanlar, huzurevinde kalan yařlılar ve yurtta kalan öđrenciler vb.) hanehalkı ferdi olarak kabul edilmezler.

Daimi ikamet; 12 ay ve daha uzun süredir oturulan veya oturulma niyetinde olunan adrestir.

Hanehalkı sorumlusu: Hanehalkının sosyo-ekonomik durumu ve hanede yařayan tüm fertlerin kiřisel özellikleri hakkında en dođru bilgiye sahip, hanenin yönetim veya geçiminden sorumlu yetiřkin hanehalkı üyesidir.

Gelir ile ilgili tanımlar

Gelir: Bir milli ekonomide belirli bir dönemde genellikle bir yılda yeniden yaratılan mal ve hizmet deđerlerinin toplamıdır. Buna milli gelir veya toplam hasıla denir. Hanehalkı ya da fert açısından gelir, üretime yapılan katkı karřılıđında belirli bir sürede elde edilen deđerlerin toplamı olarak tanımlanabilir. Hanehalkı geliri, emek karřılıđı alınan ücret ve maařlar řeklinde, toprak kirası (rant), sermaye geliri (faiz) veya teřebbüs geliri (kâr) řeklinde olabilir.

Faaliyet gelirleri:

- **Maař, ücret, yevmiye gelirleri:** Gelir referans dönemi içerisinde ücret karřılıđı bir iřte çalıřma sonucunda elde edilen nakdi veya ayni gelirlerdir. Gelir referans dönemi içerisinde alınan ikramiye, iř riski, iř güçlüğü teminindeki günlük zammı, fazla mesai, prim ve temettüleri ve ek görevlerden elde edilen gelirler maař, ücret veya yevmiye gelirine dahil edilmiřtir.
- **Müteřebbis gelirleri:** Gelir referans dönemi içerisinde müteřebbis olarak çalıřan fertlerin (iřteki durumu kendi hesabına veya iřveren olarak çalıřanlar) faaliyetleri sonucu elde ettiđi gayri safi hasılatından, gelir referans dönemi içerisinde yaptıđı tüm giderler (iřyeri masrafları, vergiler, sosyal güvenlik kesintileri vb.) düşüldükten sonra ortaklık durumu varsa kiřinin payına düşen gelir müteřebbis geliridir. yoluyla hesaplanmaktadır.

Faaliyet dışı gelirleri:

- **Gayrimenkul geliri:** Apartman dairesi, dükkân, depo vb. gayrimenkuller ile tarla, bağ, bahçe, arsa gibi mülklerin referans dönemi içerisinde kiraya verilmesi sonucu elde edilen hasılatın, bu hasılatın sağlanması için yapılan tüm giderler (sigorta, vergi, tamir ve bakım giderleri vb.) düşülerek elde edilen gelirdir.
- **Menkul kıymet geliri:** Banka hesabından elde edilen faiz, finans kurumlarından veya sermaye şirketlerinden elde edilen temettü veya kar payı vb. gelirlerdir. Menkul kıymet gelirlerinde, vergiler gelir elde edilmeden kaynaktan kesilmektedir. Ancak, menkul kıymet geliri elde eden hanehalkı fertlerinin, bu geliri elde etmek için, menkul kıymetlerin korunması ile ilgili sigorta, kiralık kasa, tahvil giderleri, vergi vb. giderler düşülerek, elde ettikleri net gelirler menkul kıymet geliri olarak kayıt edilmiştir.
- **Karşılıksız (transfer) gelirler:** Gelir referans dönemi içerisinde devlet, özel kurum/kuruluş veya başka kişi/hanehalklarından alınan karşılıksız transferlerdir. İki gruba ayrılmaktadır;
 - a. **Sosyal transferler;** Hanelerin bazı ihtiyaçlarını karşılamak veya parasal sıkıntılarından dolayı bazı risklerle karşı karşıya kalmaları durumunda gelir referans döneminde devletten veya çeşitli kurum/kuruluşlardan aldıkları ayni veya nakdi karşılıksız yardımlar olarak nitelendirilmektedir (Aile-çocuk yardımı, konut yardımı (kira vb.), yoksul hanehalklarına yapılan diğer sosyal yardımlar ile 15 ve daha yukarı yaştaki fertlerin elde ettiği emeklilik, yaşlılık, işsizlik, özürlülük, gazilik, malüllük, karşılıksız burslar vb. düzenli olarak alınan ayni veya nakdi yardımlar gibi).
 - b. **Hanelerarası transferler;** Hanelerin başka kişi veya hanehalklarından düzenli olarak aldıkları karşılıksız ayni ya da nakdi yardımlardır.

İzafi kira: Konuta mülkiyet şekli “ev sahibi”, “lojman” veya “diğer (babasının, akrabasının vb. evinde ikamet edip hiçbir şekilde ücret ödemedi oturanlar)” şeklinde olan hanehalklarının ikamet ettiğı konutun kira değeri izafi kira olarak değerlendirilmektedir. Ayrıca, kiracı olup piyasa değeri altında kira ödeyenlerin ödedikleri kira ile piyasa değeri arasındaki fark da hanenin izafi kira geliri olarak değerlendirilmektedir. İzafi kira değeri iki şekilde hesaplanmaktadır;

- Hanehalklarından anket yoluyla, benzer özelliklere sahip olan aynı semtte ve yerleşim yerinde kira ile oturan konutun kira değeri emsal alınarak saptanmak üzere beyan yoluyla alınmaktadır.
- Regresyon modeli kurularak, açıklayıcı değişkenlerin yardımıyla tahmin edilmektedir.

Hanehalkı kullanılabilir geliri hesaplanırken ikinci hesaplama kullanılmaktadır.

Nakdi Gelir: Nakit; “para” özellikle “kullanılmaya hazır para” anlamına gelmektedir. Faaliyet dışı faktörler veya faaliyet karşılığında gelirlerin nakit (para) olarak elde edilmesidir.

Aynı Gelir: Aynı, mal ile ödenen anlamına gelmektedir. Aynı gelir, emek veya sermaye karşılığında gelirin mal olarak elde edilmesidir. Burada en önemli husus, aynı gelirin hanede tüketilmesi gerekliliğidir. Aynı geliri oluşturan mal satılıp nakde çevriliyorsa bu aynı gelir olarak değil, nakdi gelir olarak değerlendirilmektedir.

Brüt gelir: Emeklilik ve sosyal sigortalar keseneğı ile vergiler düşülmeden önceki gelirdir.

Net gelir: Kişilerin ödediğı emeklilik ve sosyal sigortalar keseneğı ile vergiler hariç, eline geçen gelir net gelirdir.

Hanehalkı kullanılabilir net geliri: Hanehalkındaki her bir ferdin elde ettiğı kişisel yıllık kullanılabilir gelirlerin toplamı (maaş-ücret, yevmiye, müteşebbis geliri ile emekli maaşı, dul-yetim aylıkları ve yaşlılara yapılan ödemeler, karşılıksız burs vb. faaliyet ve faaliyet dışı aynı veya nakdi gelirlerin toplamı) ile hane

bazında elde edilen yıllık gelirlerin (gayrimenkul kira geliri, haneye yapılan karşılıksız yardımlar, 15 yaşın altındaki fertlerin elde ettiği gelirler vb.) toplamından, gelir referans döneminde ödenen vergiler ve diğer

hane veya kişilere yapılan düzenli transferler düşüldükten sonra hanehalkı kullanılabilir gelirine ulaşılmaktadır ([98]).



A.2 GYKA Fert Kesit Veri Seti Yapısı

Tablo A.1 GYKA fert kesit veri seti değişkenleri ([112])

Değişken Adı	Açıklama	Kayıt uzunluğu	Seçenekler
BULTEN	Her hane için verilen numara (Bülten numarası)	5	1...24461
FERT_ID	Fert kimlik numarası (Bülten no + Fert sıra no)	7	
FB010	Anket yılı	4	2015
FB030	Fert için ağırlık katsayısı	4	Format 4.7
FB100	Ferdin medeni durumu	1	1- Evli
			2- Hiç evlenmedi
			3- Eşi öldü
			4- Boşandı
			5- Ayrı yaşıyor
FB110	Ferdin resmi nikahının olup olmadığı [FB100="1","5" olanlar]	1	1- Evet
			2- Hayır
Eğitim			
FE010	Ferdin bir eğitim kurumuna devam etme durumu	1	1- Evet
			2- Hayır
FE020	Ferdin devam ettiği eğitim düzeyi	1	1- İlkokul
	[FE010 = "1" olanlar]		2- Ortaokul veya mesleki ortaokul
			3- Genel lise

			4- Mesleki veya teknik lise
			5- Yüksek okul veya fakülte
			6- Yüksek Lisans/Doktora
FE030	Ferdin en son bitirdiği okul (En son diploma alınan okul)	1	0- Okur-yazar olmayan
			1- Okur yazar olup, bir okul bitirmeyen
			2- İlkokul
			3- Ortaokul, mesleki ortaokul ve ilköğretim
			4- Genel lise
			5- Mesleki veya teknik lise
			6- Yüksekokul, fakülte ve üzeri
FE040	Ferdin son eğitim düzeyini tamamladığı yıl [FE030 ="2", "3", "4", "5", "6" olanlar]	4	1900..2015
Sağlık Durumu			
FS010	Ferdin genel sağlık durumu	1	1- Çok iyi
			2- İyi
			3- Orta
			4- Kötü
			5- Çok kötü
FS020	Ferdin kronik/müzmin bir hastalığının olup olmadığı (Şeker, hipertansiyon, astım, böbrek yetmezliği, romatizmal hastalıklar, vb.)	1	1- Evet
			2- Hayır

FS030	En az 6 aydan beri süregelen fiziksel veya ruhsal herhangi bir sağlık problemi nedeniyle ferдин günlük faaliyetlerinde bir sınırlama olup olmadığı	1	1- Evet, çok sınırlandı
			2- Evet, sınırlandı
			3- Hayır, sınırlanmadı
FS050	Ferdin son 12 ay içinde, tıbbi muayene veya tedavi ihtiyacı olduğu halde doktora başvuramama durumu olup olmadığı	1	1- Evet
			2- Hayır
FS060	Ferdin doktora başvuramamasının asıl sebebi [FS 050 = "1" olanlar]	1	1-Ödeme gücüğü (çok pahalı olması ya da sigorta fonu tarafından karşılanmaması)
			2- İş, çocuk ya da bakmakla yükümlü olduğu diğer kişilerden dolayı zaman bulamama
			3- Sağlık kuruluşunun uzak olması / ulaşım imkanının olmaması
			4- Ameliyat / tedavi korkusu
			5- Randevu süresinin çok geç tarihte verilmesi
			6- Rahatsızlığın kendi kendine geçmesini beklemesi
			7- İyi bir doktor veya uzman tanımaması
			8- Diğer nedenler
FS070	Ferdin son 12 ay içinde, dış hekimine başvurma ihtiyacı olduğu halde başvuramama durumu olup olmadığı	1	1- Evet
			2- Hayır
FS080	Ferdin bir dış hekimine başvuramamasının asıl sebebi	1	1- Ödeme gücüğü (çok pahalı olması ya da sigorta fonu tarafından karşılanmaması)

	[FS 070 = "1" olanlar]		2- İş, çocuk ya da bakmakla yükümlü olduğu diğer kişilerden dolayı zaman bulamama
			3- Sağlık kuruluşunun uzak olması / ulaşım imkanının olmaması
			4- Ameliyat / tedavi korkusu
			5- Randevu süresinin çok geç tarihte verilmesi
			6- Rahatsızlığın kendi kendine geçmesini beklemesi
			7- İyi bir doktor veya uzman tanımaması
			8- Diğer nedenler
Maddi Yoksunluk			
FY020	Eskimiş giysilerinizin yerine yeni giysiler alırmısınız? (İkinci el olmayan)	1	1- Evet
			2- Hayır - maddi yetersizlik
			3- Hayır - diğer nedenler
FY030	Biri her türlü hava koşuluna uygun olmak üzere; düzgün iki çift ayakkabınız var mı?	1	1- Evet
			2- Hayır - maddi yetersizlik
			3- Hayır - diğer nedenler
FY050	Ayda en az bir kere arkadaşlarınız, aileniz/akrabalarınızla yemek yemek veya birşeyler içmek için evde veya dışarıda (lokanta, pastane, kafe vb. yerlerde) biraraya gelir misiniz?	1	1- Evet
			2- Hayır - maddi yetersizlik
			3- Hayır - diğer nedenler
FY060		1	1- Evet

	Spor, sinema, konser gibi boş zaman faaliyetlerine düzenli olarak katılır mısınız?		2- Hayır - maddi yetersizlik
			3- Hayır - diğer nedenler
FY070	Kimseye danışmak zorunda kalmadan her hafta küçük bir miktar parayı kendi arzunuza göre harcar mısınız?	1	1- Evet
			2- Hayır - maddi yetersizlik
			3- Hayır - diğer nedenler
FY080	Evde kişisel amaçlı kullanımınız için internet erişiminiz var mı? (Cep telefonu, wireless, video oyun konsolu, dizüstü/masaüstü bilgisayar, TV vb. aracılığıyla)	1	1- Evet
			2- Hayır - maddi yetersizlik
			3- Hayır - diğer nedenler
İstihdam Durumu			
FI010	Ferdin kendi tanımlamasına göre şu anki durumu	1	1- Ücretli veya yevmiyeli (Tam zamanlı çalışıyor)
	(*) Ücretsiz aile işçileri dahildir.		2- Ücretli veya yevmiyeli (Yarı zamanlı çalışıyor)
			3- İşveren/kendi hesabına çalışan (*) (Tam zamanlı çalışıyor)
			4- İşveren/kendi hesabına çalışan (*) (Yarı zamanlı çalışıyor)
			5- İş arıyor
			6- Çırak/stajyer, eğitim öğretime devam ediyor
			7- Emekli, erken emekli veya işten ayrılmış
			8- Yaşlı, engelli veya çalışamaz halde
			9- Ev işleri, çocuk, yaşlı, hasta vb. kişilerin bakımı ile meşgul

			10- Diğer faal olmayan
FI020	Ferdin geçen hafta içinde aynı (mal) veya nakdi (para) bir gelir elde etmek amacıyla bir saat bile olsa bir işte çalışıp çalışmadığı veya geçici olarak başında bulunmadığı bir işinin olup olmadığı (Maaşlı, ücretli, yevmiyeli, kendi hesabına, işveren ya da ücretsiz aile işçisi olarak)	1	1- Evet
			2- Hayır
FI030	Şu an çalışmayan ferdin daha önce aynı (mal) veya nakdi (para) bir gelir elde etmek amacıyla çalışıp çalışmadığı [FI010>"4" ve FI020 = "2" olanlar]	1	1- Evet
			2- Hayır
FI040	Ferdin son dört hafta içinde iş arayıp aramadığı [FI010>"4" ve FI020 = "2" olanlar]	1	1- Evet
			2- Hayır
FI050	İş imkanı olsa ferdin iki hafta içinde işbaşı yapabilme durumu [FI010>"4" ve FI020 = "2" olanlar]	1	1- Evet
			2- Hayır
FI060	Ferdin son dört hafta içinde, iş aramak için en son kullandığı kanal [FI010>"4" ve FI020 = "2" ve FI040 = "1" olanlar]	2	1- İŞKUR veya özel istihdam ofislerine başvurdu
			2- Doğrudan bir işverene
			3- Arkadaş, akraba veya diğer kişilere
			4- Gazete, dergi veya internetten iş ilanlarına başvurma
			5- Kendi işini kurmak için teşebbüste bulunma

			6- İş amaçlı yazılı ve sözlü mülakata katıldı
			7- İş başvurusunun sonucunu bekledi
			8- İşçi simsarlarına başvurdu
			9- Hiç bir şey yapmadı
			10- Diğer
Son İş Bilgileri			
FI070	Ferdin son işindeki işteki durumu (Anonim şirket ve kooperatif ortakları fiili olarak çalışıyorlarsa "ücretli, maaşlı" olarak sınıflandırılmıştır) [FI010>"4" ve FI020 = "2" ve FI030 = "1" olanlar]	1	1- Ücretli, maaşlı
			2- Yevmiyeli
			3- İşveren
			4- Kendi hesabına
			5- Ücretsiz aile işçisi
FI080	Ferdin son çalıştığı işindeki meslek kodu (ISCO-08) [FI010>4 ve FI020 = "2" ve FI030 = "1" olanlar]	1	1- Yöneticiler
			2- Profesyonel Meslek Mensupları
			3- Teknisyenler, Teknikerler ve Yardımcı Profesyonel Meslek Mensupları
			4- Büro Hizmetlerinde Çalışan Elemanlar
			5- Hizmet ve Satış Elemanları
			6- Nitelikli Tarım, Ormancılık ve Su Ürünlerinde Çalışanlar
			7- Sanatkarlar ve İlgili İşlerde Çalışanlar
			8- Tesis ve Makine Operatörleri ve Montajcılar

			9- Nitelik Gerektirmeyen İşlerde Çalışanlar
FI090	Ferdin son işinden ayrıldığı yıl [FI010>"4" ve FI020 = "2" ve FI030 = "1" olanlar]	4	1900...2015
FI100	Ferdin son işinde 2014 yılında çalıştığı ay sayısı [FI010>"4" ve FI020 = "2" ve FI030 = "1" olanlar ve FI090 >="2014" olanlar]	2	0...12
FI110	Ferdin son çalıştığı işinden 2014 yılında elde etmiş olduğu net geliri (TL) [FI010>"4" ve FI020 = "2" ve FI030 = "1" olanlar ve FI100 >"0" olanlar]	6	-999999...999999
Esas İş Bilgileri			
FI120	Ferdin esas işindeki işteki durumu (Anonim şirket ve kooperatif ortakları fiili olarak çalışıyorlarsa "ücretli, maaşlı" olarak sınıflandırılmıştır) [FI010<"5" veya FI020 = "1" olanlar]	1	1- Ücretli, maaşlı
			2- Yevmiyeli
			3- İşveren
			4- Kendi hesabına
			5- Ücretsiz aile işçisi
FI130	Ferdin esas işindeki meslek kodu (ISCO-08) [FI010<"5" veya FI020 = "1" olanlar]	1	1- Yöneticiler
			2- Profesyonel Meslek Mensupları
			3- Teknisyenler, Teknikerler ve Yardımcı Profesyonel Meslek Mensupları
			4- Büro Hizmetlerinde Çalışan Elemanlar

			5- Hizmet ve Satış Elemanları
			6- Nitelikli Tarım, Ormancılık ve Su Ürünlerinde Çalışanlar
			7- Sanatkarlar ve İlgili İşlerde Çalışanlar
			8- Tesis ve Makine Operatörleri ve Montajcılar
			9- Nitelik Gerektirmeyen İşlerde Çalışanlar
FI140	Ferdin esas işindeki çalıştığı yer, kuruluş veya işyerinin ana faaliyet kodu (NACE REV 2) [FI010<"5" veya FI020 = "1" olanlar]	2	01- Tarım, Ormancılık, Avcılık ve Balıkçılık
			02- Madencilik ve Taşocakçılığı
			03- İmalat
			04- Elektrik, Gaz, Buhar, Su Temini ve Kanalizasyon
			05- İnşaat
			06- Toptan ve Perakende Ticaret; Motorlu Kara Taşıtlarının ve Motosikletlerin onarımı
			07- Ulaştırma ve Depolama
			08- Konaklama ve Yiyecek Hizmeti Faaliyetleri
			09- Bilgi ve İletişim
			10- Finans ve Sigorta Faaliyetleri
			11- Gayrimenkul Faaliyetleri
			12- Mesleki, Bilimsel ve Teknik Faaliyetler
			13- İdari ve Destek Hizmet Faaliyetleri

			14- Kamu Yönetimi ve Savunma
			15- Eğitim
			16- İnsan Sağlığı ve Sosyal Hizmet Faaliyetleri
			17- Kültür, Sanat, Eğlence, Dinlenme ve Spor
			18- Diğer Hizmet Faaliyetleri
FI150	Ferdin esas işindeki haftalık normal çalışma saati (Ücretlilerde, düzenli ödenen ücretli fazla mesai saatleri de eklenmiştir)	2	1..99
FI160	Ferdin ek iş/işlerindeki haftalık normal çalışma saati toplamı [FI010<"5" veya FI020 = "1" olanlar ve ek işi olanlar]	2	1..99
FI170	Ferdin haftada 30 saatten az çalışmasının nedeni [FI150+FI160<30]	1	1- Ev işleri / çocuk veya bakıma muhtaç diğer kişilerin bakımı için 2- Eğitimine devam ettiği için 3- Kendi hastalığı ya da engellilik hali nedeniyle 4- Diğer ailevi ve kişisel nedenlerden dolayı 5- Tam zamanlı bir iş bulamadığı için 6- İşin niteliği gereği 7- Daha fazla çalışmak istemediği için 8- Çalıştığı süreyi tam zamanlı bir iş olarak değerlendirdiği için

			9- Diğer
FI180	Ferdin esas işindeki yer, kuruluş veya işyerinde çalışan kişi sayısı [FI010<"5" veya FI020 = "1" olanlar]	1	1- 10 ve daha az
			2- 11-19 kişi
			3- 20-49 kişi
			4- 50 veya daha fazla kişi
			5- Bilmiyor, fakat 11'den az kişi
			6- Bilmiyor, fakat 10'dan fazla kişi
FI190	Ferdin esas işinden dolayı sosyal güvenlik kuruluşuna kayıtlılık durumu [FI010<"5" veya FI020 = "1" olanlar]	1	1- Kayıtlı
			2- Kayıtlı değil
FI210	Ücretli veya yevmiyeli çalışan ferdin esas işinin süreklilik durumu [FI010<"5" veya (FI020 ="1" ve FI120="1" veya FI120="2") olanlar]	1	1- Sürekli iş
			2- Belirli süreli sözleşmeli iş (geçici)
			3- Sözleşmesiz arada sırada çalışan
			4- Öğrenci vb. kişilerin tatil zamanında yaptıkları geçici iş
FI220	Ücretli veya yevmiyeli çalışan ferdin esas işinde, personel denetimi veya yöneticilik yapma durumu [FI010<"5" veya (FI020 ="1" ve FI120="1" veya FI120="2") olanlar]	1	1- Evet
			2- Hayır
FI240	Ferdin esas işinde 2014 yılında çalıştığı ay sayısı [FI010<"5" veya FI020 = "1" olanlar ve işe "2015" yılından önce başlayanlar]	2	1..12

FI250	Ferdin esas işinden 2014 yılı içinde elde ettiği YILLIK net gelir (TL) [FI010<"5" veya FI020 = "1" olanlar ve işe "2014" yılından önce başlayanlar]	6	-999999...999999
Çalışma Tarihi			
FI260	Ferdin 2014 yılında tam-zamanlı çalışarak geçirdiği ay sayısı	2	0..12
FI270	Ferdin 2014 yılında yarı-zamanlı çalışarak geçirdiği ay sayısı	2	0..12
FI280	Ferdin 2014 yılında işsiz olarak geçirdiği ay sayısı	2	0..12
FI290	Ferdin iktisaden faal olmaksızın 2014 yılında emekli/işten ayrılmış olarak geçirdiği ay sayısı	2	0..12
FI300	Ferdin 2014 yılında iktisaden faal olmaksızın eğitim-öğretimde (çırak/stajyerlik dahil) geçirdiği ay sayısı	2	0..12
FI310	Ferdin 2014 yılında diğer iktisaden faal olmadan geçirdiği ay sayısı (Ev işleri ile meşgul, yaşlı/engelli/çalışamaz halde, askerlik görevi vb.)	2	0..12

FI320	Ferdin ilk düzenli işinde çalışmaya başladığı yaş (Öğrenci iken geçici part-time yapılan işler hariç, ücretsiz aile işçileri dahil) [FI010<"5" veya FI020 = "1" veya FI030 = "1" olanlar] ve FI210<>"4"	2	8...99
FI330	Ferdin ilk çalışmaya başladığı zamandan beri gelir getirici bir işte çalışarak geçirdiği süre (YIL) [FI010<"5" veya FI020 = "1" veya FI030 = "1" olanlar] ve FI210<>"4"	2	0...99
2015 Yılı Fert Gelirleri			
FG010	Ferdin 2014 yılı içinde elde ettiği YILLIK toplam net nakdi ücret, maaş, yevmiye geliri (TL)	9.2	0..999999
FG020	Ferdin 2014 yılı içinde elde ettiği YILLIK toplam net aynı ücret, maaş, yevmiye geliri (TL)	9.2	0..999999
FG030	Ferdin 2014 yılı içinde elde ettiği YILLIK toplam net nakdi müteşebbis geliri (TL)	9.2	-999999..999999
FG040	Ferdin 2014 yılı içinde elde ettiği YILLIK toplam net aynı müteşebbis geliri (TL)	9.2	0..999999
FG070	Ferdin 2014 yılı içinde elde ettiği işsizlik yardımlarının değeri (TL) (Kıdem tazminatı dahil)	9.2	0..999999

FG080	Ferdin 2014 yılı içinde elde ettiği emeklilik/yaşlılık/isteğe bağlı emeklilik gelirlerinin değeri (TL)	9.2	0..999999
FG085	Ferdin 2014 yılı içinde elde ettiği emeklilik ikramiyesi	9.2	0..999999
FG090	Ferdin 2014 yılı içinde elde ettiği dul ve yetim maaşlarının değeri (TL) (Ölüm tazminatı dahil)	9.2	0..999999
FG100	Ferdin 2014 yılı içinde ücretli hastalık izninden dolayı elde etmiş olduğu sosyal yardım niteliğindeki gelirin (Rapor parası) değeri (TL)	9.2	0..999999
FG110	Ferdin 2014 yılı içinde elde ettiği sakatlık, gazilik, malülen emeklilik gelirlerinin değeri (TL)	9.2	0..999999
FG120	Ferdin 2014 yılı içinde eğitim ile ilgili elde ettiği karşılıksız burs, bağış vb. gelirlerin değeri (TL)	9.2	0..999999
FG130	Ferdin 2014 yılı içinde ödediği isteğe bağlı emeklilik primi toplam değeri (TL)	9.2	0..999999
FG140	Ferdin 2014 yılı içinde elde ettiği gelirlerin toplam değeri (TL) (FG010 + FG020 + FG030 + FG040 + FG070 + FG080 + FG085 + FG090 + FG100 + FG110 + FG120)	9.2	-999999..999999

A.3 GYKA Fert Kayıt Kesit Veri Seti Yapısı

Tablo A.2 GYKA fert kayıt kesit veri seti değişkenleri ([112])

Değişken Adı	Açıklama	Kayıt uzunluğu	Seçenekler
BULTEN	Her hane için verilen numara (Bülten numarası)	5	1...24461
FERT_ID	Fert kimlik numarası (Bülten no + Fert sıra no)	7	
FK010	Anketin yapıldığı yıl	4	2015
FK050	Fert için ağırlık katsayısı	4.7	Format 4.7
FK070	Ferdin yaşı (2014 yılı Aralık ayındaki)	2	-1..114
FK080	Ferdin doğum yılı	4	1900...2015
FK090	Cinsiyet	1	1- Erkek 2- Kadın
FK095	Hanehalkı sorumlusu	2	1- Hanehalkı sorumlusu 2- Eşi 3- Oğlu / Kızı 4- Babası / Annesi 5- Kardeşi 6- Kayınpederi / Kayınvalidesi 7- Damadı / Gelini 8- Torunu 9- Diğer akrabalar 10- Evde kalan hizmetli 11- Akraba olmayanlar
FK210	Ferdin kendi tanımlamasına göre şu anki esas faaliyet durumu	1	1- Çalışıyor 2- İş arıyor 3- Emekli/Erken emekli veya işten ayrılmış 4- Diğer aktif olmayan
FK220	Ferdin babasının kimlik numarası (hanehalkı üyesi ise)	7	Baba kimlik no - (FERT_ID)

FK230	Ferdin annesinin kimlik numarası (hanehalkı üyesi ise)	7	Anne kimlik no - (FERT_ID)
FK240	Ferdin eşinin kimlik numarası (hanehalkı üyesi ise)	7	Eş kimlik no - (FERT_ID)



A.4 GYKA Hane Kesit Veri Seti Yapısı

Tablo A.3 GYKA hane kesit veri seti değişkenleri ([112])

Değişken Adı	Açıklama	Kayıt uzunluğu	Seçenekler
BULTEN	Her hane için verilen numara (Bülten numarası)	5	1...24461
HB010	Anket yılı	4	2015
HB030	İstatistiki Bölge Birimleri Sınıflaması, 1. Düzey	3	Kod-Coğrafi Bölge Adı TR1- İstanbul TR2- Batı Marmara TR3- Ege TR4- Doğu Marmara TR5- Batı Anadolu TR6- Akdeniz TR7- Orta Anadolu TR8- Batı Karadeniz TR9- Doğu Karadeniz TRA- Kuzey Doğu Anadolu TRB- Orta Doğu Anadolu TRC- Güney Doğu Anadolu
HB031	İstatistiki Bölge Birimleri Sınıflaması, 2. Düzey	3	Kod-Coğrafi Bölge Adı TR10- İstanbul TR21- Tekirdağ, Edirne, Kırklareli TR22- Balıkesir, Çanakkale TR31- İzmir TR32- Aydın, Denizli, Muğla TR33- Manisa, Afyon, Kütahya, Uşak TR41- Bursa, Eskişehir, Bilecik TR42- Kocaeli, Sakarya, Düzce, Bolu, Yalova TR51- Ankara TR52- Konya, Karaman TR61- Antalya, Isparta, Burdur TR62- Adana, Mersin TR63- Hatay, Kahramanmaraş, Osmaniye TR71- Kırıkkale, Aksaray, Niğde, Nevşehir, Kırşehir TR72- Kayseri, Sivas, Yozgat TR81- Zonguldak, Karabük, Bartın TR82- Kastamonu, Çankırı, Sinop TR83- Samsun, Tokat, Çorum, Amasya TR90- Trabzon, Ordu, Giresun, Rize, Artvin, Gümüşhane

			TRA1- Erzurum, Erzincan, Bayburt TRA2- Ağrı, Kars, Iğdır, Ardahan TRB1- Malatya, Elazığ, Bingöl, Tunceli TRB2- Van, Muş, Bitlis, Hakkari TRC1- Gaziantep, Adıyaman, Kilis TRC2- Şanlıurfa, Diyarbakır TRC3- Mardin, Batman, Şırnak, Siirt
HB040	Hane için ağırlık katsayısı	4.7	Format 4.7
HB050	Hanehalkı tipi *: "1", "2", "3" ve "4" kodları FK070 (fert yaş değişkeni) ve FK090 (fert cinsiyet değişkeni) ile oluşturulacaktır. **: "14" kodu "6", "7" ve "8" kodlarının toplamından oluşturulmuştur. ***: "15" kodu "9", "10", "11", "12" ve "13" kodlarının toplamından oluşturulmuştur.	2	1- 64 yaşın altında* 2- 65 ve daha fazla yaşta* 3- Erkek* 4- Kadın* 5-Tek kişilik hanehalkı 6- Bağımlı çocuk olmayan, iki yetişkinli, yetişkinlerin ikisi de 65 yaşın altında 7- Bağımlı çocuk olmayan, iki yetişkinli, yetişkinlerden en az biri 65 ve daha fazla yaşta 8- Bağımlı çocuk olmayan diğer hanehalkı 9- Tek yetişkinli, en az bir bağımlı çocuk olan hanehalkı 10- İki yetişkinli, bir bağımlı çocuk olan hanehalkı 11- İki yetişkinli, iki bağımlı çocuk olan hanehalkı 12- İki yetişkinli, üç ve daha fazla bağımlı çocuk olan hanehalkı 13- Bağımlı çocuk olan diğer hanehalkı 14-Bağımlı çocuk olmayan hanehalkları (6+7+8)** 15-Bağımlı çocuk olan hanehalkları (9+10+11+12+13)*** 16- Diğer (hanehalkı tipi tespit edilemeyenler)
Konut ve Konut Kolaylıkları			
HH010	Oturulan konutun tipi	1	1- Müstakil konut 2- İkiz ya da sıralı ev 3- Apartman (10 daireden az)

			4- Apartman (10 ve daha fazla daire) 5- Diğer
HH020	Oturulan konuttaki mülkiyet şekli	1	1- Ev sahibi 2- Kiracı 3- Lojman 4- Ev sahibi değil ama kira ödemiıyor
HH030	Oturulan konuta sahip olunan veya oturmaya başlanan yıl	4	1900...2015
HH040	Oturulan konut için ödenen aylık kira tutarı (TL) (Konut için ödenen kiraya; kapıcı, yakacak, aidat, vb. diğer ödemeler dahil edilmemiştir) [HH020 "2" veya "3" olanlar]	6	0..999999
HH050	Aylık ortalama konut masrafı (TL) (Su elektrik faturaları, konut için ödenen aylık kira tutarı, yakacak, konut hizmeti, bakım masrafları, emlak vergisi vb.)	6	0..999999
HH060	Oturulan konuttaki (mutfak, banyo ve tuvalet hariç) oda sayısı	2	1- 1 2- 2 3- 3 4- 4 5- 5 6- 6 7- 7 8- 8 9- 9 10- 10 ve daha fazla
HH070	Oturulan konuttaki faydalanılan alanın metrekare olarak büyüklüğü (m2)	3	25...999
HH080	Oturulan konutun ısıtma sistemi	1	1- Soba (Kömür, gaz, doğalgaz, elektrik, vb.) 2- Kaldorifer (Müşterek veya merkezi ısıtma) 3- Kaldorifer (Kat kaldoriferi/kombi)

			4- Klima 5- Diğer 6- Isıtma sistemi yok
HH090	Oturulan konutta ısınma amacıyla en fazla kullanılan yakıt türü	1	1- Odun 2- Kömür 3- Doğalgaz 4- Fuel-Oil 5- LPG (Tüpgaz) 6- Elektrik 7- Tezek 8- Diğer
HH100	Oturulan konutta BANYO veya DUŞ'un mevcut olma durumu	1	1- Evet (Sadece kendi kullanımında) 2- Evet (Paylaşılan) 3- Hayır
HH110	Oturulan konutta TUVALET'in (ev içinde) mevcut olma durumu	1	1- Evet (Sadece kendi kullanımında) 2- Evet (Paylaşılan) 3- Hayır
HH120	Oturulan konutta BAĞIMSIZ MUTFAĞIN mevcut olma durumu	1	1- Evet 2- Hayır
HH130	Oturulan konutta BORULU SU SİSTEMİ'nin mevcut olma durumu	1	1- Evet 2- Hayır
HH140	Oturulan konutta SICAK SU SİSTEMİ'nin (Merkezi sıcak su, şofben,kombi, güneş enerjisi,vb.) mevcut olma durumu	1	1- Evet 2- Hayır
HH150	Hanehalkının TELEFON HATTI'na (sabit) sahip olma durumu	1	1- Evet 2- Hayır, maddi yetersizlik 3- Hayır, diğer nedenler
HH160	Hanehalkının CEP TELEFONU'na sahip olma durumu	1	1- Evet 2- Hayır, maddi yetersizlik 3- Hayır, diğer nedenler
HH170	Hanehalkının RENKLİ TELEVİZYON'a sahip olma durumu	1	1- Evet 2- Hayır, maddi yetersizlik 3- Hayır, diğer nedenler

HH180	Hanehalkının BİLGİSAYAR'a sahip olma durumu	1	1- Evet 2- Hayır, maddi yetersizlik 3- Hayır, diğer nedenler
HH190	Hanehalkının İNTERNET'e sahip olma durumu [HH180 = "1" olanlar]	1	1- Evet 2- Hayır, maddi yetersizlik 3- Hayır, diğer nedenler
HH200	Hanehalkının OTOMATİK ÇAMAŞIR MAKİNASI'na sahip olma durumu	1	1- Evet 2- Hayır, maddi yetersizlik 3- Hayır, diğer nedenler
HH210	Hanehalkının BUZDOLABI'na sahip olma durumu	1	1- Evet 2- Hayır, maddi yetersizlik 3- Hayır, diğer nedenler
HH220	Hanehalkının BULAŞIK MAKİNASI'na sahip olma durumu	1	1- Evet 2- Hayır, maddi yetersizlik 3- Hayır, diğer nedenler
HH230	Hanehalkının KLİMA'ya sahip olma durumu	1	1- Evet 2- Hayır, maddi yetersizlik 3- Hayır, diğer nedenler
HH240	Hanehalkının OTOMOBİL'e (iş amaçlılar hariç) sahip olma durumu	1	1- Evet 2- Hayır, maddi yetersizlik 3- Hayır, diğer nedenler
Konut ve Çevre Problemleri			
HS010	Oturulan konutta sızdıran çatı, nemli duvarlar, çürümüş pencere çerçeveleri gibi bir problemin mevcut olma durumu	1	1- Evet 2- Hayır
HS020	Oturulan konutta, konutun izolasyonundan dolayı ısınma sorunu olup olmadığı	1	1- Evet 2- Hayır
HS030	Oturulan konutta, odaların karanlık olması veya yeterli ışık almaması gibi bir sorun olup olmadığı	1	1- Evet 2- Hayır

HS040	Oturulan konutta, komşulardan veya sokaktan gelen gürültü probleminin olup olmadığı	1	1- Evet 2- Hayır
HS050	Oturulan konutun kullanım alanının yeterli olup olmadığı	1	1- Evet 2- Hayır
HS060	Yaşanılan çevrede trafik veya endüstrinin neden olduğu hava kirliliği, çevre kirliliği veya diğer çevresel sorunların olup olmadığı	1	1- Evet 2- Hayır
HS070	Yaşanılan çevrede suç veya şiddet olayları ile yoğun bir şekilde karşılaşılıp karşılaşılmadığı	1	1- Evet 2- Hayır
Ekonomik Durum			
HE010	Son 12 ay içinde EV KİRASININ, FAİZLİ BORÇ GERİ ÖDEMESİNİN veya KONUT KREDİSİNİN planlandığı gibi ödenememe durumu	1	1- Evet (Bir kez) 2- Evet (2 veya daha fazla) 3- Hayır olmadı 4- Bu tür ödeme yok
HE020	Son 12 ay içinde ELEKTRİK, SU, GAZ FATURALARININ planlandığı gibi ödenememe durumu (Borç alınarak ödeme yapılanlar ödenmiş kabul edilecektir)	1	1- Evet (Bir kez) 2- Evet (2 veya daha fazla) 3- Hayır olmadı 4- Bu tür ödeme yok
HE030	Son 12 ay içinde TAKSİT, KREDİ KARTI ve DİĞER BORÇ ÖDEMELERİNİN planlandığı gibi ödenememe durumu	1	1- Evet (Bir kez) 2- Evet (2 veya daha fazla) 3- Hayır olmadı 4- Bu tür ödeme yok

HE040	Hanenin toplam geliri ile gerekli harcamaları yapabilme durumu	1	1- Çok zor 2- Zor 3- Biraz zor 4- Biraz kolay 5- Kolay 6- Çok kolay
HE050	Bir ay boyunca geçinebilmek için hanenin sahip olması gereken en düşük aylık net gelirin ne kadar olacağı	7	1..9999999
HE060	Konuta ilişkin olarak yapılan tüm harcamalar düşünüldüğünde bu masrafların haneye nasıl bir yük getirdiği (Ev sahipleri için konut kredisi faizi geri ödemeleri, kiracılar için ödenen kira dahil olmak üzere, su, elektrik, ısınma, fatura cezaları, apartman aidatları, düzenli tamir ve bakım masrafları kapsanmıştır)	1	1- Çok yük getiriyor 2- Biraz yük getiriyor 3- Yük getirmiyor
HE070	Konut ile ilgili ödemeler dışındaki borç ve taksit ödemelerinin haneye nasıl bir yük getirdiği	1	1- Çok yük getiriyor 2- Biraz yük getiriyor 3- Yük getirmiyor 4- Taksit veya borç ödemesi yok
HE080	Haneniz ekonomik olarak; tüm hanehalkının evden uzakta bir haftalık tatil masrafını karşılayabilecek durumda mı? (Tatil köyü, otel, pansiyon, haneye ait yazlık, bir yakınının evi, kurum /devlete ait kamplar vb. yerlerde yapılan tatiller dahil edilecektir)	1	1- Evet 2- Hayır
HE090	Haneniz ekonomik olarak; 2 günde bir et, tavuk ya da balık içeren yemeği karşılayabilecek durumda mı? (Vejeteryanlar için eşdeğer yiyecekler)	1	1- Evet 2- Hayır

HE100	Haneniz ekonomik olarak; beklenmedik mecburi bir harcamayı (yaklaşık 500 TL'lik) kendi imkanlarıyla karşılatabilecek durumda mı?	1	1- Evet 2- Hayır
HE110	Haneniz ekonomik olarak; Evinizin ısıtma masrafını karşılatabilecek durumda mı?	1	1- Evet 2- Hayır
Maddi Yoksunluk			
HE120	Mobilyalarınız yıprandığında veya eskidiğinde yeniler misiniz? (İkinci el mobilya alımları dahil edilecektir)	1	1- Evet 2- Hayır- Maddi yetersizlik 3- Hayır- Diğer nedenler
Gelir			
HG010	Oturulan konutun yıllık izafi kirası (Mülkiyet şekli "ev sahibi " ve "diğer" olanlar için; HH020 ="1" veya "4" olan)	9.2	0..999999.99
HG020	Hanedeki 15 yaşından küçük fertlerin 2014 yılı içinde elde ettiği gelirin toplam değeri (TL)	9.2	0..999999.99
HG030N	2014 yılı içinde alınan NAKDİ çocuk yardımlarının DEĞERİ (TL)	9.2	0..999999.99
HG030A	2014 yılı içinde alınan AYNI çocuk yardımlarının DEĞERİ (TL)	9.2	0..999999.99
HG040	2014 yılı içerisinde alınan konut yardımlarının (kira, vb.) değeri (TL)	9.2	0..999999.99

HG050N	2014 yılı içinde alınan diğer NAKDİ sosyal yardımların DEĞERİ (TL)	9.2	0..999999.99
HG050A	2014 yılı içinde alınan diğer AYNİ sosyal yardımların DEĞERİ (TL)	9.2	0..999999.99
HG060N	2014 yılı içinde başka bir kişi veya haneden düzenli alınan NAKDİ yardımların DEĞERİ (Nafaka hariç) (TL)	9.2	0..999999.99
HG060A	2014 yılı içinde başka bir kişi veya haneden düzenli alınan AYNİ yardımların DEĞERİ (TL)	9.2	0..999999.99
HG065N	2014 yılı içinde alınan nafaka (TL)	9.2	0..999999.99
HG070	2014 yılı gayrimenkul (kira) geliri (TL)	9.2	0..999999.99
HG080	2014 yılı menkul kıymet geliri (TL)	9.2	0..999999.99
HG085	2014 yılı içinde tarımsal bir faaliyet olarak kapsanmayıp sadece hane tüketimine yönelik olarak üretilen bitkisel veya hayvansal ürünlerin piyasa fiyatı ile değeri (TL)	9.2	0..999999.99
HG090N	2014 yılı içinde haneden başka bir kişi veya haneye düzenli yapılan NAKDİ yardımların DEĞERİ (Nafaka hariç) (TL)	9.2	0..999999.99
HG090A	2014 yılı içinde haneden başka bir kişi veya haneye düzenli	9.2	0..999999.99

	yapılan AYNİ yardımların DEĞERİ (TL)		
HG095N	2014 yılı içinde verilen nafaka (TL)	9.2	0..999999.99
HG100	2014 yılında ödenen düzenli vergiler (emlak vergisi, motorlu taşıtlar vergisi ve düzenli ödenen varlık vergileri) (TL) (Gelir vergisi hariç)	9.2	0..999999.99
HG103	2014 yılında ödenen isteğe bağlı emeklilik primi (TL)	9.2	0..999999.99
HG105	Cevapsız fert/fertler için impute edilmiş yıllık gelir (TL)	9.2	-999999..999999
HG110	Toplam hanehalkı kullanılabilir geliri (FG010 + FG020 + FG030 + FG040 + FG070 + FG080 + FG085 + FG090 + FG100 + FG110 + FG120 + HG010 + HG020 + HG030N + HG030A + HG040 + HG050N + HG050A + HG060N + HG060A + HG065N + HG070 + HG080 + HG085 + HG105 - HG090N - HG090A - HG095N - HG100 - HG103)	9.2	0..999999.99

B

ALGORİTMA SONUÇLARI VE R KODLARI

B.1 Gizli Sınıf Olasılıkları

Tablo B.1 Değişken değerlerine ilişkin gizli sınıf olasılıkları

Değişken	Olasılık	Gizli Sınıf									
		1	2	3	4	5	6	7	8	9	10
STA	Pr(1)	0.00	0.52	0.84	0.92	0.03	0.00	0.35	0.00	0.69	0.89
	Pr(2)	0.00	0.02	0.04	0.00	0.02	0.00	0.60	0.00	0.14	0.05
	Pr(3)	0.00	0.14	0.04	0.06	0.03	0.00	0.00	0.00	0.00	0.01
	Pr(4)	0.00	0.30	0.08	0.03	0.83	0.00	0.04	0.00	0.15	0.03
	Pr(5)	0.00	0.01	0.00	0.00	0.10	0.00	0.01	0.00	0.02	0.01
	Pr(6)	1.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00
OCC	Pr(1)	0.00	0.14	0.01	0.62	0.00	0.00	0.00	0.00	0.00	0.06
	Pr(2)	0.00	0.07	0.10	0.28	0.00	0.00	0.00	0.00	0.01	0.19
	Pr(3)	0.00	0.60	0.01	0.09	0.00	0.00	0.02	0.00	0.47	0.35
	Pr(4)	0.00	0.00	0.41	0.00	0.98	0.00	0.12	0.00	0.02	0.11
	Pr(5)	0.00	0.09	0.36	0.00	0.00	0.00	0.40	0.00	0.11	0.17
	Pr(6)	0.00	0.11	0.11	0.00	0.02	0.00	0.46	0.00	0.38	0.12
	Pr(7)	1.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00
RGN	Pr(1)	0.30	0.24	0.38	0.28	0.16	0.20	0.11	0.25	0.28	0.32

	Pr(2)	0.16	0.12	0.16	0.13	0.16	0.13	0.08	0.12	0.19	0.14
	Pr(3)	0.19	0.16	0.16	0.19	0.12	0.12	0.10	0.16	0.13	0.17
	Pr(4)	0.11	0.13	0.09	0.11	0.09	0.12	0.15	0.14	0.14	0.10
	Pr(5)	0.14	0.10	0.11	0.12	0.21	0.16	0.08	0.08	0.13	0.10
	Pr(6)	0.07	0.13	0.05	0.11	0.20	0.15	0.26	0.12	0.06	0.09
	Pr(7)	0.04	0.11	0.06	0.06	0.07	0.12	0.22	0.12	0.07	0.09
INT	Pr(1)	0.46	0.46	0.54	0.85	0.20	0.18	0.10	0.53	0.42	0.53
	Pr(2)	0.54	0.54	0.46	0.15	0.80	0.82	0.90	0.47	0.58	0.47
MST	Pr(1)	0.87	0.95	0.94	0.81	0.90	0.38	0.75	0.21	0.80	0.21
	Pr(2)	0.13	0.05	0.06	0.19	0.10	0.62	0.25	0.79	0.20	0.79
AGE	Pr(1)	0.00	0.01	0.00	0.03	0.02	0.01	0.22	0.68	0.00	0.58
	Pr(2)	0.04	0.24	0.30	0.40	0.08	0.06	0.28	0.22	0.19	0.38
	Pr(3)	0.05	0.36	0.38	0.34	0.20	0.10	0.26	0.07	0.42	0.04
	Pr(4)	0.23	0.27	0.25	0.18	0.29	0.10	0.17	0.03	0.27	0.00
	Pr(5)	0.40	0.10	0.05	0.05	0.25	0.13	0.06	0.00	0.08	0.00
	Pr(6)	0.28	0.03	0.00	0.01	0.15	0.60	0.01	0.00	0.04	0.00
GEN	Pr(1)	0.76	0.96	0.94	0.64	0.83	0.31	0.96	0.49	0.20	0.69
	Pr(2)	0.24	0.04	0.06	0.36	0.17	0.69	0.04	0.51	0.80	0.32
SEC	Pr(1)	0.00	0.00	0.01	0.00	0.94	0.00	0.19	0.00	0.06	0.00
	Pr(2)	0.00	0.07	0.54	0.09	0.00	0.00	0.09	0.00	0.19	0.29
	Pr(3)	0.00	0.04	0.14	0.02	0.00	0.00	0.55	0.00	0.00	0.04
	Pr(4)	0.00	0.45	0.05	0.05	0.00	0.00	0.04	0.00	0.09	0.24

	Pr(5)	0.00	0.01	0.15	0.02	0.06	0.00	0.05	0.00	0.00	0.03
	Pr(6)	0.00	0.35	0.06	0.43	0.00	0.00	0.07	0.00	0.43	0.32
	Pr(7)	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.01
	Pr(8)	0.00	0.08	0.04	0.22	0.00	0.00	0.01	0.00	0.00	0.02
	Pr(9)	0.00	0.01	0.01	0.11	0.00	0.00	0.00	0.00	0.22	0.05
	Pr(10)	1.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00
EDU	Pr(1)	0.03	0.04	0.01	0.00	0.18	0.67	0.23	0.06	0.19	0.03
	Pr(2)	0.51	0.44	0.49	0.00	0.65	0.32	0.40	0.01	0.63	0.00
	Pr(3)	0.12	0.19	0.19	0.01	0.09	0.00	0.29	0.42	0.08	0.41
	Pr(4)	0.10	0.17	0.10	0.09	0.04	0.00	0.05	0.21	0.06	0.18
	Pr(5)	0.09	0.12	0.17	0.08	0.03	0.00	0.03	0.15	0.03	0.19
	Pr(6)	0.15	0.05	0.03	0.82	0.02	0.00	0.01	0.15	0.01	0.18
IG	Pr(1)	0.03	0.02	0.01	0.02	0.22	0.50	0.29	0.77	0.33	0.24
	Pr(2)	0.23	0.14	0.08	0.03	0.20	0.39	0.40	0.16	0.41	0.27
	Pr(3)	0.37	0.21	0.25	0.05	0.16	0.08	0.18	0.05	0.18	0.33
	Pr(4)	0.28	0.34	0.43	0.14	0.22	0.02	0.12	0.01	0.07	0.15
	Pr(5)	0.09	0.30	0.23	0.77	0.21	0.00	0.01	0.01	0.01	0.01

B.2 10 Sınıflı GSA Model Çözümünde Gizli Sınıf Tanımları

Tablo B.2 GSA model çözümünde gizli sınıf tanımları

Sınıf	Bölüm	Tanım
1: "Orta gelirli, düşük eğitilmiş yaşlı erkekler, ikinci en büyük dilim"	%14 (n = 5185)	İstihdamda değil
		Mesleği yok
		Marmara Bölgesinde
		İnternet bağlantısına sahip değil
		Evli
		55-64 yaş grubunda
		Erkek
		Sektörü yok
		İlkokul mezunu
2: "Ticaret sektöründe ücretli çalışan"	%10 (n = 3574)	Ücretli çalışan
		Hizmet ve satış elemanı
		Marmara Bölgesinde
		İnternet bağlantısına sahip değil
		Evli
		35-44 yaş grubunda
		Erkek

		Ticaret sektöründe çalışan
		İlkokul mezunu
		Dördüncü %20'lik gelir grubunda
3: "Sanayi sektöründe ücretli çalışan"	%11 (n = 4115)	Ücretli çalışan
		Tesis ve Makine Operatörü/Montajcı
		Marmara Bölgesinde
		İnternet bağlantısına sahip
		Evli
		35-44 yaş grubunda
		Erkek
		Sanayi sektöründe çalışan
		İlkokul mezunu
		Dördüncü %20'lik gelir grubunda
4: "En yüksek gelirli ve en eğitilmiş"	%14 (n = 5142)	Ücretli çalışan
		Yönetici/Profesyonel
		Marmara Bölgesinde
		İnternet bağlantısına sahip
		Evli
		25-34 yaş grubunda
		Erkek

		Hizmet sektöründe çalışan
		Yüksek öğrenim mezunu
		En yüksek %20'lik gelir grubunda
5: "Karadeniz bölgesinde tarım sektöründe kendi hesabına çalışan"	%9 (n = 3106)	Kendi hesabına çalışan
		Tarımsal meslek sahibi
		Karadeniz bölgesinde
		İnternet bağlantısına sahip değil
		Evli
		45-54 yaş grubunda
		Erkek
		Tarım sektöründe çalışan
		İlkokul mezunu
		Dördüncü %20'lik gelir grubunda
6: "En düşük gelirli ve en eğitimsiz, en büyük dilim"	%15 (n = 5466)	İstihdamda değil
		Mesleği yok
		Marmara Bölgesinde
		İnternet bağlantısına sahip değil
		Evli değil
		65 yaş ve üstü yaş grubunda

		Kadın
		Sektörü yok
		Okur yazar olmayan/bir okul bitirmeyen
		En düşük %20'lik gelir grubunda
7: "Düşük gelirli, Doğu Anadolu bölgesinde inşaat sektöründe yevmiyeli çalışan"	%6 (n = 1988)	Yevmiyeli çalışan
		Nitelik gerektirmeyen işlerde çalışan
		Doğu Anadolu Bölgesinde
		İnternet bağlantısına sahip değil
		Evli
		25-34 yaş grubunda
		Erkek
		İnşaat sektöründe çalışan
		İlkokul mezunu
		İkinci %20'lik gelir grubunda
8: "En düşük gelirli ve istihdamda olmayan genç bayanlar"	%6 (n = 2306)	İstihdamda değil
		Mesleği yok
		Marmara Bölgesinde
		İnternet bağlantısına sahip
		Evli değil

		15-24 yaş grubunda
		Kadın
		Sektörü yok
		Ortaokul mezunu
		En düşük %20'lik gelir grubunda
9: "Düşük gelirli, orta yaşta evli olan bayanlar"	%6 (n = 1992)	Ücretli çalışan
		Hizmet ve satış elemanı
		Marmara Bölgesinde
		İnternet bağlantısına sahip değil
		Evli
		35-44 yaş grubunda
		Kadın
		Hizmet sektöründe çalışan
		İlkokul mezunu
		İkinci %20'lik gelir grubunda
10: "Orta gelirli, evli olmayan genç yaşta erkekler"	%9 (n = 3160)	Ücretli çalışan
		Hizmet ve satış elemanı
		Marmara Bölgesinde
		İnternet bağlantısına sahip
		Evli değil

		15-24 yaş grubunda
		Erkek
		Hizmet sektöründe çalışan
		Ortaokul mezunu
		Üçüncü %20'lik gelir grubunda



B.3 10 Kümelili k-modlar Model Çözümünde Küme Tanımları

Tablo B.3 k-modlar model çözümünde küme tanımları

Küme	Büyüklük	Tanım
1: "En yüksek gelirli ve en eğitilmiş genç bayanlar"	%9 (n = 3405)	Ücretli çalışan
		Yönetici/Profesyonel
		Karadeniz Bölgesinde
		İnternet bağlantısına sahip
		Evli
		25-34 yaş grubunda
		Bayan
		Hizmet ve satış sektöründe çalışan
		Yüksek öğrenim mezunu
		En yüksek %20'lik gelir grubunda
2: "En düşük gelirli ve en eğitimsiz, en büyük dilim"	%22 (n = 7840)	İstihdamda değil
		Mesleği yok
		Marmara Bölgesinde
		İnternet bağlantısına sahip değil
		Evli değil
		65 yaş ve üstü yaş grubunda
		Kadın
		Sektörü yok

		Okur yazar olmayan/bir okul bitirmeyen
		En düşük %20'lik gelir grubunda
3: "Yüksek gelirli, Akdeniz Bölgesinde sanayi sektöründe ücretli çalışan genç erkekler"	%8 (n = 2893)	Ücretli çalışan
		Sanatkâr ve ilgili işlerde çalışan
		Akdeniz Bölgesinde
		İnternet bağlantısına sahip değil
		Evli
		15-24 yaş grubunda
		Erkek
		Sanayi sektöründe çalışan
		Ortaokul mezunu
		En yüksek %20'lik gelir grubunda
4: "Doğu Anadolu Bölgesinde tarım sektöründe kendi hesabına çalışan"	%13 (n = 4723)	Kendi hesabına çalışan
		Tarımsal meslek sahibi
		Doğu Anadolu Bölgesinde
		İnternet bağlantısına sahip değil
		Evli
		35-44 yaş grubunda
		Erkek
		Tarım sektöründe çalışan
		İlkokul mezunu

		İkinci %20'lik gelir grubunda
5: "Sanayi sektöründe ücretli çalışan"	%10 (n = 3587)	Ücretli çalışan
		Tarımsal meslek sahibi
		Marmara Bölgesinde
		İnternet bağlantısına sahip değil
		Evli
		25-34 yaş grubunda
		Erkek
		Sanayi sektöründe çalışan
		İlkokul mezunu
		Dördüncü %20'lik gelir grubunda
6: "Düşük gelirli orta yaşlı bayanlar"	%5 (n = 1653)	Ücretli çalışan
		Hizmet ve satış elemanı
		Marmara Bölgesinde
		İnternet bağlantısına sahip değil
		Evli
		45-54 yaş grubunda
		Bayan
		Hizmet sektöründe çalışan
		İlkokul mezunu
		En düşük %20'lik gelir grubunda
	%5 (n = 1899)	Ücretli çalışan

7: "Yüksek gelirli ve en eğitilmiş erkekler"		Teknisyen/Yardımcı Profesyonel
		Marmara Bölgesinde
		İnternet bağlantısına sahip
		Evli değil
		25-34 yaş grubunda
		Erkek
		Hizmet sektöründe çalışan
		Yüksek öğrenim mezunu
		Dördüncü %20'lik gelir grubunda
8: "Düşük gelirli, ticaret sektöründe ücretli çalışan evli olmayan erkekler"	%4 (n = 1434)	Ücretli çalışan
		Hizmet ve satış elemanı
		İç Anadolu Bölgesinde
		İnternet bağlantısına sahip değil
		Evli değil
		25-34 yaş grubunda
		Erkek
		Ticaret sektöründe çalışan
		Ortaokul mezunu
		İkinci %20'lik gelir grubunda
9: "Orta gelirli, eğitim seviyesi düşük yaşlı erkekler, ikinci en büyük dilim"	%15 (n = 5390)	İstihdamda değil
		Mesleği yok
		Marmara Bölgesinde

		İnternet bağlantısına sahip değil
		Evli
		55-64 yaş grubunda
		Erkek
		Sektörü yok
		İlkokul mezunu
		Üçüncü %20'lik gelir grubunda
10: "En yüksek gelirli ve en eğitilmiş orta yaşlı evli erkekler"	%9 (n = 3212)	Ücretli çalışan
		Yönetici/Profesyonel
		Marmara Bölgesinde
		İnternet bağlantısına sahip
		Evli
		35-44 yaş grubunda
		Erkek
		Hizmet sektöründe çalışan
		Yüksek öğrenim mezunu
		En yüksek %20'lik gelir grubunda

B.4 RO Algoritması R Kodu ve Çıktıları

Veri seti eğitim ve test verisi olmak üzere 70/30 oranında bölümlenmekte ve bağımsız örneklemeler olarak “ind” değişkeninde saklanmaktadır. Eğitim verisinin RO modeli için hata matrisi çıktısı Tablo B.4’te yer almaktadır.

```
> set.seed(123)

> ind <- sample(2, nrow(Data_RF), replace = TRUE, prob = c(0.7, 0.3))

> train <- Data_RF[ind==1,]

> test <- Data_RF[ind==2,]

# Başlangıç RO Modeli elde edilmektedir

> library(randomForest)

> set.seed(123)

> rf <- randomForest(IG ~., data=train)

> print(rf)
```

Call:

```
randomForest(formula = IG ~ ., data = train)
```

Type of random forest: classification

Number of trees: 500

No. of variables tried at each split: 3

OOB estimate of error rate: 50.5%

Confusion matrix:

Tablo B.4 Eğitim verisinin RO modeli için hata matrisi çıktısı

	1_Quantile	2_Quantile	3_Quantile	4_Quantile	5_Quantile	class.error
1_Quantile	3332	1207	401	230	133	0.3716764
2_Quantile	1225	2283	1057	653	288	0.5853614

3_Quantile	499	1150	1593	1165	409	0.6692276
4_Quantile	238	647	986	1951	965	0.5924379
5_Quantile	90	174	340	963	3406	0.3151015

Eğitim verisi için hata matrisi ve tahmin modeli elde edilmektedir. Eğitim verisinin tahmini için Tablo B5'teki hata matrisi çıktısı ile Tablo B6'daki model istatistikleri çıktısı elde edilmektedir.

```
> library(caret)
```

```
> p1 <- predict(rf, train)
```

```
> confusionMatrix(p1, train$IG)
```

Reference

Tablo B.5 Eğitim verisinin tahmini için hata matrisi çıktısı

Prediction	1_Quantile	2_Quantile	3_Quantile	4_Quantile	5_Quantile
1_Quantile	3973	879	342	161	66
2_Quantile	834	3365	746	395	107
3_Quantile	250	688	2824	584	238
4_Quantile	150	397	662	3140	578
5_Quantile	96	177	242	507	3984

Overall Statistics

Accuracy : 0.681

95% CI : (0.6752, 0.6867)

No Information Rate : 0.2169

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.6009

McNemar's Test P-Value : 2.078e-07

Statistics by Class:

Tablo B.6 Eğitim verisinin tahmini için model istatistikleri çıktısı

Class:	1_Quantile	2_Quantile	3_Quantile	4_Quantile	5_Quantile
Sensitivity	0.7492	0.6112	0.5864	0.6559	0.8011
Specificity	0.9279	0.8953	0.9144	0.9132	0.9499
Pos Pred Value	0.7329	0.6178	0.6161	0.6373	0.7958
Neg Pred Value	0.9334	0.8926	0.9042	0.9195	0.9515
Prevalence	0.2089	0.2169	0.1897	0.1886	0.1959
Detection Rate	0.1565	0.1326	0.1112	0.1237	0.1569
Detection Prevalence	0.2136	0.2146	0.1806	0.1941	0.8385
Balanced Accuracy	0.8385	0.7532	0.7504	0.7846	0.8755

Test verisi için hata matrisi ve tahmin modeli elde edilmektedir. Test verisi tahmini için Tablo B7'deki hata matrisi çıktısı ile Tablo B8'deki model istatistikleri çıktısı elde edilmektedir.

```
> p2 <- predict(rf, test)
```

```
> confusionMatrix(p2, test$IG)
```

Confusion Matrix and Statistics

Reference

Tablo B.7 Test verisi tahmini için hata matrisi çıktısı

Prediction	1_Quantile	2_Quantile	3_Quantile	4_Quantile	5_Quantile
1_Quantile	1407	480	208	107	38
2_Quantile	527	985	424	271	73
3_Quantile	141	433	676	454	132
4_Quantile	101	250	458	832	399
5_Quantile	65	121	160	386	1523

Overall Statistics

Accuracy : 0.5092

95% CI : (0.4996, 0.5187)

No Information Rate : 0.213

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.3857

Mcnemar's Test P-Value : 3.732e-05

Statistics by Class:

Tablo B.8 Test verisi tahmini için model istatistikleri çıktısı

Class:	1_Quantile	2_Quantile	3_Quantile	4_Quantile	5_Quantile
Sensitivity	0.6278	0.43411	0.35099	0.40585	0.7035
Specificity	0.9010	0.84550	0.86705	0.85955	0.9137
Pos Pred Value	0.6281	0.43202	0.36819	0.40784	0.6754
Neg Pred Value	0.9008	0.84661	0.85820	0.85855	0.9235
Prevalence	0.2104	0.21303	0.18083	0.19247	0.2033

Detection Rate	0.1321	0.09248	0.06347	0.07811	0.1430
Detection Prevalence	0.2103	0.21406	0.17238	0.19153	0.2117
Balanced Accuracy	0.7644	0.63981	0.60902	0.63270	0.8086

RO modelindeki hata oranını öğrenmek için RO modelini kullanarak grafik oluşturulmaktadır.

```
> plot(rf)
```

En uygun parametre değerleri belirlenerek RO modeli iyileştirilmektedir.

```
> t <- tuneRF(train[, -10], train[, 10], stepFactor = 0.5, plot = TRUE, ntreeTry = 310, trace = TRUE, improve = 0.05)
```

Çıktılar:

$m_{try} = 3$ OOB error = 19.72%

Searching left ...

$m_{try} = 6$ OOB error = 21.13%

-0.07192807 0.05

Searching right ...

$m_{try} = 1$ OOB error = 22.4%

-0.1360639 0.05

ntree=300 ve $m_{try} = 3$ için başlangıç RO modeli tekrar oluşturulmaktadır. Eğitim verisinin iyileştirilmiş modeli için hata matrisi çıktısı Tablo B9'da yer almaktadır.

```
> library(randomForest)
```

```
> set.seed(123)
```

```
> rf1 <- randomForest(IG ~., data=train, ntree = 310,  $m_{try} = 3$ )
```

```
> print(rf1)
```

Çıktılar:

Call:

```
randomForest(formula = IG ~ ., data = train, ntree = 310, mtry = 3)
```

Type of random forest: classification

Number of trees: 310

No. of variables tried at each split: 3

OOB estimate of error rate: 50.45%

Confusion matrix:

Tablo B.9 Eğitim verisi için iyileştirilmiş modelin hata matrisi çıktısı

	1_Quantile	2_Quantile	3_Quantile	4_Quantile	5_Quantile	class.error
1_Quantile	3348	1184	402	235	134	0.3686592
2_Quantile	1234	2276	1072	645	279	0.5866328
3_Quantile	508	1137	1592	1178	401	0.6694352
4_Quantile	239	645	993	1950	960	0.5926468
5_Quantile	93	174	341	952	3413	0.3136939

Eğitim verisinin hatası için model tahmini yapılmaktadır. Eğitim verisinin tahmini için iyileştirilmiş modelin hata matrisi çıktısı Tablo B10'da, model istatistikleri çıktısı ise Tablo B.11'de elde edilmektedir.

```
> library(caret)
```

```
> p11 <- predict(rf1, train)
```

```
> confusionMatrix(p1, train$IG)
```

Çıktılar:

Confusion Matrix and Statistics

Reference

Tablo B.10 Eğitim verisinin tahmini için iyileştirilmiş modelin hata matrisi çıktısı

Prediction	1_Quantile	2_Quantile	3_Quantile	4_Quantile	5_Quantile
1_Quantile	3973	879	342	161	66
2_Quantile	835	3365	746	393	106
3_Quantile	250	688	2826	584	239
4_Quantile	149	398	661	3140	577
5_Quantile	96	176	241	509	3985

Overall Statistics

Accuracy : 0.6811

95% CI : (0.6753, 0.6868)

No Information Rate : 0.2169

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.601

Mcnemar's Test P-Value : 2.423e-07

Statistics by Class:

Tablo B.11 Eğitim verisinin tahmini için iyileştirilmiş modelin model istatistikleri çıktısı

Class:	1_Quantile	2_Quantile	3_Quantile	4_Quantile	5_Quantile
Sensitivity	0.7492	0.6112	0.5868	0.6559	0.8013
Specificity	0.9279	0.8954	0.9144	0.9133	0.9499
Pos Pred Value	0.7329	0.6180	0.6161	0.6376	0.7959

Neg Pred Value	0.9334	0.8926	0.9043	0.9195	0.9515
Prevalence	0.2089	0.2169	0.1897	0.1886	0.1959
Detection Rate	0.1565	0.1326	0.1113	0.1237	0.1570
Detection Prevalence	0.2136	0.2145	0.1807	0.1940	0.1972
Balanced Accuracy	0.8385	0.7533	0.7506	0.7846	0.8756

Test verisi için iyileştirilmiş modelin hata matrisi hesaplanmaktadır. Hata matrisine ilişkin Tablo B12'deki çıktı, model istatistikleri için Tablo B13'teki çıktı elde edilmektedir.

```
> p22 <- predict(rf1, test)
```

```
> confusionMatrix(p22, test$IG)
```

Çıktılar:

Confusion Matrix and Statistics

Reference

Tablo B.12 Test verisi için iyileştirilmiş modelin hata matrisi çıktısı

Prediction	1_Quantile	2_Quantile	3_Quantile	4_Quantile	5_Quantile
1_Quantile	1410	479	210	104	38
2_Quantile	517	980	430	270	76
3_Quantile	146	431	664	449	131
4_Quantile	105	258	462	845	401
5_Quantile	63	121	160	382	1519

Overall Statistics

Accuracy : 0.5087

95% CI : (0.4991, 0.5182)

No Information Rate : 0.213

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.3852

Mcnemar's Test P-Value : 0.0002484

Statistics by Class:

Tablo B.13 Test verisi için iyileştirilmiş modelin model istatistikleri çıktısı

Class:	1_Quantile	2_Quantile	3_Quantile	4_Quantile	5_Quantile
Sensitivity	0.6292	0.43191	0.34476	0.41220	0.7016
Specificity	0.9012	0.84574	0.86739	0.85746	0.9144
Pos Pred Value	0.6292	0.43115	0.36463	0.40802	0.6766
Neg Pred Value	0.9012	0.84614	0.85708	0.85956	0.9232
Prevalence	0.2104	0.21303	0.18083	0.19247	0.2033
Detection Rate	0.1324	0.09201	0.06234	0.07934	0.1426
Detection Prevalence	0.2104	0.21341	0.17097	0.19444	0.2108
Balanced Accuracy	0.7652	0.63882	0.60607	0.63483	0.8080

Ağaçlar için düğüm sayısı histogram ile elde edilmektedir.

```
> hist(treesize(rf1), main = "No. of Nodes for the Trees", col = "green")
```

Değişken önemliliği grafiği oluşturulmaktadır.

```
> varImpPlot(rf1, sort = T, main = "Variable Importance")
```

Kısmi bağımlılık grafikleri değişken ve sınıf bazında elde edilmektedir.

```
> partialPlot(rf1, train, EDU, "1")
```

```
> partialPlot(rf1, train, EDU, "5")
```

```
> partialPlot(rf1, train, OCC, "1")
```

```
> partialPlot(rf1, train, EDU, "5")
```

Modelden seçilen tek bir ağacın ağaç çıktısı Tablo B.14'te yer almaktadır.

```
> getTree(rf1, 1, labelVar = TRUE)
```

Çıktı:

Tablo B.14 Modelden seçilen tek bir ağaca ilişkin RO ağaç çıktısı

No	left daughter	daughter split	right daughter	split var	split point	prediction
1	2	3	OCC	55	1	<NA>
2	4	5	SEC	40	1	<NA>
3	6	7	AGE	3	1	<NA>
4	8	9	AGE	7	1	<NA>
5	10	11	SEC	64	1	<NA>
6	12	13	EDU	1	1	<NA>
7	14	15	AGE	24	1	<NA>
8	16	17	STA	5	1	<NA>
9	18	19	EDU	4	1	<NA>
10	20	21	MST	1	1	<NA>
11	22	23	STA	26	1	<NA>
12	24	25	OCC	8	1	<NA>
13	26	27	SEC	68	1	<NA>

14	28	29	STA	10	1	<NA>
15	30	31	RGN	62	1	<NA>
16	32	33	RGN	30	1	<NA>
17	34	35	EDU	12	1	<NA>
18	36	37	AGE	16	1	<NA>
19	38	39	EDU	26	1	<NA>
20	40	41	EDU	3	1	<NA>
21	42	43	OCC	5	1	<NA>
22	44	45	STA	8	1	<NA>
23	46	47	AGE	1	1	<NA>
24	48	49	MST	1	1	<NA>
25	50	51	RGN	47	1	<NA>
26	52	53	RGN	23	1	<NA>
27	54	55	MST	1	1	<NA>
28	56	57	EDU	1	1	<NA>
29	58	59	STA	16	1	<NA>
30	60	61	SEC	402	1	<NA>
31	62	63	EDU	17	1	<NA>
32	64	65	EDU	20	1	<NA>
33	66	67	EDU	11	1	<NA>
34	68	69	AGE	2	1	<NA>
35	70	71	RGN	47	1	<NA>

36	72	73	SEC	8	1	<NA>
37	74	75	SEC	8	1	<NA>
38	76	77	GEN	1	1	<NA>
39	78	79	INT	1	1	<NA>
40	80	81	OCC	5	1	<NA>
41	82	83	RGN	55	1	<NA>
42	84	85	EDU	2	1	<NA>
43	86	87	INT	1	1	<NA>
44	88	89	MST	1	1	<NA>
45	90	91	GEN	1	1	<NA>
46	92	93	SEC	385	1	<NA>
47	94	95	GEN	1	1	<NA>
48	96	97	SEC	402	1	<NA>
49	0	0	<NA>	0	-1	5_Quantile
50	98	99	AGE	1	1	<NA>
51	100	101	SEC	414	1	<NA>
52	102	103	OCC	8	1	<NA>
53	104	105	INT	1	1	<NA>
54	106	107	EDU	26	1	<NA>
55	108	109	RGN	2	1	<NA>
56	110	111	SEC	479	1	<NA>
57	112	113	MST	1	1	<NA>

58	114	115	OCC	8	1	<NA>
59	116	117	GEN	1	1	<NA>
60	118	119	MST	1	1	<NA>
61	120	121	OCC	8	1	<NA>
62	122	123	SEC	257	1	<NA>
63	124	125	STA	17	1	<NA>
64	126	127	RGN	12	1	<NA>
65	128	129	AGE	1	1	<NA>
66	130	131	GEN	1	1	<NA>
67	132	133	AGE	1	1	<NA>
68	134	135	OCC	1	1	<NA>
69	136	137	RGN	31	1	<NA>
70	138	139	AGE	2	1	<NA>
71	140	141	OCC	4	1	<NA>
72	142	143	INT	1	1	<NA>
73	144	145	RGN	40	1	<NA>
74	146	147	AGE	8	1	<NA>
75	148	149	MST	1	1	<NA>
76	150	151	OCC	16	1	<NA>
77	152	153	INT	1	1	<NA>
78	154	155	EDU	1	1	<NA>
79	156	157	EDU	1	1	<NA>

80	158	159	GEN	1	1	<NA>
81	160	161	GEN	1	1	<NA>
82	162	163	RGN	17	1	<NA>
83	164	165	INT	1	1	<NA>
84	166	167	OCC	1	1	<NA>
85	168	169	AGE	1	1	<NA>
86	170	171	RGN	16	1	<NA>
87	172	173	GEN	1	1	<NA>
88	174	175	INT	1	1	<NA>
89	176	177	RGN	53	1	<NA>
90	178	179	SEC	16	1	<NA>
91	180	181	STA	2	1	<NA>
92	182	183	GEN	1	1	<NA>
93	184	185	MST	1	1	<NA>
94	186	187	MST	1	1	<NA>
95	188	189	SEC	1	1	<NA>
96	190	191	SEC	128	1	<NA>
97	192	193	AGE	1	1	<NA>
98	194	195	INT	1	1	<NA>
99	196	197	SEC	386	1	<NA>
100	198	199	INT	1	1	<NA>
101	200	201	MST	1	1	<NA>

102	202	203	SEC	4	1	<NA>
103	204	205	EDU	16	1	<NA>
104	206	207	RGN	40	1	<NA>
105	208	209	OCC	8	1	<NA>
106	210	211	SEC	27	1	<NA>
107	212	213	STA	2	1	<NA>
108	214	215	SEC	138	1	<NA>
109	216	217	EDU	20	1	<NA>
110	218	219	SEC	18	1	<NA>
111	220	221	OCC	8	1	<NA>
112	222	223	AGE	8	1	<NA>
113	224	225	RGN	18	1	<NA>
114	226	227	RGN	30	1	<NA>
115	228	229	RGN	10	1	<NA>
116	0	0	<NA>	0	-1	2_Quantile
117	230	231	SEC	128	1	<NA>
118	232	233	EDU	1	1	<NA>
119	234	235	INT	1	1	<NA>
120	236	237	RGN	18	1	<NA>
121	238	239	MST	1	1	<NA>
122	240	241	GEN	1	1	<NA>
123	242	243	OCC	8	1	<NA>

124	244	245	OCC	8	1	<NA>
125	246	247	SEC	82	1	<NA>
126	248	249	MST	1	1	<NA>
127	250	251	MST	1	1	<NA>
128	252	253	EDU	8	1	<NA>
129	254	255	MST	1	1	<NA>
130	256	257	EDU	1	1	<NA>
131	258	259	MST	1	1	<NA>
132	260	261	GEN	1	1	<NA>
133	262	263	EDU	4	1	<NA>
134	0	0	<NA>	0	-1	4_Quantile
135	264	265	EDU	4	1	<NA>
136	266	267	OCC	5	1	<NA>
137	268	269	EDU	4	1	<NA>
138	270	271	OCC	1	1	<NA>
139	272	273	RGN	12	1	<NA>
140	274	275	AGE	2	1	<NA>
141	276	277	AGE	2	1	<NA>
142	278	279	MST	1	1	<NA>
143	0	0	<NA>	0	-1	2_Quantile
144	280	281	INT	1	1	<NA>
145	282	283	GEN	1	1	<NA>

146	0	0	<NA>	0	-1	2_Quantile
147	0	0	<NA>	0	-1	2_Quantile
148	284	285	RGN	55	1	<NA>
149	286	287	GEN	1	1	<NA>
150	288	289	INT	1	1	<NA>
151	290	291	MST	1	1	<NA>
152	292	293	RGN	58	1	<NA>
153	294	295	RGN	10	1	<NA>
154	296	297	MST	1	1	<NA>
155	298	299	RGN	39	1	<NA>
156	300	301	GEN	1	1	<NA>
157	302	303	RGN	2	1	<NA>
158	0	0	<NA>	0	-1	4_Quantile
159	304	305	INT	1	1	<NA>
160	306	307	EDU	1	1	<NA>
161	308	309	AGE	1	1	<NA>
162	310	311	OCC	1	1	<NA>
163	0	0	<NA>	0	-1	5_Quantile
164	312	313	OCC	1	1	<NA>
165	314	315	OCC	5	1	<NA>
166	0	0	<NA>	0	-1	5_Quantile

```

# Modele ait ilk dört ağaca ilişkin grafikler elde edilmektedir.

> library(randomForest)

> to.dendrogram <- function(dfrep,rownum=1,height.increment=0.1){
  if(dfrep[rownum,'status'] == -1){
    rval <- list()
    attr(rval,"members") <- 1
    attr(rval,"height") <- 0.0
    attr(rval,"label") <- dfrep[rownum,'prediction']
    attr(rval,"leaf") <- TRUE
  }
  else {
    left <- to.dendrogram(dfrep,dfrep[rownum,'left daughter'],height.increment)
    right <- to.dendrogram(dfrep,dfrep[rownum,'right
daughter'],height.increment)
    rval <- list(left,right)
    attr(rval,"members") <- attr(left,"members")+attr(right,"members")
    attr(rval,"height") <- max(attr(left,"height"),attr(right,"height")) +
height.increment
    attr(rval,"leaf") <- FALSE
    attr(rval,"edgetext") <- dfrep[rownum, 'split var']
  }
  class(rval) <- "dendrogram"
  return(rval)
}

## Aşağıdaki kod modele ait ilk dört ağacı görselleştirmektedir.

```

```
> for(i in 1:4) {  
  tree <- getTree(rf1,i,labelVar=T)  
  d <- to.dendrogram(tree)  
  plot(d,center=TRUE,leaflab='none',edgePar=list(t.cex=1,p.col=NA,p.lty=0))  
}
```



B.5 Boruta Algoritması ile Değişken Seçim İşlemi ve R Kodu

Birden fazla çeşit kategoride olan yaş ve sektör değişkenlerini azaltmak için Boruta algoritması ile değişken seçimi yapılmaktadır.

```
> set.seed(123)
```

```
> boruta.Merge_15 <- Boruta(IG~. ,data = Merge_15, doTrace = 2)
```

Çıktılar:

1. run of importance source...

Growing trees.. Progress: 70%. Estimated remaining time: 13 seconds.

Computing permutation importance.. Progress: 34%. Estimated remaining time: 59 seconds.

Computing permutation importance.. Progress: 69%. Estimated remaining time: 27 seconds.

2. run of importance source...

Growing trees.. Progress: 77%. Estimated remaining time: 9 seconds.

Computing permutation importance.. Progress: 36%. Estimated remaining time: 55 seconds.

Computing permutation importance.. Progress: 72%. Estimated remaining time: 24 seconds.

3. run of importance source...

Growing trees.. Progress: 78%. Estimated remaining time: 8 seconds.

Computing permutation importance.. Progress: 36%. Estimated remaining time: 55 seconds.

Computing permutation importance.. Progress: 72%. Estimated remaining time: 23 seconds.

4. run of importance source...

Growing trees.. Progress: 77%. Estimated remaining time: 9 seconds.

Computing permutation importance.. Progress: 36%. Estimated remaining time: 55 seconds.

Computing permutation importance.. Progress: 72%. Estimated remaining time: 23 seconds.

5. run of importance source...

Growing trees.. Progress: 77%. Estimated remaining time: 9 seconds.

Computing permutation importance.. Progress: 36%. Estimated remaining time: 55 seconds.

Computing permutation importance.. Progress: 73%. Estimated remaining time: 23 seconds.

6. run of importance source...

Growing trees.. Progress: 78%. Estimated remaining time: 8 seconds.

Computing permutation importance.. Progress: 36%. Estimated remaining time: 55 seconds.

Computing permutation importance.. Progress: 72%. Estimated remaining time: 23 seconds.

7. run of importance source...

Growing trees.. Progress: 77%. Estimated remaining time: 9 seconds.

Computing permutation importance.. Progress: 36%. Estimated remaining time: 55 seconds.

Computing permutation importance.. Progress: 71%. Estimated remaining time: 25 seconds.

8. run of importance source...

Growing trees.. Progress: 77%. Estimated remaining time: 9 seconds.

Computing permutation importance.. Progress: 36%. Estimated remaining time: 55 seconds.

Computing permutation importance.. Progress: 73%. Estimated remaining time: 23 seconds.

9. run of importance source...

Growing trees.. Progress: 77%. Estimated remaining time: 9 seconds.

Computing permutation importance.. Progress: 36%. Estimated remaining time: 54 seconds.

Computing permutation importance.. Progress: 73%. Estimated remaining time: 23 seconds.

10. run of importance source...

Growing trees.. Progress: 77%. Estimated remaining time: 9 seconds.

Computing permutation importance.. Progress: 36%. Estimated remaining time: 55 seconds.

Computing permutation importance.. Progress: 72%. Estimated remaining time: 23 seconds.

11. run of importance source...

Growing trees.. Progress: 77%. Estimated remaining time: 9 seconds.

Computing permutation importance.. Progress: 36%. Estimated remaining time: 55 seconds.

Computing permutation importance.. Progress: 72%. Estimated remaining time: 23 seconds.

After 11 iterations, +25 mins:

confirmed 19 attributes: AGE11, AGE3, AGE6, EDU, EMP and 14 more;

no more attributes left.

Önemli ve önemsiz değişkenlerin hangileri olduğu belirlenmektedir.

```
> print(boruta.Merge_15)
```

Çıktılar:

Boruta performed 11 iterations in 24.77739 mins.

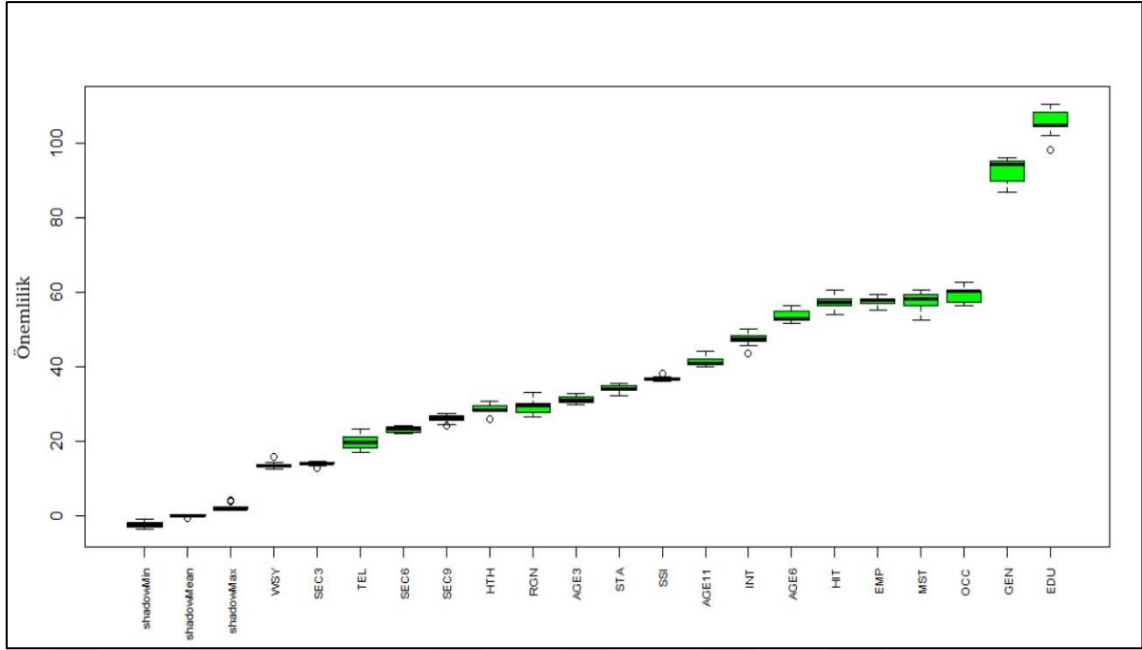
19 attributes confirmed important: AGE11, AGE3, AGE6, EDU, EMP and 14 more;

No attributes deemed unimportant.

Değişken önemliliklerine ilişkin kutu grafiklerini gösteren grafik Şekil B.1'de oluşturulmaktadır.

```
> plot(boruta.Merge_15, xlab = "", xaxt = "n")
> lz<-lapply(1:ncol(boruta.Merge_15$ImpHistory),function(i)
+   boruta.Merge_15$ImpHistory[is.finite(boruta.Merge_15$ImpHistory[,i]),i])
> names(lz) <- colnames(boruta.Merge_15$ImpHistory)
> Labels <- sort(sapply(lz,median))
> axis(side = 1,las=2,labels = names(Labels),
+   at = 1:ncol(boruta.Merge_15$ImpHistory), cex.axis = 0.7)
```

Çıktılar:



Şekil B.1 Boruta ile elde edilen değişken önemliliği grafiği

Değişken önemliliği grafiğinde onaylanmış değişkenleri gösteren yeşil renkteki kutu grafiklerine bakıldığında tüm değişkenlerin onaylanmış olduğu görülmektedir. Boruta ile değişken seçimi işlemi öncesi SEC3, SEC6 ve SEC9 değişkenlerindeki eksik değerli hücreler “Sektörü Yok” olarak kodlanmış olup bu değişkenlerin isimleri sonradan SEC4, SEC7 ve SEC10 olarak değiştirilmiştir.

Onaylanmış değişkenlerin listesi elde edilmektedir.

```
> getSelectedAttributes(boruta.Merge_15, withTentative = FALSE)
```

Çıktılar:

```
[1] "MST" "EDU" "HTH" "EMP" "STA" "OCC" "SEC3" "SEC9" "SEC6"
```

```
[10] "SSI" "AGE11" "AGE6" "AGE3" "GEN" "RGN" "WSY" "TEL" "INT"
```

```
[19] "HIT"
```

Değişkenler önem düzeyine göre sıralanmaktadır. Onaylanmış değişkenlere ilişkin istatistikler Tablo B.16’da yer almaktadır.

```
> boruta.df <- attStats(boruta.Merge_15)
```

```
> class(boruta.df)
```

```
> df_table <- print(boruta.df)

> df_table[order(df_table$maxImp),]
```

Çıktılar:

Tablo B.15 Boruta ile elde edilen onaylanmış değişkenlere ilişkin istatistikler

	meanImp	medianImp	minImp	maxIm	normHits	decision
SEC3	13.97153	14.1366	12.94411	14.63551	1	Confirmed
WSY	13.55983	13.37455	12.39134	15.89243	1	Confirmed
TEL	19.763	19.69685	16.99834	23.13669	1	Confirmed
SEC6	23.06807	23.1888	22.17021	24.15883	1	Confirmed
SEC9	26.06066	26.16171	24.11078	27.53114	1	Confirmed
HTH	28.59498	28.39261	25.84496	30.73196	1	Confirmed
AGE3	31.16645	31.04675	29.88173	32.6455	1	Confirmed
RGN	29.28226	29.41586	26.4831	33.13706	1	Confirmed
STA	33.98375	33.87083	32.09149	35.40021	1	Confirmed
SSI	36.8482	36.63701	36.04098	38.2333	1	Confirmed
AGE11	41.50594	40.8106	40.07315	44.18452	1	Confirmed
INT	47.45302	47.54062	43.6231	49.97149	1	Confirmed
AGE6	53.63267	52.80545	51.44063	56.42523	1	Confirmed
EMP	57.42711	57.90291	55.03696	59.39932	1	Confirmed
HIT	57.31493	57.28227	53.83574	60.4666	1	Confirmed
MST	57.6943	58.27748	52.51592	60.48251	1	Confirmed
OCC	59.26978	60.24813	56.23342	62.62175	1	Confirmed

GEN	92.50264	94.10256	86.87323	95.88997	1	Confirmed
EDU	105.4372	104.68071	98.04286	110.4927	1	Confirmed

Boruta paketinde standart değeri 100 olan maxRuns parametre değeri 200 yapıldığında da aynı sonuçlar elde edilmektedir. Boruta ile değişken seçim işlemi sonrası değişkenlerin tümü önemli bulunmaktadır. Sonuçlarına göre, yaş değişkenlerinden AGE3 ve AGE11, sektör değişkeninden ise SEC6 ve SEC3 değişkenleri veri setinden çıkarılmaktadır.



B.6 LCAvarsel Algoritması ile Değişken Seçim İşlemi ve R Kodu

GSA öncesi değişkenleri belirlemek amacıyla LCAvarsel ile değişken seçimi gerçekleştirilmektedir.

```
> sel_LCAvarsel_Data_ga <- LCAvarsel(Data, search = "ga", parallel = TRUE)
```

Arama yöntemi olarak genetik algoritma kullanıldığında aşağıdaki sonuç elde edilmektedir.

Çıktı:

GA search | Iter = 10 Mean = -940178.9 | Best = -930421.3

Clustering set: STA OCC WSY INT HIT EHSIZE_4 AGE_6 GEN

GA search | Iter = 20 Mean = -952695.0 | Best = -927064.6

Clustering set: STA OCC RGN WSY INT AGE_6 GEN

GA search | Iter = 30 Mean = -966728.1 | Best = -926783.6

Clustering set: STA OCC RGN INT AGE_6 GEN

GA search | Iter = 40 Mean = -938738.7 | Best = -926783.6

Clustering set: STA OCC RGN INT AGE_6 GEN

GA search | Iter = 50 Mean = -937143.6 | Best = -926268.6

Clustering set: STA OCC RGN AGE_6 GEN

GA search | Iter = 60 Mean = -941328.5 | Best = -926268.6

Clustering set: STA OCC RGN AGE_6 GEN

GA search | Iter = 70 Mean = -935953.5 | Best = -926268.6

Clustering set: STA OCC RGN AGE_6 GEN

GA search | Iter = 80 Mean = -938023.7 | Best = -926268.6

Clustering set: STA OCC RGN AGE_6 GEN

GA search | Iter = 90 Mean = -929206.3 | Best = -926268.6

Clustering set: STA OCC RGN AGE_6 GEN

GA search | Iter = 100 Mean = -946474.9 | Best = -926268.6

Clustering set: STA OCC RGN AGE_6 GEN

100 iterasyon sonunda genetik algoritma yoluyla oluşan değişken seti: STA
OCC RGN AGE_6 GEN

```
> sel_LCAvarsel_Data_swap <- LCAvarsel(Data, search = "backward", swap =  
TRUE, parallel = TRUE)
```

Arama yöntemi olarak swap algoritması kullanıldığında Tablo B.17'deki
sonuçlar elde edilmektedir.

Çıktı:

Starting clustering set:

all variables

Tablo B.16 LCAvarsel ile değişken seçim işleminde swap algoritması sonuçları

Step	Variable	BICdiff	Decision
1 Remove	AGE_6	29503.46	Accepted
2 Remove	SEC_9	25599.45	Accepted
3 Swap	OCC-SEC_9	-3994.5	Rejected
4 Add	AGE_6	-29315.99	Rejected
5 Swap	EDU-AGE_6	-2450.036	Rejected
6 Remove	IG	15702.03	Accepted
7 Swap	EDU-IG	-784.7666	Rejected
8 Add	SEC_9	-23829.35	Rejected
9 Swap	OCC-SEC_9	-1801.872	Rejected

10 Remove	EDU	12196.66	Accepted
11 Swap	EMP-EDU	-1996.099	Rejected
12 Add	IG	-12787.44	Rejected
13 Swap	OCC-IG	1683.052	Accepted
14 Remove	EMP	10463.83	Accepted
15 Swap	STA-EDU	31.52865	Accepted
16 Add	STA	-9888.687	Rejected
17 Swap	EDU-STA	-31.5283	Rejected
18 Remove	EDU	8481.277	Accepted
19 Swap	IG-EDU	-1883.869	Rejected
20 Add	STA	-8301.017	Rejected
21 Swap	HIT-STA	-1163.275	Rejected
22 Remove	IG	4729.986	Accepted
23 Swap	INT-IG	-810.4044	Rejected
24 Add	EDU	-6764.405	Rejected
25 Swap	RGN-EDU	-2607.975	Rejected
26 Remove	MST	2726.921	Accepted
27 Swap	GEN-STA	-626.2798	Rejected
28 Add	AGE_6	-3598.219	Rejected
29 Swap	HIT-AGE_6	-920.3341	Rejected
30 Remove	GEN	291.6303	Accepted
31 Swap	RGN-GEN	-79.27958	Rejected

32 Add	MST	-894.3554	Rejected
33 Swap	INT-MST	-58.7295	Rejected
34 Remove	RGN	62.78139	Accepted
35 Swap	WSY-RGN	-41.35084	Rejected
36 Add	GEN	-189.8782	Rejected
37 Swap	WSY-GEN	-219.9194	Rejected
38 Remove	WSY	-24.13852	Rejected
39 Swap	WSY-RGN	-41.35083	Rejected
40 Add	RGN	-62.7814	Rejected
41 Swap	WSY-RGN	-41.35082	Rejected

swap algoritması yoluyla oluşan değişken seti: WSY, INT, HIT, EHSIZE_4

```
> sel_LCAvarsel_Data <- LCAvarsel(Data, search = "backward", parallel = TRUE)
```

Arama yöntemi olarak stepwise algoritması kullanıldığında aşağıdaki sonuç elde edilmektedir.

Çıktı:

Starting clustering set:

all variables

Tablo B.17 LCAvarsel ile değişken seçim işleminde stepwise algoritması sonuçları

Step	Variable	BICdiff	Decision
1 Remove	AGE_6	29503.46	Accepted

2 Remove	SEC_9	25599.45	Accepted
3 Add	AGE_6	-29315.99	Rejected
4 Remove	IG	15702.03	Accepted
5 Add	SEC_9	-23829.35	Rejected
6 Remove	EDU	12196.66	Accepted
7 Add	IG	-12787.44	Rejected
8 Remove	OCC	10393.14	Accepted
9 Add	IG	-8744.337	Rejected
10 Remove	EMP	6831.195	Accepted
11 Add	IG	-5541.013	Rejected
12 Remove	STA	7363.898	Accepted
13 Add	IG	-4729.983	Rejected
14 Remove	MST	2726.921	Accepted
15 Add	AGE_6	-3598.219	Rejected
16 Remove	GEN	291.6304	Accepted
17 Add	MST	-894.3554	Rejected
18 Remove	RGN	62.78672	Accepted
19 Add	GEN	-189.8829	Rejected
20 Remove	WSY	-24.14496	Rejected
21 Add	RGN	-62.78673	Rejected

stepwise algoritması yoluyla oluşan değişken seti: WSY, INT, HIT, EHSIZE_4

TEZDEN ÜRETİLMİŞ YAYINLAR

İletişim Bilgisi: olgunozdemir35@gmail.com

Makaleler

1. O. Özdemir ve İ. Demir, “Data mining of SILC data: Turkey case,” International Journal of Sciences: Basic and Applied Research, vol. 48(7), pp. 110-138, 2019.

Konferans Bildirileri

1. O. Özdemir ve İ. Demir, “Data mining of SILC data: Turkey case”, 11. International Statistics Congress, ISC2019, Muğla, Türkiye, pp. 237, 4-8 Ekim 2019.