Çankırı Karatekin Üniversitesi
İktisadi ve İdari Bilimler
Fakültesi Dergisi
Y. 2018, Cilt 8, Sayı 2, ss. 91-114

Cankırı Karatekin University
Journal of the Faculty of Economics
and Administrative Sciences
Y. 2018, Volume 8, Issue 2, pp. 91-114

Türkiye İşgücü Verilerinin Karar Ağacı Yöntemleriyle Analizi[*†]

Engin YILDIZTEPE
Sorumlu Yazar: Dokuz Eylül Üniversitesi, İstatistik Bölümü,
engin.yildiztepe@deu.edu.tr

Ahmet KOCATAŞ
TÜİK İzmir Bölge Müdürlüğü,
ahmet.kocatas@tuik.gov.tr

Öz

İstihdam, işsizlik ve bunlara bağlı sorunlara odaklanan ülkeler, bu sorunları çözebilmek ve sağlıklı politikalar oluşturabilmek için konu ile ilgili veri derleme ve araştırma çalışmalarına önem vermektedirler. Bu konudaki araştırmalarda çeşitli veri analiz yöntemlerinden faydalanılmaktadır. Klasik istatistiksel analiz yöntemlerinin yanı sıra, elde edilen veri miktarının artması nedeniyle, büyük veri tabanları üzerinde en etkili analiz yöntemlerini içinde barındıran veri madenciliği yöntemleri de yaygın olarak kullanılmaktadır. Bu çalışmada, Türkiye İstatistik Kurumu tarafından yapılan "Hanehalkı İşgücü Araştırması" 2013 yılı verilerini kullanarak, nüfusun işgücü açısından durumunu ortaya koymak ve işgücü durumu için karar ağacı modelleri oluşturmak amaçlanmıştır. Bu amaçla, işgücündeki nüfus için istatistikler elde edilmiş ve bir önceki yıl çalışmayan bireylere ait veriler kullanılarak karar ağacı modelleri oluşturulmuştur. Bir yıl önce çalışmadığını belirten bireylerin cinsiyetleri, yaş grupları, mezuniyet durumları, medeni durumları, bir okula devam edip etmedikleri ve ikamet ettikleri coğrafi bölge işgücü durumlarını etkileyen en önemli değişkenler olarak bulunmuştur.

Anahtar Kelimeler: *Hanehalkı İşgücü Araştırması, İşgücü, İstihdam, İşsizlik, Karar Ağaçları, Veri Madenciliği*

JEL Sınıflandırma Kodları: J00, J21, C38

Analysis of Turkey Labour Force Data by Decision Tree Methods[‡]

Abstract

The countries which have focused on employment, unemployment and the problems related to these issues attach importance to gathering data and research efforts about these issues in order to solve these problems and create efficient policies. Various data analysis methods are utilized in these researches. In addition to classical statistical data analysis methods, data mining methods which contain the most effective analysis methods on large databases are widely used because of the increase in the amount of data obtained. In this study, it is aimed to present the situation of labour force of the population and to provide detailed information by using the year 2013 data of "Household Labour Force Survey" conducted by Turkish Statistical Institute. For this purpose, statistics were obtained for the population in the labour force, and decision tree models were created using data that belong to the persons whose labour force situation was "not working" at the previous year. Gender, age groups, graduation status, marital status, attending a school and the geographical region where they reside were found to be the most important variables affecting the labour force situation.

Keywords: *Household Labour Force Survey, Labour Force, Employment, Unemployment, Decision Trees, Data Mining*
JEL Classification Codes: J00, J21, C38

[*] Çalışmada elde edilen bulgular ve değerlendirmeler yazarların kişisel görüşleri olup Türkiye İstatistik Kurumu'nu bağlamamaktadır.
[†] Çalışma, yazarlardan Ahmet Kocataş'ın TÜİK Uzmanlık tez çalışması kapsamında yapılan araştırmalardan üretilmiştir.
[‡] Extended abstract is presented at the end of the article.

Atıfta bulunmak için/Cite this paper:
Yıldıztepe, E. ve Kocataş, A. (2018). Türkiye işgücü verilerinin karar ağacı yöntemleriyle analizi. *Çankırı Karatekin Üniversitesi İİBF Dergisi*. 8 (2), 91-114.

1. Giriş

Toplumların gelişmişlik düzeylerine bakılmaksızın hemen hemen her ülkede, işsizlikle ve neden olduğu sosyal ve ekonomik sorunlarla karşılaşılmaktadır. Nüfus artışı, otomasyona dayalı üretimlerin artması ve işgücünün etkin kullanılamaması gibi temel nedenlerden dolayı, istihdamda istenilen düzeye ulaşılamamaktadır. Ülkemizde işsizlik sorunu ciddiyetini korumakta ve sürekli olarak sebepleri araştırılmaktadır. Politika belirleyiciler de işsizlik sorununu azaltmak için çeşitli sosyal ve ekonomik önlemler almaktadırlar. Ancak, çok çeşitli etmenlerin etkisinde kalan işgücü piyasası zaman içerisinde değişiklik göstermektedir. Küresel çapta yaşanan ekonomik sıkıntılar ve sınır komşusu olduğumuz ülkelerde yıllarca yaşanan iç karışıklar nedeniyle o ülkelerden gelen göçmenler işgücünde ciddi değişikliğin yaşanmasına neden olabilmektedir.

Ülkemizde işgücü piyasasını sürekli olarak takip etmek amacıyla, Türkiye İstatistik Kurumu (TÜİK) tarafından, aylık olarak Hanehalkı İşgücü Araştırması (HİA) gerçekleştirilmektedir. Tüm ülke genelini yansıtacak belli bir örneklemin, örnekleme yöntemleriyle seçilmesiyle uygulanan anket çalışması ile ülke genelinin işgücü rakamları ortaya koyulmaya çalışılmaktadır. HİA ile, istihdam edilenlerin; ekonomik faaliyet, meslek, işteki durum ve çalışma süresi; işsizlerin ise, iş arama süresi, iş aranan meslek ve benzer özellikleri hakkında veriler derlenmektedir. TÜİK topladığı verilerle elde edilen işgücü istatistiklerini içeren haber bültenlerini web sitesinde yayınlamaktadır. Ayrıca sınırlı sayıda değişken için çeşitli özet istatistiklerin sorgulanmasına imkân sağlamaktadır (http://www.tuik.gov.tr/PreTabloArama.do?metod=search&araType=vt). Ancak TÜİK HİA kapsamında elde edilen ham verileri web sitesinden yayınlamamaktadır. B grubu mikro veri kapsamında olan bu veriler, istenilen izinler alınarak, kurumdan talep edilebilir (http://www.tuik.gov.tr/UstMenu.do?metod=bilgiTalebi). Bu veriler üzerinde yapılan istatistiksel değerlendirmeler, ekonomik ve sosyal politikaların yönleri konusunda yardımcı olabilir. Bu veriler ile kurulan modeller ise planlanan politikaların meydana getireceği değişimi tahmin etmede kullanılabilir.

Bu çalışmanın amacı, HİA verilerini kullanarak işgücünün yapısı ile ilgili ayrıntılı ve açıklayıcı sonuçlara ulaşmak ve HİA'da yer alan "işgücü durumu" değişkenini hedef değişken olarak belirleyerek karar ağaçları yöntemleri ile işgücü yapısı için model geliştirmektir. Çalışmanın, HİA'yı ve içerdiği verileri tanıtmada katkı sağlayacağı ayrıca TÜİK tarafından sağlanan verilerle yapılacak benzer çalışmalar için yol gösterici olacağı düşünülmektedir. Araştırmada Microsoft Excel 2013 ve IBM SPSS Modeler 16.0 yazılımları kullanılmıştır. Çalışmanın ikinci bölümünde literatür özetine yer verilmiştir. Üçüncü bölümde HİA hakkında detaylı bilgi verilmiştir. Dördüncü bölümde çalışmada kullanılan yöntemler özetlenmiş ve verilerin seçimi, önişleme ve dönüştürme aşamalarından bahsedilmiştir. Beşinci

bölümde araştırmanın bulgularına yer verilmiştir. Son bölümde elde edilen sonuçlar değerlendirilmiştir.

2. Literatür Özeti

TÜİK tarafından her ay sürekli olarak gerçekleştirilen HİA ilk defa 1966 yılında sanayisi gelişmiş sekiz ilde deneme niteliğinde yapılmıştır. 1987 yılına kadar HİA düzensiz aralıklarla uygulanmış olup 1988 yılının Ekim ayından itibaren Uluslararası Çalışma Örgütü tarafından belirlenen tanım ve kavramlar çerçevesinde yılda iki kez düzenli olarak uygulanmaya başlanmıştır. 2000 yılından itibaren her ay düzenli olarak uygulanmaktadır (TÜİK, 2014b, s. 174).

TÜİK, HİA sonucunda elde ettiği istatistikleri, ayrıca araştırmaya ilişkin çerçeve, tanımlar ve sınıflandırmaları, web sitesinde özet tablolar halinde yayınlamaktadır. Literatürdeki çalışmalar incelendiğinde bu özet tabloların çok sayıda çalışmada kullanıldığı görülmektedir. Tansel, kadınların işgücüne katılımı ve ekonomik gelişmeler ile ilgili çalışmasında 2000 yılı HİA istatistiklerinden yararlanmıştır (Tansel, 2001). Tunalı ve Ercan, 2003 yılındaki Türkiye işgücü piyasası hakkındaki çalışmalarında ana veri kaynağı olarak 1988-2000 yılları arasında yapılan HİA'ları kullandıklarını belirtmişlerdir (Tunalı & Ercan, 2003). Özkan, işgücü piyasası hareketliliğinin modellenmesi ile ilgili çalışmasında 2000-2002 yıllarına ait HİA bulgularından yararlanmıştır (Özkan, 2013). TÜSİAD tarafından yayımlanan bir başka çalışmada 2002-2006 yıllarına ait HİA sonuçları kullanılmıştır (Yükseler & Türkan, 2008).

HİA'da kullanılan soru formu ve mikro veriler ise ancak kurumdan alınan izin ile akademik amaçlı kullanılabilmektedir. Bu çalışmalar daha az sayıdadır ve bazıları işgücü, istihdam konuları dışındadır. Örneğin, Türkiye'deki ücret eğrisi ile ilgili bir çalışmada 2005-2008 yılları arasındaki HİA mikro verileri kullanılmıştır (Baltagi, Baskaya, & Hulagub, 2012). TÜİK HİA mikro verilerinin sınıflama amacıyla kullanıldığı çalışmalara örnek olarak Oğuzlar (2004) ve Yılmaz (2012) verilebilir. Oğuzlar çalışmasında, karar ağacı algoritmalarından biri olan sınıflandırma ve regresyon ağacı (Classification and Regression Tree – CART) algoritmasını kullanarak 2002 yılı III. Dönem Hanehalkı İşgücü Anketi mikro verilerini analiz etmiştir (Oğuzlar, 2004). Oğuzlar, araştırmasında iş arama durumunu yanıt değişkeni olarak belirlemiş, modelde cinsiyet, yaş, hanehalkı reisine yakınlık, en son bitirilen okul, medeni durum değişkenlerini kullanmıştır. Yılmaz, çalışmasında 2009 ve 2010 yılları HİA verilerine Chi-squared Automatic Interaction Detection (CHAID) algoritmasını uygulamış ve işgücü durumu yanıt değişkeni için elde ettiği karar ağacı modellerini karşılaştırmıştır (Yılmaz, 2012). Yılmaz, karar ağacı modellerini, cinsiyet, referans kişiye yakınlık durumu, en son bitirilen okul, bir öğrenim kurumuna devam etme durumu, medeni durum, bir yıl önceki işgücü durumu ve istatistiki bölge birimleri sınıflaması değişkenlerini kullanarak oluşturmuştur.

3. Hanehalkı İşgücü Araştırması

HİA'nın amacı ülkedeki işgücünün yapısını ortaya koymaktır. Bu amaçla, istihdam edilen kişilerin meslek, çalışma süresi, istihdam edilemeyenlerin ise iş arama süresi ve aradıkları iş ve benzer özellikleri hakkında veriler toplanır. Araştırmada, ülke genelindeki tüm yerleşim yerleri örnek seçimi için kapsama dâhil edilmektedir ve ülke sınırları içinde yaşayan hanelerde bulunan tüm kişiler kapsanmaktadır. Okul, yurt, otel, çocuk yuvası, huzurevi, hastane, hapishane, kışla ve orduevleri gibi kurumsal yerlerde ikamet edenler araştırmaya dâhil edilmemektedir. Araştırmada, 20001 ve daha fazla nüfuslu yerleşim yerleri kent, 20000 ve daha az nüfuslu yerleşim yerleri ise kır olarak tanımlanmıştır. Araştırma, örnekleme yöntemi ile gerçekleştirilen bir çalışmadır ve her ay alan uygulaması yapılacak şekilde üçer aylık hareketli ortalamalar üzerinden dönemlik ve yıllık tahminlere ulaşmak üzere tasarlanmıştır. Araştırma verilerinin ağırlıklandırması en güncel nüfus kestirimlerine göre yapılmaktadır. Araştırma sonuçları belirlenen ağırlıklar kullanılarak güncel nüfus projeksiyonlarına göre yayımlanmaktadır. Araştırmanın örnekleme birimi adres, gözlem birimi ise hanehalkıdır. Hanehalkı, aralarındaki akrabalık bağına bakılmaksızın, aynı konutta veya aynı konutun bir bölümünde yaşayan bir ya da daha çok kişinin oluşturduğu topluluktur. Veriler seçilen konutlarda yaşayan hanehalklarından derlendiğinden kullanılan istatistiki birim "hanehalkları"dır. İşgücü durumunun tespitine yönelik sorular 15 ve daha yukarı yaştaki kişilere sorulmaktadır ancak hanehalkında bulunan tüm bireylere ilişkin demografik veriler toplanmaktadır (TÜİK, 2014a).

HİA'da kullanılan bazı tanım ve kavramlar TÜİK, HİA 2013'te (TÜİK, 2014a) ve Türkiye İstatistik Yıllığı'nda (TÜİK, 2014b, s. 175-179) aşağıdaki gibi açıklanmaktadır:

- **Kurumsal Olmayan Nüfus:** Okul, yurt, otel, çocuk yuvası, huzurevi, hastane, hapishane, kışla ya da orduevinde ikamet edenler dışında kalan nüfustur.
- **Çalışma Çağındaki Nüfus:** Kurumsal olmayan nüfustaki 15 ve daha yukarı yaştaki nüfustur.
- **İşbaşında Olanlar:** Ücretli, maaşlı, yevmiyeli, kendi hesabına, işveren ya da ücretsiz aile işçisi olarak referans dönemi içerisinde en az bir saat bir iktisadi faaliyette bulunan kişilerdir.
- **İşbaşında Olmayanlar:** İşi ile bağlantısı devam ettiği halde, araştırma haftası içerisinde çeşitli nedenlerle işinin basında olmayan kendi hesabına veya işveren olarak çalışan kişilerdir. Bu kişiler istihdamda kabul edilir.
- **İstihdam Edilenler:** Çalışma çağında, işbaşında olan veya işbaşında olmayan tüm nüfus istihdam edilen nüfustur.
- **İşsiz:** Araştırma dönemi içinde istihdam edilenler dışındaki çalışma çağındaki kişilerden iş aramak için son üç ay içinde iş arama kanallarından en az birini kullanmış ve iki hafta içinde işbaşı yapabilecek durumda olan tüm kişilerdir.

Bu kişilere ek olarak, üç ay içinde başlayabileceği bir iş bulmuş ya da kendi işini kurmuş ancak işe başlamak için çeşitli eksikliklerini tamamlamayı bekleyenler de işsiz sayılırlar.

- **İşgücü:** İstihdam edilen nüfus ile işsiz nüfusun toplamından oluşur.
- **İşgücüne Katılım Oranı:** İşgücünün, çalışma çağındaki nüfusa oranıdır.
- **İşsizlik Oranı:** İşsiz nüfusun işgücüne oranıdır.
- **İşgücüne Dâhil Olmayanlar:** İşsiz veya istihdamda bulunmayan çalışma çağındaki nüfustur. Aşağıdaki gruplara ayrılmıştır.
- **İki hafta içinde işbaşı yapmaya hazır olanlar:**
- **İş aramayıp çalışmaya hazır olanlar:** Çeşitli nedenlerle iş aramayan, ancak iki hafta içinde işbaşı yapmaya hazır olduğunu belirten kişilerdir. Daha önce iş aradığı halde bulamayan veya kendi vasıflarına uygun bir iş bulabileceğine inanmadığı için iş aramayan kişiler ile diğer nedenlerle (mevsimlik çalışma, ev hanımı, öğrenci, gelir sahibi, emekli, çalışamaz halde) iş aramayan kişiler bu gruptadır.
- **İş başı yapmaya hazır olmayanlar:**
- **Mevsimlik çalışanlar:** Mevsimlik çalışması nedeniyle iş aramayan kişilerdir.
- **Ev işleriyle meşgul:** Kendi ev işleri nedeniyle iş aramayan kişilerdir.
- **Eğitim/Öğretime devam ediyor:** Bir öğrenim kurumuna, kursa vb. devam etmesi nedeniyle iş aramayan kişilerdir.
- **Emekli:** Bir sosyal güvenlik kuruluşundan emekli olduğu için iş aramayan kişilerdir.
- **Çalışamaz halde:** Bedensel özür, hastalık veya yaşlılık nedeniyle iş aramayan kişilerdir.
- **Diğer:** Ailevi, kişisel veya bunların dışındaki diğer nedenler ile iş aramayan kişilerdir.

HİA'da kullanılan anket iki formdan oluşmaktadır. İlk formda hanehalkındaki tüm bireylerin demografik özellikleri sorulmaktadır. İkinci formda, hanedeki 15 yaş ve üzeri bireylerin işgücü durumunu ölçmek için sorular yer almaktadır. 2013 yılı HİA soru kâğıdında ikinci formdaki sorular altı bölüme ayrılmıştır (TÜİK, 2014a);

1. Hanehalkı fertlerinin kişisel nitelikleri
2. İstihdam ile ilgili sorular
3. Gelir bilgileri
4. İşsizlik ve iktisadi faal olmama ile ilgili sorular
5. Geçmişteki iş deneyimi ile ilgili sorular
6. Bir yıl önceki işgücü durumu

TÜİK, HİA'da bireylerin eğitim durumu, işyerlerinin ekonomik faaliyetleri, meslek durumları ve işteki durumları için çeşitli sınıflama bilgileri kullanmaktadır. Araştırmada, kişilerin eğitim durumları altı ana grupta ele alınmaktadır.

1. Bir okul bitirmeyen
2. İlkokul (5 yıl)
3. Ortaokul, mesleki ortaokul ve ilköğretim (8 yıl)
4. Genel lise
5. Mesleki veya teknik lise
6. Yüksekokul, fakülte ve üzeri

Araştırmada kuruluş veya işyerinde yapılan iş, Uluslararası Standart Meslek Sınıflamasına (ISCO-08) göre belirlenen 40 kategoriye göre sorulmaktadır. Bazı sorularda bu kategoriler Tablo 1'de belirtildiği gibi dokuz ana grupta ifade edilmektedir.

Tablo 1: Uluslararası Standart Meslek Sınıflaması

Grup	Tanım
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 ürünleri çalışanları
7	Sanatkârlar ve ilgili işlerde çalışanlar
8	Tesis ve makine operatörleri ve montajcılar
9	Nitelik gerektirmeyen işlerde çalışanlar

HİA'da çalışılan kuruluş, işyerinin ana faaliyeti, Avrupa Topluluğunda Ekonomik Faaliyetlerin İstatistiki Sınıflaması'na (NACE Rev.2) göre 87 kategoride sorulmaktadır. Tablo 2'de bu faaliyet alanları on ana grupta özetlenmiştir.

Tablo 2: Ekonomik Faaliyet Alanları

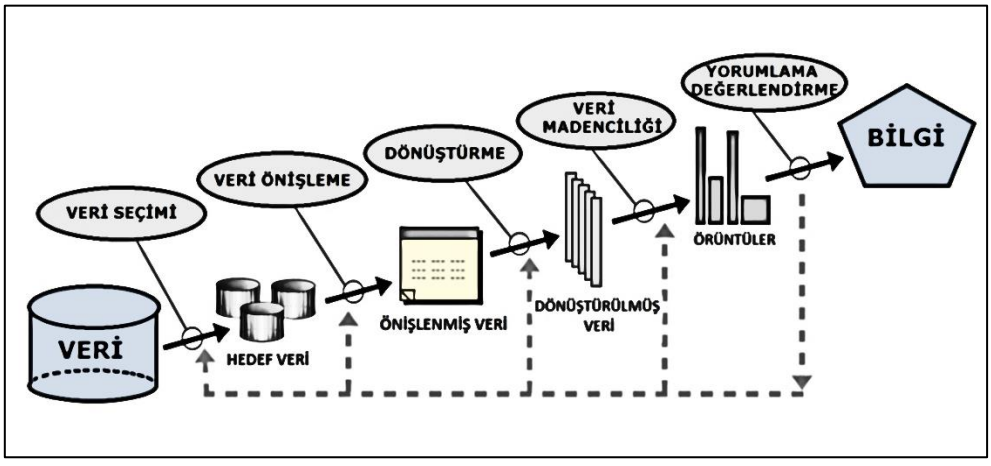
Alan	Tanım
1	Tarım
2	İmalat sanayi, madencilik ve taş ocakçılığı ve diğer sanayi
3	İnşaat
4	Toptan ve perakende ticaret, ulaştırma ve depolama, konaklama ve yiyecek hizmetleri
5	Bilgi ve iletişim
6	Finans ve sigorta faaliyetleri
7	Gayrimenkul faaliyetleri
8	Mesleki, bilimsel ve teknik faaliyetler, idari ve destek hizmetleri
9	Kamu yönetimi ve savunma, eğitim hizmetleri, insan sağlığı ve sosyal hizmetler
10	Diğer hizmet faaliyetleri

4. Yöntem ve Veriler

Çalışmanın bu bölümünde, veri madenciliği süreci ve kullanılan karar ağaçları yöntemlerine ilişkin bilgi verilecektir. Ayrıca uygulamada kullanılan mikro veriler ile ilgili tanımlamalar yapılacak, veri önişleme ve dönüşüm süreçleri anlatılacaktır.

4.1. Veri Madenciliği

Veri madenciliği, bir dizi istatistiksel yöntemin veri tabanlarında anlamlı ve faydalı bilgi keşfi amacıyla kullanılması sürecini tanımlayan, 1990'lı yıllarda ortaya atılan ve 2000'li yıllarda popüler olan bir kavramdır. Veri madenciliği bir yöntemi değil bir süreci tanımlar. Bu süreç içerisinde istatistik ve bilgisayar bilimleri alanlarında geliştirilmiş yöntemler kullanılır. Bilinen tanımlardan yola çıkarak veri madenciliği süreci, klasik istatistiksel yöntemler ile makine öğrenmesi yöntemlerinin bir arada kullanılabildiği, büyük veri yığınları içerisinden önceden öngörülemeyen anlamlı ve faydalı örüntülerin-bilgilerin bulunmaya çalışıldığı bir süreç olarak tanımlanabilir (Han & Kamber, 2001). İlgili literatürde veri madenciliği sürecinin aşamaları için farklı tanımlamalar yapılsa da en yaygını, Şekil 1'deki gibidir. Şekle göre, veri madenciliği sürecinde ilk önce, araştırılmak istenen problem ile ilgili elde edilen veritabanı içerisinden hedeflenen verinin seçimi gerçekleştirilir. Seçilen örneklem üzerinde veri önişleme işlemleri gerçekleştirilerek hatalı veri, eksik değer gibi sorunlar giderilir. Daha sonra gerekirse veri üzerinde dönüştürme işlemleri yapılabilir. Sonra çeşitli yöntemler kullanılarak modeller elde edilir ve elde edilen modeller test edilir. Uygun modeller ile elde edilen bilgi yorumlanarak süreç tamamlanır. Bu süreç doğrusal ilerleyen bir süreç değildir ve farklı aşamalardan daha önceki aşamalara geri dönmek gerekebilir. Daha detaylı bilgi için Akpınar (2014) ve Silahtaroğlu (2013) incelenebilir.



Şekil 1. Veri madenciliği süreci
Kaynak: (Fayyad, Piatetsky-Shapiro, & Smyth, 1996, s. 41)

Günümüzde, her alandaki ciddi veri artışı, bu verilerin etkili analiz yöntemleri ile anlamlandırılmasını da gerekli kılmaktadır. İşletmeler yoğun rekabet içerisinde bir adım öne geçebilmek için verileri etkin bir şekilde kullanmak istemekte, bu noktada da veri madenciliği büyük önem kazanmaktadır. Veri madenciliği bankacılık, sigortacılık, elektronik ticaret, sağlık, iletişim, ulaştırma, savunma, dolandırıcılık tespiti ve eğitim gibi pek çok farklı alanda kullanılmaktadır. Çünkü veri madenciliği her sektörün ihtiyacına cevap verecek yöntemlere sahiptir. Bu yöntemler uygulamadaki kullanımlarına göre genellikle iki gruba ayrılır. Ancak tanımlayıcı yöntemlerin tahminleme amacıyla kullanıldığı durumlar söz konusu olabilmektedir (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Tablo 3: Veri madenciliği yöntemleri
Kaynak: (Dunham, 2006)

Tahminleyici yöntemler	Tanımlayıcı yöntemler
Sınıflama (Karar ağaçları, k-en yakın komşu, yapay sinir ağları, destek vektör makineleri)	Kümeleme (Hiyerarşik, Bölümleyici, Yoğunluk tabanlı, Bulanık)
Zaman serisi analizi	Özet tablo ve grafikler
Regresyon analizi	Birliktelik kuralları
Nokta ve Aralık kestirim yöntemleri	Ardışık zamanlı örüntüler

Veri madenciliği sürecindeki aşamaların gerçekleştirilmesinde kullanılan yazılım büyük bir öneme sahiptir. Özellikle büyük miktarda verilerle çalışılan durumlarda kullanılan yazılımların hızlı ve tutarlı sonuçlar vermesi beklenmektedir. 2015 yılında 72 ülkeden 1220 uzman ile yapılan bir araştırmanın sonuçlarına göre en yaygın kullanılan ilk üç veri madenciliği aracı sırasıyla R, IBM SPSS Statistics ve SAS yazılımlarıdır (Rexer Analytics, 2016). Aynı araştırmada IBM SPSS Modeler önceki yıllara göre gerileyerek altıncı sırada yer almıştır. R'nin açık kaynak kodlu ve ücretsiz olması ayrıca veri madenciliği yöntemleri için kullanılabilen çok sayıda paketinin olması bu popülerliğin sebebi olarak görülebilir.

4.2. Karar Ağaçları

Karar ağaçları kök, dal ve yapraklardan meydana gelmektedir. Karar ağaçlarında bu yapı kullanılan algoritmaya göre karar verilen düğümler ile oluşturulur. Algoritmaların ağacı oluşturabilmesi için sınıf değeri belirli olan eğitim veri kümesi kullanılır. Bu nedenle karar ağaçları denetimli yöntemler olarak tanımlanır. Karar ağaçlarını oluşturmak için birçok algoritma geliştirilmiştir. ID3, C4.5, C5.0, "Classification and Regression Trees" (CART), "Chi-squared Automatic Interaction Detection" (CHAID) ve "Quick, Unbiased, Efficient, Statistical Tree" (QUEST) en bilinenleridir. Bu çalışmada en yüksek doğruluk oranları CART ve C5.0 algoritmaları ile elde edildiğinden sadece bu algoritmaların sonuçlarına yer verilmiştir. CART algoritması**,** Breiman, Friedman,

Olshen ve Stone tarafından 1984 yılında önerilmiştir. Bu algoritmada ağaç kök düğümden başlayarak her düğümde ikiye ayrılarak büyür. Hem kategorik hem de sürekli değişkenler ile kullanılabilir. Kategorik değişkenlerle çalışırken, sınıflandırma ağacı, sürekli değişkenlerle çalışırken ise regresyon ağacı olarak nitelendirilir. Bu algoritmada amaç bir önceki ayrımdan daha homojen bir ayrım yapabilmektir. Kategorik değişkenler için Twoing, Gini ve sürekli değişkenler için en küçük kareler sapması gibi kriterleri kullanır. Her düğümde iki ayrım yapılabildiğinden uzun karar ağaçları oluşturma eğilimindedir (Akpınar, 2014).

C4.5, Iterative Dichotomiser 3 (ID3) olarak adlandırılan ve 1986 yılında açıklanan algoritmanın Ross Quinlan tarafından geliştirilmiş halidir. Düğümler belirlenirken entropi veya enformasyon kazancı değerlerini kullanır. Hem kategorik hem de sürekli veriler ile çalışabilir. C5.0 olarak adlandırılan ve bazı konularda daha gelişmiş özelliklere sahip olan ticari sürümü yaygın olarak kullanılmaktadır. Kategorik değişkenler ile kullanılırken her bir farklı kategori için ayrı bir dal oluşturduğu için sonuçta geniş ağaçlar (çalı) elde edilebilir. Bunu engellemek için bazı kategoriler birleştirilebilir (Larose, 2005).

Bir karar ağacı, düğümdeki tüm örneklerin aynı sınıfa ait olması, dallanma için başka bir değişkenin kalmaması veya değişkeninin söz konusu değerine sahip kayıt bulunmaması durumları gerçekleşinceye kadar büyümeye devam eder. Ancak bu süreç istenilenden uzun sürebilir veya oluşan ağaç çok büyüyüp kullanışsız hale gelebilir. Bu durumlarda durdurma kuralları ile ağacın büyümesi kontrol altına alınabilir. Ağacın ulaşabileceği maksimum derinliğin veya bir düğümdeki minimum kayıt sayısının belirlenmesi yaygın kullanılan durdurma kurallarıdır. Bu kuralların ağacın olabildiğince büyümesine izin verecek şekilde tanımlanması ve sonrasında ağacın budanması ise kullanılan bir başka stratejidir. Budama sürecinde, karar ağacındaki sonucu etkilemeyen ve sınıflamaya katkısı olmayan dallar veya yapraklar budanır. Bu işlemde budanan dal veya yapraklardaki kayıtlar çıkartılmaz daha üstte yer alan düğümlere eklenir. Budama ağaç oluşturulduktan sonra yapılabildiği gibi ağacın oluşumu sırasında da yapılabilir (Akpınar, 2014; Silahtaroğlu, 2013).

Sınıflama yöntemlerinde verinin bir kısmı öğrenme amacıyla eğitim kümesi olarak diğer kısmı ise modelin doğruluk oranının belirlenmesi için test kümesi olarak kullanılır. Bir sınıflama modelinde yanlış olarak sınıflanan kayıt sayısının, toplam kayıt sayısına bölünmesi ile hata oranı, doğru olarak sınıflanan kayıt sayısının toplam kayıt sayısına bölünmesi ile doğruluk oranı bulunur. Basit geçerlilik yaklaşımında tüm verilerin bir kısmı rasgele seçilerek eğitim kümesi oluşturulur, kalan kısım test kümesi olarak kullanılır. Model eğitim kümesi kullanılarak geliştirilir. Modelin doğruluk oranı test kümesi kullanılarak belirlenir. Çapraz geçerlilik yaklaşımında, öğrenim veri kümesi rasgele eşit genişlikte k alt kümeye ayrılır. k alt kümenin birisi test için ve diğer $(k-1)$ alt küme ise öğrenim için kullanılır. Bu işlem tüm k alt küme test için en az bir kez kullanılacak şekilde

tekrarlanır. Bu işlemler sonucunda elde edilen hata oranlarının ortalaması modelin doğruluk oranı olarak kullanılır. Bunların dışında, modelin doğruluk oranının belirlenmesinde yeniden örnekleme yaklaşımını kullanan "bootstrapping" ve "leave-one-out" yöntemleri de kullanılmaktadır (Han & Kamber, 2001).

4.3. Veri Seçimi, Önişleme ve Dönüştürme

Bu çalışmada, TÜİK tarafından her ay düzenli olarak gerçekleştirilen HİA'nın 2013 yılı mikro verileri kullanılmıştır (TÜİK, 2014a). 2013 yılı HİA'da, 146.055 haneden toplam 502.426 kişi için veri toplanmıştır. Araştırmada kullanılan soru formunda 100 değişken bulunmaktadır. Yapılan ön incelemeler sonucunda aşağıda açıklanan 15 değişken analizler için uygun bulunmuştur. Çalışma kapsamında işgücü incelendiği için 15 yaş ve üstü kurumsal olmayan nüfusa (379.742 kişi) ait yanıtlar kullanılmıştır. HİA en güncel nüfus projeksiyonlarına göre ağırlıklandırılmaktadır. Bu nedenle, tanımlayıcı istatistikler için veri kümesinde yer alan "Faktör" değişkeni dikkate alınmış ve ağırlıklandırılmış veriler üzerinden istatistikler hesaplanmıştır. Karar ağacı modellerinde, işgücünde yer alan ve bir yıl önce (2012'de) çalışmadığını ifade eden kişilere ait veriler arasından rasgele seçilen 22.206 kayıt kullanılmıştır. Yanıt değişkeni, "işgücü durumu" olarak belirlenmiştir. Çalışmada kullanılan değişkenler, kategorileri ve yapılan dönüşümler aşağıda tanımlanmıştır:

- İşgücü durumu: Bireylerin istihdamda, işsiz veya işgücüne dâhil olmayan nüfus içinde yer alma durumunu belirten, kategorik bir değişkendir. 15 yaşından küçük olan bireyler, çalışma çağındaki nüfus içerisinde yer almadığından dolayı, analizlere dâhil edilmemiştir. Değişkenin değerleri: 1- İstihdam / 2- İşsiz / 3- İşgücüne dâhil olmayan.
- Bir yıl önceki işgücü durumu: Bireyin bir yıl önceki işgücü durumunu belirten, 10 değerli kategorik değişken. Değişkenin değerleri: 1- Bir işte çalışıyordu / 2- Şu anki işinde çalışıyordu / 3- Emekliydi / 4- İş arıyordu / 5- Ev işleriyle meşguldü / 6- Eğitim/öğretimine devam ediyordu / 7- Özürlü veya hastaydı / 8- Askerdi / 9- Yaşlı (65 ve daha yukarı yaş) / 10- Diğer
- Cinsiyet: Değişkenin değerleri: 1- Erkek / 2- Kadın.
- Referans kişiye yakınlık: Hanehalkının geçiminden sorumlu olan kişi, hanenin referans kişisi olarak belirlenmektedir. Hanede yaşayan diğer kişilerin, bu kişiye olan yakınlık durumlarının belirtildiği değişkendir. Değişkenin değerleri: 1- Referans kişi / 2- Eşi / 3- Çocuğu / 4- Gelini veya damadı / 5- Torunu / 6- Kendi veya eşinin anne/babası / 7- Diğer akrabalar/ 8-Akraba olmayanlar.
- Yaş grupları: Kişilerin yaşlarına göre oluşturulmuş olan gruplardır. HİA verilerindeki yaş grubu değişkeninin 14 düzeyi bulunmaktadır. Bu çalışma için bazı kategoriler birleştirilerek, sınıflama algoritmalarının performanslarının arttırılması hedeflenmiştir. 8 düzeyi olan değişkenin değerleri: 1- (0-14) yaş

arası / 2- (15-19) yaş arası / 3- (20-24) yaş arası / 4- (25-29) yaş arası / 5- (30-39) yaş arası / 6- (40-49) yaş arası / 7- (50-59) yaş arası / 8- 60 yaş ve üzeri.

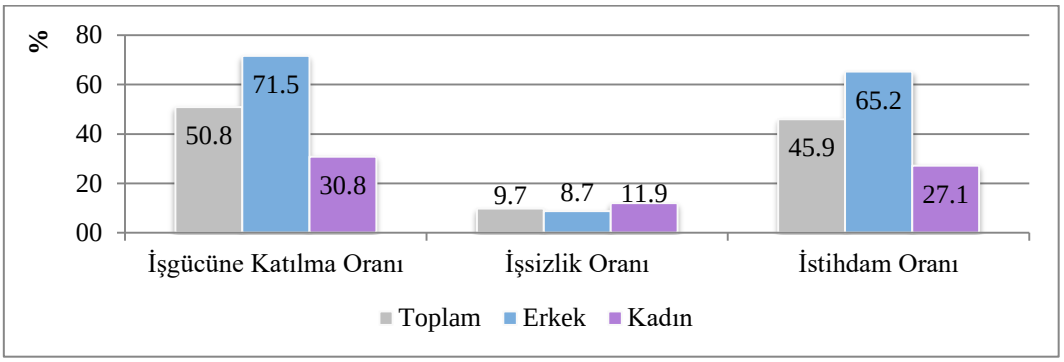
- Medeni durum: Bireylerin evli veya bekâr olmak üzere medeni durumlarını belirten değişkendir. HİA'da 4 kategoriye sahip olan değişken, evli ve bekâr olmak üzere 2 kategoride analize dâhil edilmiştir. Değişkenin değerleri: 1- Evli / 2- Bekâr.
- Kır/kent: Kişilerin ikamet ettiği yerleşim biriminin bulunduğu yerin kır veya kent olma durumlarını belirten değişkendir. Değişkenin değerleri: 1- Kır / 2- Kent.
- Mezuniyet: Kişilerin en son mezun oldukları eğitim kurumunun seviyesini belirten değişkendir. HİA'da 6 kategoriye sahip olan değişken, 4 kategoriye düşürülerek analize dâhil edilmiştir. Değişkenin değerleri: 1- Bir okul bitirmeyen / 2- İlköğretim / 3- Ortaöğretim / 4- Yükseköğretim.
- Okula devam: Kişilerin herhangi bir örgün eğitim kurumuna devam edip/etmeme durumlarını belirten değişkendir. Değişkenin değerleri: 1-Evet / 2-Hayır.
- Aynı ilde yaşama: Bireylerin doğdukları andan itibaren, anket yapılma zamanına kadar geçen süre içinde, aynı ilde yaşama durumlarını belirten değişkendir. Değişkenin değerleri: 1- Evet / 2- Hayır.
- NUTS-1: Kişilerin ikamet ettiklerin yerlerin, İstatistiki Bölge Birimleri Sınıflaması 1. Düzey durumlarını belirten değişkendir. Değişkenin değerleri: 1- İstanbul / 2- Batı Marmara / 3- Ege / 4- Doğu Marmara / 5-Batı Anadolu / 6- Akdeniz / 7-Orta Anadolu / 8- Batı Karadeniz / 9-Doğu Karadeniz / 10- Kuzeydoğu Anadolu / 11-Ortadoğu Anadolu / 12-Güneydoğu Anadolu.
- Faaliyet: Bireyin çalışmış olduğu işyerinin faaliyetini ifade eden ve NACE Rev.2 kodlamasından elde edilen değişkendir. HİA'da 2'li kodlar halinde 88 kategoride verilen değişken, tarım, sanayi ve hizmet olmak üzere 3 kategori haline dönüştürülerek analize dâhil edilmiştir. Değişkenin değerleri: 1- Tarım / 2- Sanayi / 3- Hizmet.
- Meslek: Bireyin çalışmış olduğu işyerinde, yapmış olduğu işi ifade eden ve ISCO-08 kodlamasından elde edilen değişkendir. HİA'da 2'li kodlar halinde 43 kategoride verilen değişken, tekli kodlar olarak 9 ana kategori haline dönüştürülerek analize dâhil edilmiştir. Değişkenin değerleri: 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 ürünleri çalışanları / 7-Sanatkârlar ve ilgili işlerde çalışanlar / 8-Tesis ve makine operatörleri ve montajcılar / 9-Nitelik gerektirmeyen işlerde çalışanlar.
- Mezun olduğu alan: En son bitirilen okulda mezun olunan bölümü belirten değişkendir. 21 farklı değeri bulunmaktadır.
- Faktör: Ağırlık katsayısı. Ağırlıklandırma, cevapsızlık düzeltmeleri, dışsal dağılım kontrolleri (İBBS kır kent, yaş grubu - Cinsiyet bazında) ve nihai düzeltme katsayısı dikkate alınarak hesaplanmaktadır. Bu değişken ile güncel nüfusa göre tahminlerin yapılması sağlanmaktadır.

5. Bulgular

Bu bölümde, ilk olarak, çalışmada kullanılan bazı değişkenler için 15 yaş ve üzeri kurumsal olmayan nüfusa ait verilerden elde edilen tanımlayıcı istatistiklere ve grafiklere yer verilecektir[§]. Bölüm 5.2.'de "işgücü durumu" yanıt değişkeni için geliştirilen karar ağacı modellerinden bahsedilecektir.

5.1. Tanımlayıcı İstatistikler ve Grafikler

2013 yılı HİA 502.426 kişi ile gerçekleştirilmiştir. Bu kişilerin %75,58'i (379.742 kişi) 15 yaş ve üstü kurumsal olmayan nüfustadır. Çalışmaya katılanların demografik özellikleri Tablo 4'de verilmiştir. HİA'da işgücü kapsamında kurumsal olmayan nüfustaki 15 yaş ve üstü kişilerden veri toplanmaktadır. Bu nedenle takip eden bölümdeki grafiklerde ve tablolarda kurumsal olmayan nüfustaki 15 yaş ve üstü kişilere ait veriler kullanılmıştır. Bu bölümdeki istatistiklerin hesaplanmasında ve grafiklerde örneklemden elde edilen ham veriler değil, "Faktör" değişkeni dikkate alınarak ağırlıklandırılmış veriler kullanılmıştır. Bir başka deyişle bu bölümde verilen istatistikler 2013 yılı Türkiye nüfusu için kestirim değerlerini yansıtmaktadır.



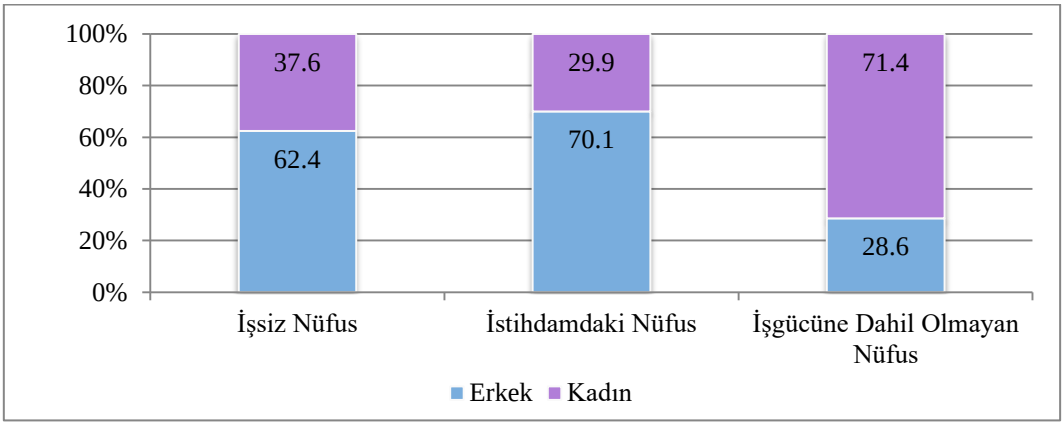
Şekil 2. 2013 yılında, Türkiye'de nüfusun cinsiyete göre işgücü durumu

2013 yılında, Türkiye'de işgücüne katılma oranının %50,8 olduğu Şekil 2'den görülmektedir. Bu oran, erkeklerde %71,5 iken kadınlarda ise %30,8 olarak gerçekleşmiştir. İşsizlik ve istihdam oranları incelendiğinde, kurumsal olmayan nüfusun %45,9'u istihdamda yer alırken, işgücüne katılan nüfusun %9,7'si işsiz nüfusta yer almıştır. Erkeklerin istihdam oranı %65,2 ile kadınların istihdam oranına göre çok daha fazla olurken, kadınların işsizlik oranı da %11,9 ile erkeklerin işsizlik oranına göre daha fazla olmuştur.

[§] Bu bölümdeki tanımlayıcı istatistikler ve grafikler, TÜİK tarafından yayımlanan 2013 işgücü istatistikleri haber bülteninde yer almamaktadır. Mikro veriden elde edilen değerlere ve grafiklere yer verilmiştir.

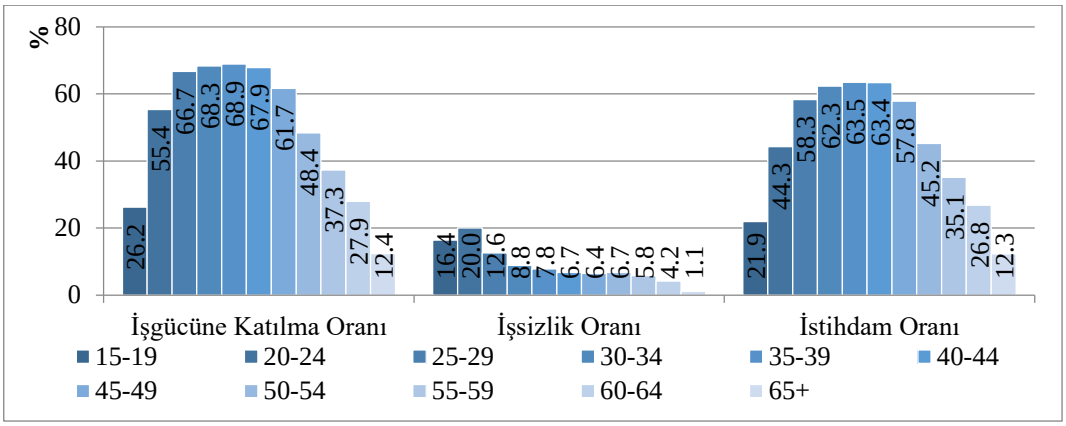
Tablo 4: Örneklemin demografik özellikleri

	Değişken		Sayı	Yüzde
Tüm örneklem (n=502.426)	Cinsiyet	1-Erkek	245.173	%48,80
		2-Kadın	257.253	%51,20
	Yaş Grubu	1- 00-04 yaş arası	36.572	%7,28
		2- 05-11 yaş arası	57.721	%11,49
		3- 12-14 yaş arası	28.391	%5,65
		4- 15-19 yaş arası	43.096	%8,58
		5- 20-24 yaş arası	32.217	%6,41
		6- 25-29 yaş arası	35.021	%6,97
		7- 30-34 yaş arası	38.808	%7,72
		8- 35-39 yaş arası	37.174	%7,40
		9- 40-44 yaş arası	35.730	%7,11
		10- 45-49 yaş arası	33.534	%6,67
		11- 50-54 yaş arası	30.352	%6,04
		12- 55-59 yaş arası	26.492	%5,27
		13- 60-64 yaş arası	21.099	%4,20
		14- 65+ yaş	46.219	%9,20
	En son bitirilen okul	0- 5 yaşından küçük olanlar	36.572	%7,28
		1- Bir okul bitirmeyen	112.723	%22,44
		2- İlkokul (5 yıl)	164.772	%32,80
		3- Ortaokul, mesleki ortaokul ve ilköğretim (8 yıl)	79.731	%15,87
		4- Genel lise	36.018	%7,17
		5- Mesleki veya teknik lise	28.760	%5,72
		6- Yüksekokul, fakülte ve üzeri	43.850	%8,73
	Kır-Kent	1-Kır	139.723	%27,81
		2-Kent	362.703	%72,19
15 ve daha yukarı yaştakiler (n=379.742)	Medeni Durum	1-Hiç evlenmedi	92.203	%24,28
		2-Evli	253.422	%66,74
		3-Boşandı	8.652	%2,28
		4-Eşi öldü	25.465	%6,71
	İşgücü Durumu	1- İstihdam	164.176	%43,23
		2- İşsiz	16.734	%4,41
		3- İşgücüne dâhil olmayan	198.832	%52,36



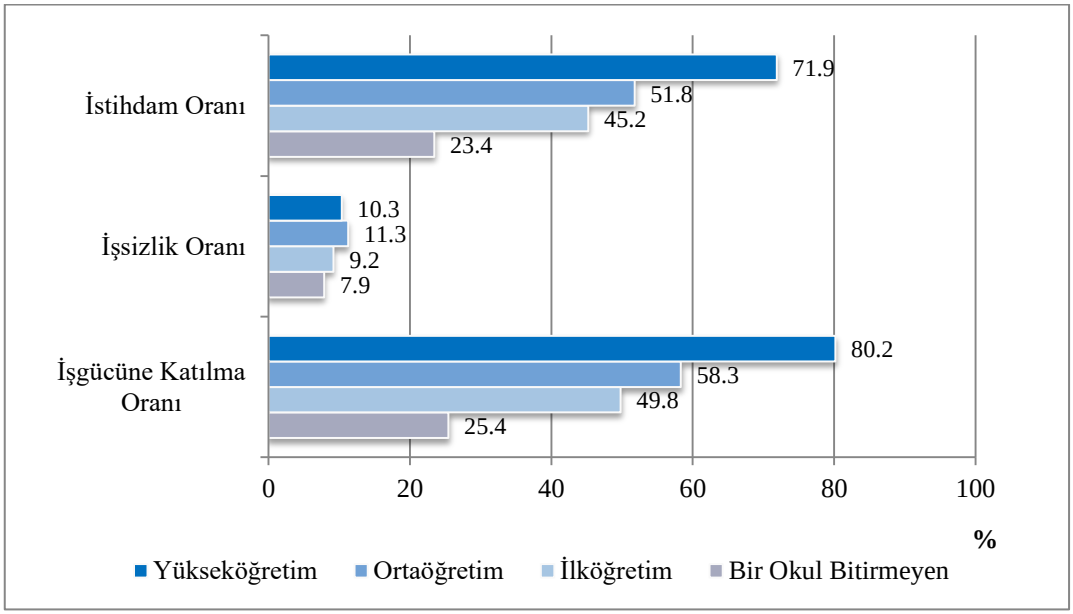
Şekil 3. Cinsiyet göre işgücü durumlarının dağılımı

İşgücü durumlarının cinsiyete göre dağılımları Şekil 3'te verilmiştir. Bu grafikten kadınların istihdamdaki oranının %29,9'da kaldığı görülmektedir.



Şekil 4. Yaş gruplarına göre işgücü durumlarının dağılımı

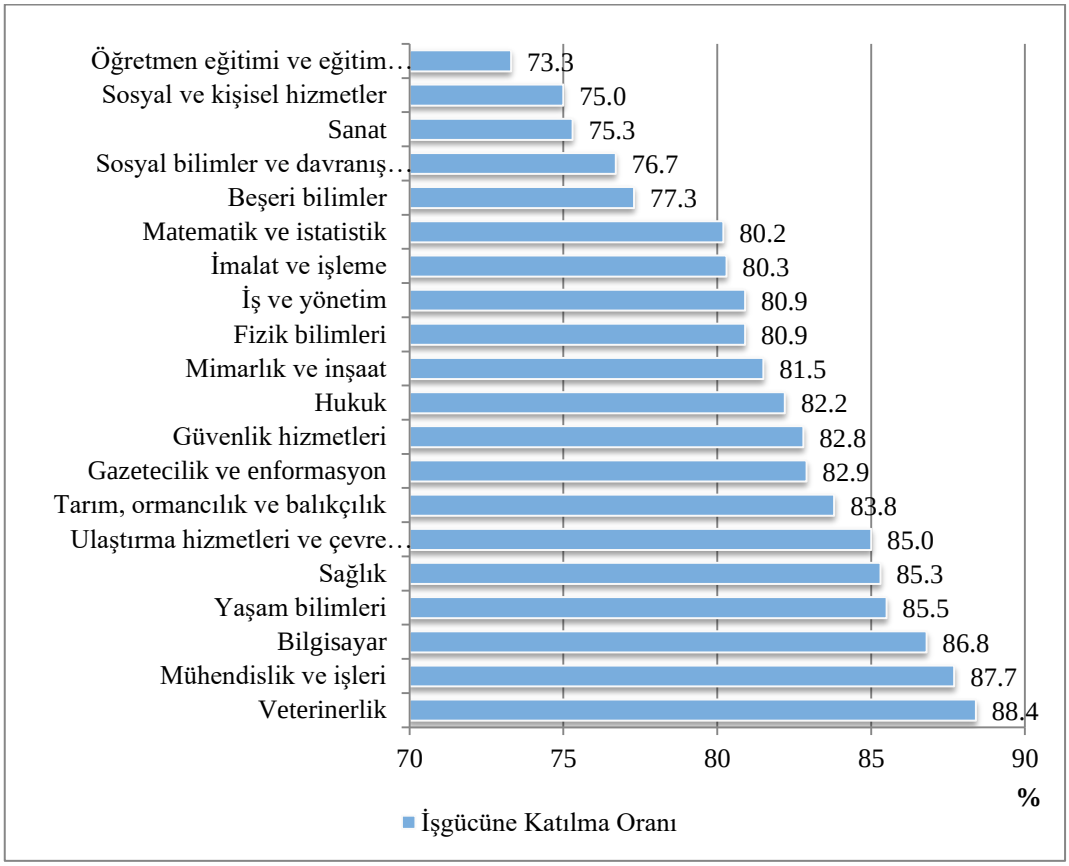
Yaş gruplarına göre işgücüne katılım oranları incelendiğinde (Şekil 4), 35-39 yaş grubundaki nüfusun %68,9 ile işgücüne dâhil olma oranının en yüksek grup olduğu görülmektedir. İşgücüne katılma oranının en düşük olduğu yaş grubunu ise %12,4 ile 65 yaş ve üzerindeki nüfusun yer aldığı yaş grubu oluşturmaktadır. Grafikten anlaşıldığı üzere, 50 yaşından itibaren nüfusun işgücüne katılım oranı düşmektedir. İşsizlik oranına göre yaş grupları incelendiğinde, en yüksek işsizlik oranının %20 ile 20-24 yaş grubunda olduğu görülmektedir. İstihdam oranının en yüksek olduğu yaş grubu ise %63,5 ile 35-39 yaş grubu olmuştur. 60-64 yaş aralığındaki istihdam oranının (%26,8), 15-19 yaş grubundaki nüfusun istihdam oranından daha yüksek olduğu görülmektedir.



Şekil 5. Eğitim durumuna göre işgücü durumlarının dağılımı

Şekil 5'te, mezun olunan eğitim kurumu seviyesine göre işgücü durumları gösterilmektedir. Grafiğe göre yükseköğretim mezunu nüfusun %80,2'si gibi büyük bir kısmı işgücüne dâhil olmakta iken bu oran ortaöğretim için %58,3 ve ilköğretim için %49,8 olarak gerçekleşmiştir. Bir okul bitirmeyen nüfusta ise bu oran %25,4 düzeyinde kalmıştır. İstihdam oranı, yükseköğretim mezunu nüfusta %71,9 ile en yüksek orana sahipken bir okul bitirmemiş nüfusta %23,4 ile en düşük seviyede gerçekleşmiştir. İşsizlik oranına göre eğitim durumları incelendiğinde, en düşük işsizlik oranının %7,9 ile bir okul bitirmemiş nüfusta olduğu görülmektedir. İşsizlik oranının eğitim durumuna göre değişimine dikkat edildiğinde, işgücünde yer alan düşük eğitim seviyesine sahip olan nüfustaki işsizlik oranının, yüksek eğitim seviyesindeki nüfusun işsizlik oranına göre daha az olduğu dikkat çekmektedir.

Yükseköğretim mezunlarının en son mezun oldukları alana göre işgücüne katılma oranları Şekil 6'da gösterilmektedir. Grafiğe göre işgücüne katılma oranının en yüksek olduğu alan %88,4 ile veterinerlik bölümüdür. Bunu sırasıyla %87,7 ile mühendislik ve işleri ve %86,8 ile bilgisayar bölümleri takip etmektedir. İşgücüne katılma oranının en düşük seviyede olduğu alan ise %73,3 ile öğretmen eğitimi ve eğitim bilimleri bölümü olmuştur. Buradan, öğretmen eğitimi ve eğitim bilimleri alanından mezun olanların %26,7'si gibi ciddi bir kısmının, çeşitli nedenlerle işgücüne dâhil olmadığı anlaşılmaktadır.

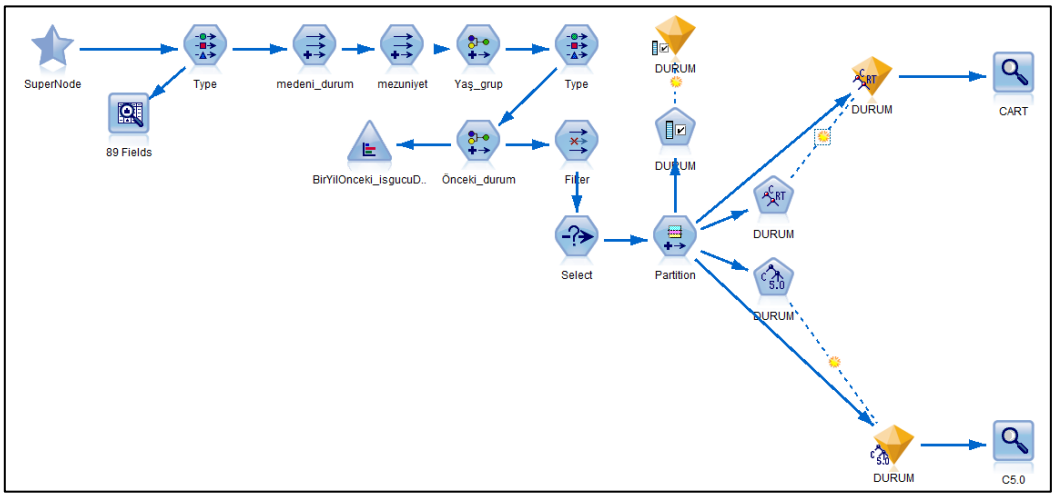


Şekil 6. Yükseköğretim mezunlarının mezun oldukları alana göre işgücüne katılma oranları

5.2. Karar Ağacı Uygulaması

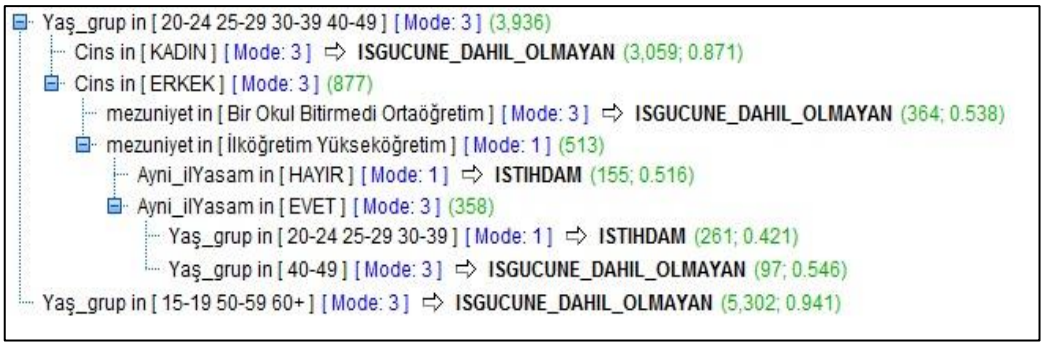
Bu bölümde, işgücünde yer alan ve bir yıl önce (2012'de) çalışmadığını ifade eden kişilerin, 2013 yılındaki işgücü durumları hedef değişken olarak belirlenerek CART ve C5.0 algoritmaları ile kurulan karar ağacı modellerine yer verilmiştir. 2012 yılında çalışmadığını belirten nüfus içerisinde 2013 yılında erkeklerin %71,9'u, kadınların ise %90,9'u işgücüne dâhil olmayan nüfusta yer almıştır. 2012'de çalışmadığını belirten nüfusun 2013 yılında istihdamda olma durumuna bakıldığında erkeklerin %17,4'ünün, kadınların ise %5,1'inin istihdamda yer aldığı görülmüştür (Bu oranlar "Faktör" değişkeni kullanılarak ağırlıklandırılmış veriler ile hesaplanmıştır). Bu çalışma kapsamında, işgücünde yer alan ve bir yıl önce çalışmadığını ifade eden kişilere ait veriler arasından rasgele seçilen 22.206 kayıt kullanılmıştır. Yanıt değişkeni, "işgücü durumu" olarak belirlenmiştir. Bir yıl önceki çalışma durumları belirlenirken, bir yıl önceki işgücü durumunu belirten, 10 değerli kategorik değişkenden faydalanılmıştır. Değişkenin kategorileri "çalışan" ve "çalışmayan" olmak üzere iki kategoride birleştirilmiş ve sadece "çalışmayan" bireylere ait veriler analize dâhil edilmiştir. Böylece son bir

yıl içerisinde işgücü durumu değişen veya değişmeyen kişilere ait özellikleri keşfetmek amaçlanmıştır. Modeller için IBM SPSS Modeler programında yapılan işlemler Şekil 7'de gösterilmiştir. Şekil 7'de görülen SuperNode nesnesi karmaşıklığı azaltmak için kullanılmıştır. Bu nesne içerisinde verilerin kaynak dosyadan çekilmesi, filtrelenmesi, örneklem çekilmesi özet tabloların oluşturulması gibi işlemler bulunmaktadır.



Şekil 7. IBM SPSS Modeler ile modelleme süreci

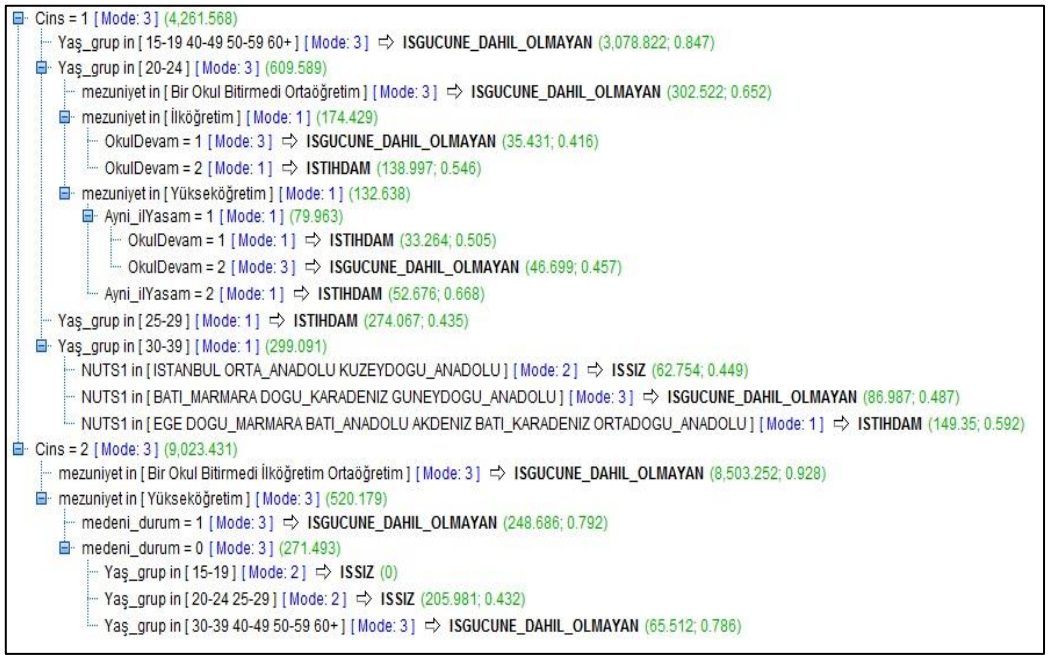
Öncelikle CART algoritması ile elde edilen sonuçlara yer verilecektir. İlk olarak, ön çalışmalar neticesinde belirlenen değişkenler arasından yanıt değişkeni üzerinde en etkili değişkenler belirlenmiştir. Buna göre "yaş grubu", "cinsiyet", "mezuniyet", "aynı ilde yaşama", "okula devam", "NUTS1", "medeni durum" değişkenleri CART algoritması için modele alınmıştır. Şekil 8'de CART algoritması ile elde edilen kural çıktısı gösterilmektedir. CART ile elde edilen karar ağacı oldukça uzun olduğundan ağaç görünümüne yer verilememiştir.



Şekil 8. CART ile elde edilen kural çıktısı

CART karar ağacı modelinde ilk ayrım "yaş grubu" değişkenine göre gerçekleşmiştir. İlk ayrıma göre, bir önceki yıl çalışmayan bireylerden 15-19, 50-59 ve 60 yaş ve üzeri yaş gruplarında bulunanlar 2013 yılında işgücüne dâhil olamamışlardır (%94,1). 20-49 yaş gruplarındaki kişiler için ikinci ayrım "cinsiyet" değişkenine göre yapılmıştır. Bu kurala göre bir önceki yıl çalışmadığını beyan eden kadınlar 2013 yılı için de işgücünde yer alamamışlardır (%87,1). Sonraki ayrım 20-49 yaş gruplarındaki erkek nüfusun "mezuniyet" durumuna göre yapılmıştır. Bu gruptaki kişilerin "mezuniyet" durumu bir okul bitirmeyen veya ortaöğretim olanlarının %53,8'i işgücüne dâhil olmamıştır. İlköğretim veya yükseköğretim mezunu kişileri için sonraki dallanma "aynı ilde yaşama" değişkenine göre yapılmıştır. Bu kurala göre doğduklarından beri aynı ilde yaşadığını belirten bireylerin 20-39 yaş gurubunda olanların %42,1'inin işgücü durumu değişmiş ve bu yıl "istihdamda" olarak sınıflandırılmışlardır. Bu gruptakilerin %26,8'i "işsiz", %31,1'i "işgücüne dâhil olmayan" şeklinde sınıflandırılmıştır.

C5.0 algoritması için önemli değişkenler "cinsiyet", "yaş grubu", "mezuniyet", "medeni durum" ,"okula devam", "aynı ilde yaşama" ve "NUTS1" olarak bulunmuştur. Bu değişkenler için elde edilen karar ağacının kural çıktısı Şekil 9'da verilmektedir. C5.0 ile oluşturulan karar ağacı oldukça geniş olduğundan ağaç görünümüne yer verilememiştir.



Şekil 9. C5.0 ile elde edilen kural çıktısı

C5.0 algoritmasında ilk ayrım, "cinsiyet" değişkenine göre yapılmıştır. Modelde, cinsiyetten sonraki dallara ayırma erkeklerde yaş gruplarına göre, kadınlarda ise

mezuniyet durumlarına göre gerçeklemiştir. C5.0 algoritmasına göre, bir önceki sene (2012'de) çalışmadığını belirten bireyler içerisinde, 15-19 veya 40 ve üzeri yaştaki erkekler 2013 yılında "işgücüne dâhil olmayan" sınıfında yer alırken, 25-29 yaş aralığındaki erkekler "istihdamda" olarak sınıflandırılmışlardır. 20-24 yaş aralığındaki yükseköğretim mezunu ve doğduğundan beri aynı ilde yaşayan erkeklerden, 2013 yılında, bir öğrenim kurumuna devam edenler istihdamda yer alırken, okula devam etmeyenler işgücüne dâhil olmayan nüfusta yer almıştır. 20-24 yaş aralığındaki yükseköğretim mezunu ve doğduğundan beri aynı ilde yaşamayan erkekler de, 2013 yılında, istihdamda yer almışlardır. 30-39 yaş aralığında bulunan erkeklerden, İstanbul, Orta Anadolu ve Kuzeydoğu Anadolu bölgelerinde ikamet edenler işsiz nüfusta yer alırken, Batı Marmara, Doğu Karadeniz ve Güneydoğu Anadolu bölgelerinde ikamet edenler işgücüne dâhil olmayan nüfusta ve Ege, Doğu Marmara, Batı Anadolu, Akdeniz, Batı Karadeniz ve Ortadoğu Anadolu bölgelerinde ikamet edenler ise istihdamdaki nüfusta yer almışlardır. Buradan, coğrafi bölgelerin istihdam durumları üzerinde etkili olduğu anlaşılmaktadır. Bir yıl önce (2012'de) çalışmadığını belirten kadınların, 2013 yılında istihdamdaki nüfusta yer alamadıkları görülmektedirler. 15-29 yaş aralığında bulunan yükseköğretim mezunu bekâr kadınların işsiz nüfusta yer aldığı, 30 ve üzeri yaştaki kadınların ise işgücüne dâhil olmayan nüfusta yer aldığı görülmektedir. Yükseköğretim mezunu evli kadınların ise işgücüne dâhil olamadıkları dikkat çekmektedir.

Elde edilen modellerin doğruluk oranları basit doğrulama yöntemine göre bulunmuştur. Bu yöntemde modelleme için kullanılan verilerin %10 ila %40'lık bir bölümü rasgele seçilerek test kümesi olarak ayrılır ve modellemede kullanılmaz. Geriye kalan eğitim kümesi kullanılarak model kurulur. Modelin doğruluk oranı test kümesi ile belirlenir. Bu çalışmada verilerin yaklaşık %40'ı test için kullanılmıştır. CART algoritması ile elde edilen modelin doğruluk oranı, %87,13 olarak bulunmuştur. C5.0 ile elde edilen modelin doğruluk oranı ise %86,85 seviyesinde gerçekleşmiştir.

6. Sonuç

Ülkemizde son yıllarda artış gösteren işsizlik sorununun çözümü için geliştirilecek istihdam politikalarının doğru belirlenebilmesinde işgücü konusunda yapılacak araştırmalara ihtiyaç vardır. TÜİK bu araştırmalara ülke çapında topladığı verilerle olanak sağlamaktadır. HİA, 2000 yılından itibaren düzenli olarak tüm ülke genelini yansıtacak şekilde yapılmaktadır. Sağlanan veriler, ülkenin işgücü durumunu ortaya koymada, istihdam edilenlerin ya da işsizlerin özelliklerini belirlemede kullanılabilmektedir. Bu çalışmada, 2013 yılı TÜİK HİA verilerini kullanarak Türkiye'de işgücünün yapısı hakkında bilgi verilmiş ve işgücü durumu için karar ağacı modelleri kurulmuştur. Araştırmada, işgücü durumu incelendiğinden 15 yaş ve üzeri olan nüfus analizlere dâhil edilmiştir. Elde edilen istatistiklere göre, 2013 yılında Türkiye'de kurumsal olmayan 15 yaş

ve üzeri nüfusun işgücüne katılım oranı %50,8, istihdam oranı %45,9 ve işsizlik oranı %9,7 seviyesinde gerçekleşmiştir. Yılmaz tarafından verilen sonuçlara göre bu oranlar sırasıyla 2009 yılında %46, %39,8, %13,3 ve 2010 yılında %46,7, %41,4, %11,3 olarak gerçekleşmiştir (Yılmaz, 2012). Bu geçmiş değerler ile karşılaştırıldığında 2013 yılında istihdama katılımın arttığı ve işsizlik oranının azaldığı görülmektedir.

Erkeklerin işgücüne katılımı %71,5 ile kadınlara (%30,8) oranla çok daha yüksektir. Yılmaz (2012), çalışmasında 2009 ve 2010 yıllarında kadınların işgücüne katılım oranının erkeklerin üçte biri kadar olduğunu belirtmiştir. Buna göre kadınların işgücüne katılım oranında bir iyileşme olduğu söylenebilir. Bulgulara göre, kadınların işgücüne katılımının erkeklerden çok daha düşük olmasının yanı sıra, işgücüne katılan nüfustaki işsiz kadınların oranı işsiz erkeklere göre daha yüksektir. Ayrıca istihdamdaki kadınların oranı istihdamdaki erkeklere göre daha düşük bir seviyededir. Bu oranlar, Türkiye'de erkek nüfus ağırlıklı bir işgücü yapısı olduğunu bir kez daha göstermektedir. Kadınların işgücüne katılımının artması amacıyla planlamalar yapılmalı, kadınlara yönelik meslek edindirici kursların sayısı ve çeşitliliği artırılmalıdır. İşgücünde yer alanlar yaş gruplarına göre incelendiğinde, işgücüne katılımın en yüksek 35-39 yaş aralığındaki nüfusta olduğu görülmüştür. 15-24 yaş aralığında yer alan genç nüfusta ise işgücüne katılımın diğer yaş gruplarına göre düşük ancak işsizlik oranının daha yüksek olduğu belirlenmiştir. Ülke nüfusunun önemli bir kısmını oluşturan genç nüfusun, işgücündeki oranlarının istenilen seviyeye çıkarılamadığı anlaşılmaktadır. Genç nüfustaki işsizliğe ve düşük istihdama odaklı çalışmalar genç nüfusun işgücüne katılımını arttırıcı politikaların belirlenmesine yardımcı olabilir. İşgücünde yer alan nüfusun eğitim durumu incelendiğinde, eğitim seviyesinin işgücüne katılımı etkilediği tespit edilmiştir. Özellikle yükseköğretim mezunu nüfusta işgücüne katılımın %80,2 ile en yüksek oranda olduğu görülmüştür. Ayrıca yüksek eğitimli nüfusun istihdam oranının düşük eğitimli nüfusa göre daha yüksek olduğu belirlenmiştir. Ancak, istihdamda yer alan nüfusun %52,3 gibi büyük bir kısmı ilköğretim mezunu nüfustan oluşmaktadır.

Bu çalışma kapsamında, bir yıl önce (2012'de) çalışmayan bireylerin 2013 yılındaki işgücü durumları CART ve C5.0 algoritmaları ile modellenmeye çalışılmıştır. CART algoritması için en önemli değişkenler, yaş grubu, cinsiyet, mezuniyet, aynı ilde yaşama olarak belirlenmiştir. CART algoritmasına göre; 2013 yılında kurumsal olmayan nüfusta yer alan ve bir yıl önce çalışmadığını beyan eden bireylerin %86,8 ile büyük bir kısmı işgücüne dâhil olmayan şeklinde sınıflanmıştır. Bu kişilerin büyük çoğunluğu, bir önceki sene de işgücüne dâhil olmayan nüfusta yer alan emekliler, ev hanımları, özürlü / hastalar ve yaşlılardan oluşmaktadır. Ayrıca bir önceki sene işsiz nüfusta yer alan, fakat çeşitli nedenlerle işgücüne dâhil olmayan nüfusa geçen kişiler de bu oran içinde yer almaktadır. Bir önceki sene çalışmadığını beyan edenlerin içinde, istihdamda olanların oranı %7,9 ve işsizlerin oranı ise %5,3 düzeyinde bulunmuştur. CART sonuçlarına göre, bir

sene önce çalışmadığını belirten bireylerin, cinsiyetleri, mezuniyet durumları, ikamet ettikleri ildeki süreklilik durumları ve yaş grupları işgücü durumlarını etkileyen en önemli faktörler olarak bulunmuştur. C5.0 algoritması ile elde edilen sonuçlara göre; 2013 yılında kurumsal olmayan nüfusta yer alan ve bir yıl önce çalışmadığını belirten nüfusun %84,8 ile önemli bir kısmı işgücüne dâhil olmayan nüfusta, %9'u istihdamda ve %6,1'i ise işsiz nüfusta tahmin edilmiştir. C5.0 karar ağacına göre, bir yıl önce çalışmadığını belirten bireylerin cinsiyetleri, yaş grupları, mezuniyet durumları, medeni durumları, bir okula devam edip etmedikleri, ikamet ettikleri ildeki süreklilik durumları ve ikamet ettikleri coğrafi bölge işgücü durumlarını etkileyen en önemli faktörlerdir.

Bu çalışmadaki uygulamada işgücü durumuna odaklanılmıştır. Sonraki çalışmalarda, HİA kapsamında toplanılan veriler ile istihdam edilen bireylerin sektörlere, mesleklere göre dağılımı, çalışma süreleri ve gelir durumları incelenerek istihdamdaki nüfusun özellikleri araştırılabilir. Benzer şekilde işsiz nüfusun nitelikleri, iş arama süreleri, iş arama kanalları gibi değişkenler ileriki araştırmalara konu edilebilir. Ayrıca sonraki yıllara ait HİA verilerinin kullanıldığı benzer çalışmalar ile 2013 yılı bulguları karşılaştırılabilir.

Kaynakça

Akpınar, H. (2014). *Data veri madenciliği veri analizi.* İstanbul: Papatya Yayıncılık.

Baltagi, B. H., Baskaya, Y. S., ve Hulagub, T. (2012). The Turkish wage curve: Evidence from the household labor force survey. *Economics Letters, 114*(1), 128-131.

Dunham, M. H. (2006). *Data mining: Introductory and advanced topics.* Pearson Education India.

Fayyad, U., Piatetsky-Shapiro, G., ve Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine, 17*(3), 37-54.

Han, J., ve Kamber, M. (2001). *Data mining concepts and techniques.* Academic Press.

Larose, D. T. (2005). *Discovering knowledge in data: An introduction to data mining.* John Wiley & Sons.

Oğuzlar, A. (2004). CART analizi ile hanehalkı işgücü anketi sonuçlarının özetlenmesi. *Atatürk Üniversitesi İİBF Dergisi, 18*, 79-90.

Özkan, H. Ö. (2013). *Labor market mobility and marginal attachment in Turkey: Evidence from hlfs, 2000-2002.* Yayınlanmamış Yüksek Lisans Tezi, Koç Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Rexer Analytics. (2016, Nisan). *2015 Data science survey highlights.* (Erişim tarihi:13.06.2017), http://www.rexeranalytics.com/files/Rexer_Data_ Science_Survey_Highlights_Apr-2016.pdf .

Silahtaroğlu, G. (2013). *Veri madenciliği kavram ve algoritmalar*. İstanbul: Papatya Yayıncılık.

Tansel, A. (2001). Economic development and female labor force participation in Turkey: time-series evidence and cross-province estimates. *Economic Research Forum for the Arab Countries.*

Tunalı, İ., ve Ercan, H. (2003). *Background study on labour market.* European Training Foundation: Torino.

TÜİK. (2014a). Hanehalkı işgücü araştırması mikro veri seti 2013. Ankara: TÜİK Matbaası.

TÜİK. (2014b). *Türkiye İstatistik Yıllığı 2013.* Ankara: TÜİK Matbaası.

Yılmaz, E. (2012). İ*statistiksel analiz yöntemi olarak veri madenciliğinde CHAID algoritması ve Türkiye'de işgücü piyasasının durumunun ve bunun nedenlerinin belirlenmesine ilişkin bir uygulama*. Yayımlanmamış Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Yükseler, Z., ve Türkan, E. (2008). *Türkiye'da hanehalkı: İşgücü, Gelir, Harcama ve Yoksulluk Açısından Analizi.* İstanbul: TÜSİAD.

Analysis of Turkey Labour Force Data by Decision Tree Methods

Extended Abstract

1. Introduction

Regardless of the developmental level of societies, unemployment and economic problems are encountered in almost every country. The governments take various social and economic measures to reduce the unemployment rate. However, the labour market, which is influenced by a wide variety of factors, changes over time. In Turkey, the Household Labour Force Survey (LFS) is carried out by Turkish Statistical Institute (TSI) in order to monitor the labour market. TSI publishes the bulletin containing the labour force statistics on the website. However, TSI does not publish the raw data obtained from the LFS. These data which is in the scope of B group microdata can be requested from the institution by obtaining the required permissions. Statistical studies on these data can help the direction of economic and social policies. This study aims to present the situation of the labour force of the population and to provide detailed information by using the year 2013 data of "Household Labour Force Survey" conducted by TSI. For this purpose, statistics are obtained for the population in the labour force, and decision tree models are created using data that belong to the persons whose labour force situation was "without work" at the previous year.

2. Method

Starting from the well-known definitions, the data mining process can be defined as a process in which classical statistical methods and machine learning methods can be used together, trying to find meaningful and useful patterns-information that cannot be foreseen in large data mass (Han & Kamber, 2001). Data mining is used in many different areas such as banking, insurance, electronic commerce, health, communication, transportation, defence, fraud detection, and education. Data mining has various methods to carry the needs of each sector. Decision trees are one of the commonly used methods. Decision trees are constituted of roots, branches, and leaves. In order to create a decision tree, a training data with a class label is required. Therefore, decision trees are defined as supervised methods. Many algorithms have been developed to create decision trees. ID3, C4.5, C5.0, "Classification and Regression Trees" (CART), "Chi-squared Automatic Interaction Detection" (CHAID) and "Quick, Interesting, Efficient, Statistical Tree" (QUEST) are the most known. In this study, the highest accuracy rates are obtained by CART and C5.0 algorithms, so only the results of these algorithms are included. In the study, the microdata of the LFS in 2013, which are conducted by TSI on a monthly basis, are used (TUIK, 2014a). Data was collected for a total of 502426 people in 146055 households in the 2013 LFS. As a result of the preliminary investigations, 15 variables (employment, labour force status in previous year, gender, relationship to reference person, age group, marital status, rural/urban area, educational attainment, school attendance, living in the same city, geographical location, activity, occupation, undergraduate major, factor) are found to be suitable for the analysis. Responses to the non-institutional population aged 15 and over (379742) are used in the study.

3. Results and Discussion

Data was collected for a total of 502426 people in the 2013 LFS. 75.58% of these people (379742 persons) are in the non-institutional population aged 15 and over. In the analysis, data belonging to persons aged 15 and over in the non-institutional population are used. In 2013, the labour force participation rate in Turkey was found to be 50.8%. This rate was 71.5% for men and 30.8% for women. The employment rate for men was 65.2%, which is much higher than that of women (27.1%). Female unemployment rate (11.9%) was higher than the male unemployment rate (8.7%). The rate of participation in the labour force is the highest with the rate of 68.9% in the 35-39 age group. While 80.2% of the higher education graduates were included in the labour force, this rate was 58.3% for secondary education and 49.8% for primary education. In the decision tree

application, a randomly selected 22206 data was used among the data of persons who stated that they were "without work" one year ago. The response variable is set as "labour force status". Thus, it is aimed to discover the characteristics of the people whose labour force status have changed or not changed in the last year.

4. Conclusion

In this study, a detailed information about the structure of the labour force in Turkey is provided using the year 2013 household labour force survey data and, decision tree models are constructed for labour force status. The results indicate that the labour structure in Turkey mainly consists of the males. Plans should be made to increase women's participation in the labour force, and the number and diversity of vocational training courses for women should be increased. Also, the ratio of the young population constituting a significant part of the country's population is not at the desired level in the labour force. In the study, the labour force status of the persons who stated that they were "without work" one year ago is tried to be modelled with CART and C5.0 algorithms. According to CART results, gender, educational attainment, continuity in the city they live in and age group are found to be the most important factors affecting the labour force status. The C5.0 algorithm results indicate that gender, age group, educational attainment, marital status, school attendance, continuity in the city they live in, and geographical location are the most important factors affecting the labour force status. This study focuses on labour force. In future studies, the characteristics of the employment can be researched by examining the distribution of employed population by sector, occupation, working time and income status.