HANEHALKI İŞGÜCÜ ARAŞTIRMA VERİLERİ İLE VERİ MADENCİLİĞİ YÖNTEMLERİNİN UYGULANMASI VE MODELLERİN KARŞILAŞTIRILMASI

IMPLEMENTATION OF DATA MINING METHODS ON HOUSEHOLD LABOR RESEARCH DATA AND COMPARISON OF MODELS

MERVE BARAN KILIÇALAN

DOÇ.DR. ÇAĞDAŞ HAKAN ALADAĞ
Tez Danışmanı

Hacettepe Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin İstatistik Anabilim Dalı için Öngördüğü YÜKSEK LİSANS TEZİ olarak hazırlanmıştır.

2018

MERVE BARAN KILIÇALAN'ın hazırladığı "**Hanehalkı İşgücü Araştırma Verileri İle Veri Madenciliği Yöntemlerinin Uygulanması Ve Modellerin Karşılaştırılması**" adlı bu çalışma aşağıdaki jüri tarafından **İSTATİSTİK ANABİLİM DALI**' nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Prof.Dr. Turhan MENTEŞ
Başkan ……

Doç.Dr. Çağdaş Hakan ALADAĞ
Danışman ……

Doç.Dr. Kerim ÖZCAN
Üye ……

Doç.Dr. Sinan SARAÇLI
Üye ……

Dr.Öğr.Üyesi Şükrü ACITAŞ
Üye ……

Bu tez Hacettepe Üniversitesi Fen Bilimleri Enstitüsü tarafından **YÜKSEK LİSANS TEZİ** olarak onaylanmıştır.

Prof. Dr. Menemşe GÜMÜŞDERELİOĞLU
Fen Bilimleri Enstitüsü Müdürü

YAYINLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanması zorunlu metinlerin yazılı izin alarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

☒ **Tezimin/Raporumun tamamı dünya çapında erişime açılabilir ve bir kısmı veya tamamının fotokopisi alınabilir.**
(Bu seçenekle teziniz arama motorlarında indekslenebilecek, daha sonra tezinizin erişim statüsünün değiştirilmesini talep etseniz ve kütüphane bu talebinizi yerine getirse bile, tezinin arama motorlarının önbelleklerinde kalmaya devam edebilecektir.)

☐ **Tezimin/Raporumun ..... tarihine kadar erişime açılmasını ve fotokopi alınmasını (İç Kapak, Özet, İçindekiler ve Kaynakça hariç) istemiyorum.**
(Bu sürenin sonunda uzatma için başvuruda bulunmadığım taktirde, tezimin/raporumun tamamı her yerden erişime açılabilir, kaynak gösterilmek şartıyla bir kısmı ve ya tamamının fotokopisi alınabilir)

☐ **Tezimin/Raporumun ..... tarihine kadar erişime açılmasını istemiyorum, ancak kaynak gösterilmek şartıyla bir kısmı veya tamamının fotokopisinin alınmasını onaylıyorum.**

☐ **Serbest Seçenek/Yazarın Seçimi**

28./06./2018

(İmza)

Öğrencinin Adı Soyadı

Merve BARAN KILIÇALAN

ETİK

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversite veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

08/06/2018

Merve BARAN KILIÇALAN

Annem'e ve Babam'a

ÖZET

HANEHALKI İŞGÜCÜ ARAŞTIRMA VERİLERİ İLE VERİ MADENCİLİĞİ YÖNTEMLERİNİN UYGULANMASI VE MODELLERİN KARŞILAŞTIRILMASI

Merve BARAN KILIÇALAN

Yüksek Lisans, İstatistik Bölümü

Tez Danışmanı: Doç. Dr. Çağdaş Hakan ALADAĞ

Haziran 2018, 81 sayfa

Günümüz koşullarında büyüyen ve genişleyen veri hacmiyle birlikte verilerin değerlendirilmesi, analiz edilmesi ve ileriye yönelik tahminlerin yapılması gibi ihtiyaçların artması veri madenciliği yöntemlerine olan yönelimi artırmıştır.

Bu çalışma kapsamında veri madenciliğinin tahmin edici yöntemlerinden sınıflama ve regresyon yöntemleri ele alınmış olup C5.0 karar ağacı, CHAID karar ağacı, Lojistik Regresyon ile Bayes Ağları yöntemleri incelenmiş ve gerçek bir veri seti üzerinde uygulanarak modelleme başarıları karşılaştırılmıştır.

Çalışmada TÜİK tarafından derlenen ve uygulanan Hanehalkı İşgücü Araştırması (HİA)' nın 2014, 2015 ve 2016 yıllarına ait verileri ele alınmıştır. Ayrıca bu çalışma kapsamında Türkiye'de işgücü durumunun belirlenmesi, hem istihdam kapsamında hem de istihdam dışında olan birey profiline ilişkin sınıflamaların yapılması ile işgücü durumu bilinmeyen bir bireyin model sonucunda işgücü durumunun doğru tahmin edilmesi amaçlanmıştır.

Çalışma sonuçları değerlendirildiğinde, çeşitli ölçütlerle yapılan kıyaslamalarda model başarıları birbirine oldukça yakın olarak edilmiş olmasına karşın C5.0 karar ağacı yöntemi

sonucunda elde edilen modelin en başarılı sınıflama tahminine sahip olduğu sonucuna ulaşılmıştır.

Anahtar Kelimeler: Veri Madenciliği, Karar Ağaçları, Lojistik Regresyon Yöntemi, Bayes Ağları, Hanehalkı İşgücü Araştırması, İstihdam

ABSTRACT

IMPLEMENTATION OF DATA MINING METHODS ON HOUSEHOLD LABOR RESEARCH DATA AND COMPARISON OF MODELS

Merve BARAN KILIÇALAN

Master of Science, Department of Statistics

Supervisor: Assoc. Prof. Dr. Çağdaş Hakan ALADAĞ

June 2018, 81 Pages

In today' s conditions, increasing the need for such as assessing, analyzing and making forward estimates of data, has increased the demand for data mining methods.

In this study, classification and regression estimating methods of data mining are discussed, C5.0 decision tree, CHAID decision tree, Logistic Regression and Bayesian Networks methods are examined and the modeling successes were compared by applying them on a real data set.

Additionally, data on the Household Labor Force Survey compiled and implemented by Turkish Statistical Institute for 2014, 2015 and 2016 years. Also, within the scope of this study to determine the labor situation in Turkey, both the employed labor force status by making regarding the classification of individuals outside employment profile is intended to be an accurate estimate of the labor situation as a result of unknown individual models.

When the study results were evaluated, it was calculated that the model obtained as a result of the C5.0 decision tree method had the most successful classification prediction even though the model successes were relatively close to each other in comparison with various criteria.

Key Words: Data Mining, Decision Trees, Logistic Regression Method, Bayesian Networks, Household Labor Force Survey, Employment

TEŞEKKÜR

Tez çalışmam süresince bana yol gösteren, bilgisini en iyi şekilde aktaran danışmanım Doç. Dr. Çağdaş Hakan ALADAĞ' a, çalışmalarım boyunca beni destekleyen Dr. Demet MECİT' e, bu süreçte görüşlerini ve yardımlarını esirgemeyen bölüm hocalarıma, mesai arkadaşlarıma ve dostlarıma teşekkürlerimi sunarım.

Bugünlere gelmem için ellerinden geleni fazlasıyla yapan ve her koşulda yanımda olan güzel yürekli anneme ve babama, başaracağıma olan inancıyla beni destekleyen halama, beni sevgisiyle her daim yüreklendiren varlığına şükrettiğim, hayattaki en büyük şansım kıymetli eşim K. Yavuz KILIÇALAN' a sonsuz teşekkürü bir borç bilirim.

İÇİNDEKİLER

Sayfa

ÇİZELGELER

ŞEKİLLER DİZİNİ

KISALTMALAR

AB	Avrupa Birliği
CART	Classification And Regression Trees
CHAID	Chi-square Automatic Interaction Detector
CRISP-DM	Cross Industry Standard Process Model for Data Mining
HİA	Hanehalkı İşgücü Araştırması
TÜİK	Türkiye İstatistik Kurumu
ENIAC	Electrical Numerical Integrator And Calculator
İBBS	İstatistiki Bölge Birimleri Sınıflaması
UNDP	Birleşmiş Milletler Kalkınma Programı
İPES	İşgücü Piyasası Enformasyon Sistemi

1. GİRİŞ

Günümüzde depolanan veri miktarında görülen artışlardan dolayı birçok ihtiyaç ortaya çıkmaya başlamıştır. Büyük miktarda veriyi depolayabilecek boyutta veri tabanlarına ve büyük boyutlardaki verinin analiz edilebilmesi için etkin yaklaşımlara gereksinim duyulmuştur. Gelişen teknolojiyle birlikte hacmi büyüyen verilerin yorumlanması, eldeki verilerden anlamlı bilgiler edinilmesi ve geleceğe yönelik tahminler yapılması amacıyla veri madenciliği kavramı gün geçtikçe önem kazanmıştır.

Veri madenciliğinin ortaya çıkması geçmiş yıllara dayansa da kullanımı günümüzde oldukça yaygınlaşmış ve popüler bir hale gelmiştir. Büyük hacimli verilerin analiz edilmesinde, değerlendirilmesinde, veri setiyle ilgili anlamlı bilgilerin ortaya çıkarılmasında büyük rol oynayan veri madenciliği yöntemleri birçok alanda başarıyla kullanılmaktadır.

Bu çalışmada, veri madenciliğinin tahmin edici yöntemlerinden çeşitli sınıflama ve regresyon yöntemleri ele alınmıştır. Bu yöntemlerin uygulanması ve karşılaştırılması amacıyla TÜİK tarafından uygulanan Hanehalkı İşgücü Araştırması (HİA)'nın 2014, 2015 ve 2016 yıllarına ait verileri düzenlenmiş ve veri setine ilişkin tanımlayıcı istatistikler elde edilmiştir. Analize hazır hale getirilen veri seti C5.0 karar ağacı, CHAID karar ağacı, Lojistik Regresyon ile Bayes Ağları yöntemleri uygulanarak analiz edilmiştir. Belirtilen yöntemlerden elde edilen sonuçlar detaylı olarak yorumlanmış, elde edilen modeller çeşitli karşılaştırma kriterleri kullanılarak kıyaslanmıştır. Yapılan karşılaştırma sonucunda, veri seti için en uygun yöntemin C5.0 Karar ağacı yöntemi olduğu görülmüştür. Veri seti için sınıflama oranları en düşük olan yöntemin ise Bayes Ağları yöntemi olduğu belirlenmiştir.

Veri madenciliği yöntemi birçok alan üzerinde uygulanmıştır. Bu yöntemlerin ele alındığı çalışmaların bir kısmı aşağıdaki şekilde özetlenmiştir:

Kara [1] çalışmasında, kadın istihdamını etkileyen değişkenleri lojistik regresyon yöntemleri ile ele almış, model sonucunda kadın istihdamını en çok etkileyen değişkenin eğitim düzeyi olduğunu belirlemiştir. Çalışmada evli kadınlar kategorisinde eğitim düzeyi değişkeninin evli olmayanlara göre daha önemli olduğunu belirlemiştir.

Altunkaya [2] çalışmasında, ülkelerin kredi derecelendirme puanlarının değerlendirilmesi amacıyla 62 ülkeye ait 5 yıllık veri setini kullanarak; bu değerlendirmeyi etkileyen faktörleri yorumlamıştır. Bu amaçla Sinir ağları, CART, CHAID, C5.0 ve Quest

yöntemlerini ele almıştır. Yöntem başarıları test edildiğinde yıllar bazında C5.0, CHAID ve CART yöntemlerinin daha başarılı olduğu sonucuna varılmıştır.

Kocabaş [3] çalışmasında, çeşitli veri madenciliği yöntemlerinin incelenmesi amacıyla karar ağaçları, birliktelik kuralları, kümeleme yöntemlerini ele almış, veri madenciliği uygulaması için internet hizmeti alan müşterilerden abonelik iptalinde bulunan müşteri profilini incelemiştir. C5.0, CHAID karar ağaçları ile lojistik regresyon yöntemlerini kullanılarak yapılan analizler sonucunda abonelik iptalinde etkili olan faktörleri değerlendirmiştir.

Yılmaz [4] çalışmasında, 2009-2010 yıllarına ait Hanehalkı işgücü anket verilerini kullanarak bu veri setine CHAID algoritmasını uygulamıştır. İşgücü durumunu en çok etkileyen değişkenin, bir yıl önceki iş gücü durumu değişkeni olduğunu belirlemiştir.

Yakut [5] çalışmasında, işletme başarılarına ilişkin doğru tahmin yapılması amacıyla C5.0 karar ağacı, destek vektör makineleri ve yapay sinir ağlarını kullanmıştır. Kullanılan çeşitli sınıflama yöntemlerinden yapay sinir ağları yönteminin, hem C5.0 hem de Destek vektör makineleri yöntemlerine göre daha iyi sonuçlar verdiği çıkarımını yapmıştır.

Kuzey [6] çalışmasında, 2011 yılında bilgi teknolojileri alanında çalışanlara uygulanan anket sonucunda elde edilen veri setinde, çeşitli değişkenlerin çalışanlar üzerindeki etkilerinin yorumlanması amacıyla karar ağaçları ve destek vektör makineleri yöntemlerini kullanmıştır. Bu yöntemlerin uygulanmasıyla elde edilen modellerin başarılarını, çeşitli kriterlere göre karşılaştırmıştır. Doğruluk oranı, eğri altında kalan alan ve en yüksek kâr karşılaştırma kriterlerine göre yorumlayarak en iyi modeli belirlemiştir.

Çakır [7] çalışmasında, banka müşterilerine ait veri kümesine sınıflama yöntemlerinden C5.0 karar ağacı, yapay sinir ağları ve lojistik regresyon yöntemlerini uygulamıştır. Uygulanan yöntemlerin karşılaştırılması amacıyla hız, ölçeklenebilirlik, kesinlik oranlarını kullanmıştır. Bu kriterler yardımıyla verilerin kesinlik kriterine göre farklılık göstermemesine rağmen hızlı karar verilme ihtiyacı sonucunda C5.0 karar ağacı yönteminin en uygun yöntem olduğunu belirlemiştir.

Kıyak [8] çalışmasında, Çapraz Endüstri Standart Süreci (CRISP-DM) döngüsü adımlarını uygulayarak Deniz Kuvvetleri Komutanlığı personelinin üniformalarındaki sipariş adımlarını incelemiştir. Bu veri setine Naive Bayes, karar ağacı ve yapay sinir ağı yöntemlerini uygulamıştır. Bu yöntemlerin başarıları doğruluk oranı, ortalama mutlak hata

ve ortalama kareler hatalarını kullanarak kıyaslamıştır. Çalışma sonucunda doğruluk oranlarına göre, veri setine en uygun modeli karar ağacı olarak belirlenmiştir.

1.1. Veri ve Bilgi

Bilgi yığınının artması ile veri ve bilgi kavramları gündeme gelmeye başlamaktadır. Veri, enformasyon ve bilgi terimleri birbirleriyle benzerlik gösteren ve sıkça karıştırılan terimlerdir. Bu terimler kısaca aşağıdaki şekilde tanımlanmaktadır [9,10,11,12]:

Veri (data): Bilgisayarlar tarafından işlem gören her türlü olgu, rakam veya metin olarak tanımlanmaktadır. Veri, araştırmalar, gözlemler, internet, sosyal medya gibi çok farklı ortamlardan elde edilen genel bir terimi ifade etmektedir.

Enformasyon (information): Veri içindeki örüntüler, birliktelikler veya ilişkiler enformasyon sağlayabilmektedir.

Örneğin, bir işletmenin satış işlemlerine ait verinin analiz edilmesi ile hangi ürünün ne zaman satıldığı enformasyonu elde edilmektedir. Bir anlam ifade eden ve kullanılabilen yapılara enformasyon olarak tanımlanmaktadır.

Bilgi (knowledge): Veriden enformasyonun oluşma süreci gibi, bilgi de enformasyondan oluşmaktadır. Bilgi; veri ve enformasyondan farklı olarak sahip olduğu bilginin sonucunda karar vericilere destek olup, hayata geçirilmesini sağlamaktadır.

Yukarıda belirtilen kavramlara bağlı olarak önem kazanan bir diğer terim veri tabanıdır. Veri tabanı, birleşik bir şema altındaki çoklu kaynaklardan alınan bilgilerin oluşturduğu bir depo olarak tanımlanabilir.

Veri Tabanı: Veri tabanı, birleşik bir şema altındaki çoklu kaynaklardan alınan bilgilerin oluşturduğu bir depodur.

Veri ambarları; Veri temizleme, veri entegrasyonu, veri dönüşümü, veri Yükleme ve periyodik veri yenileme süreci ile inşa edilmiştir.

Veri tabanı sistemi aşağıda verilen 3 ana fonksiyon ile değerlendirilmektedir [12]:

-Veri toplama ve veri tabanı oluşturma

-Veri yönetimi (veri depolama, bilgi keşfi, veri alışverişi)

-Veri çözümlemesi ve veri analizi (veri ambarı ve veri madenciliği)

1990'lı yıllardan itibaren bilgi keşfine ilişkin en önemli adımlar atılarak, veri ambarı, veri tabanı yapıları ele alınmıştır. Ortaya çıkan ihtiyaçlar neticesinde veri madenciliği kavramı

gündeme gelmiş olup veri madenciliği alanı bilgi keşfi olarak yorumlanmaya başlanmıştır [13]. Gelişen teknoloji hızına bağlı olarak kullanım imkânı sağlayan çeşitli yeniliklere ihtiyaç duyulmuştur. Her yapılan işlemden hareketle her geçen gün artan veri yığını ile verilerin tutulması, saklanması için verilerin bilgiye dönüştürmesine ihtiyaç duyulmuştur [14].

2. VERİ MADENCİLİĞİ

2.1. Veri Madenciliği Nedir?

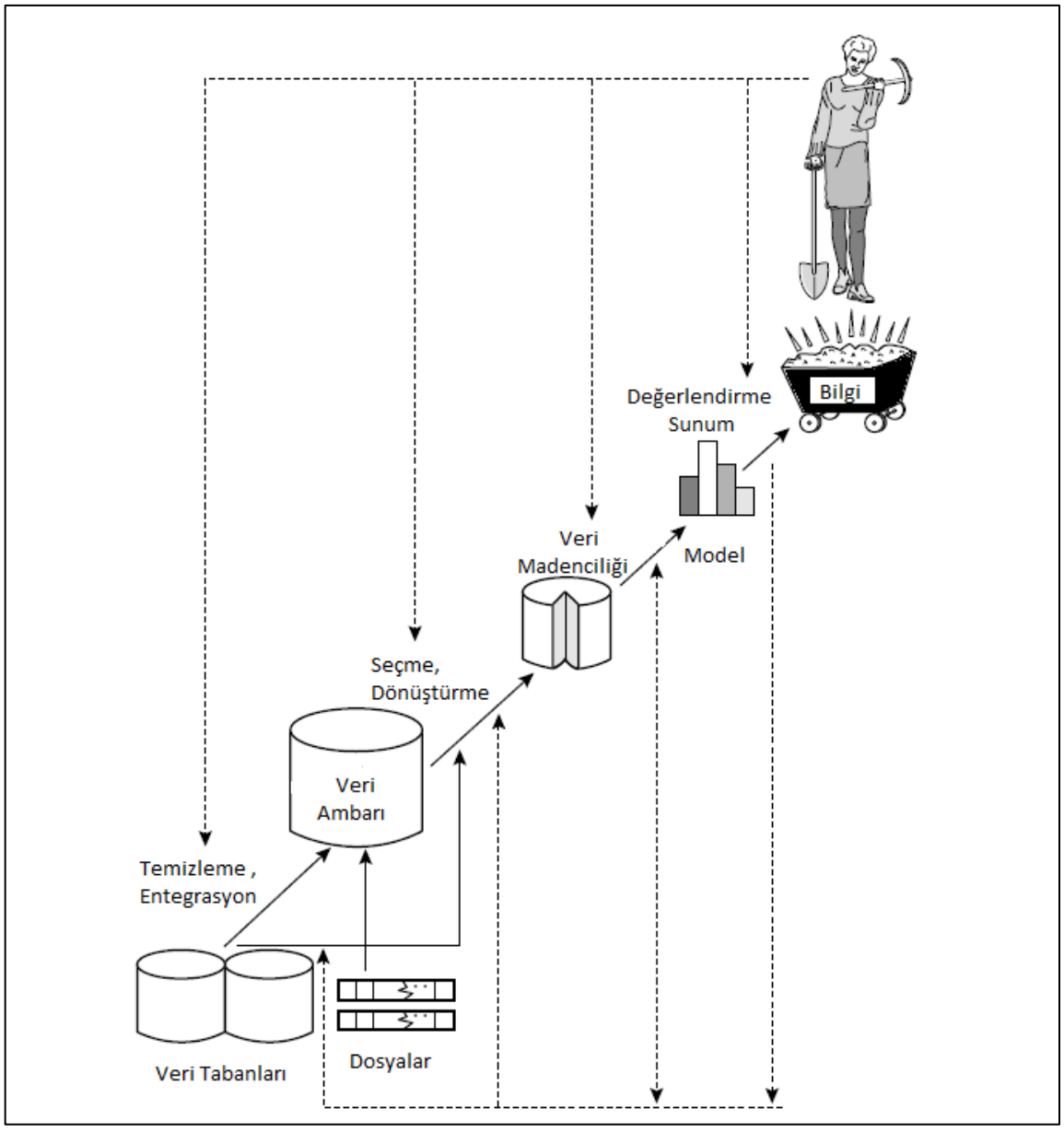
'Veri madenciliği' terimi; gizli, değerli kaynaklarını çıkaran operasyonlar anlamına gelmektedir. Mevcut veri yığını içinde daha önceden keşfedilmeyen, fark edilmeyen bilgilerin elde edilmesi olarak yorumlanmaktadır. Bilimsel araştırma açısından, veri madenciliği, esas olarak bilgisayar, pazarlama ve istatistik gibi diğer disiplinlerde yapılan çalışmalardan geliştirilen bir disiplindir. Veri madenciliğinde kullanılan metodolojilerin birçoğu, makine öğrenmede geliştirilen iki alandan gelmektedir, biri makine öğrenimine göre diğeri ise istatistiksel değerlere göre; özellikle de çok değişkenli ve hesaplamalı istatistiklerdir. Makine öğrenimi, bilgisayar bilimi ve yapay zekâyla bağlantılıdır. Makine öğreniminin amacı, veri üretme sürecinde, analistlere gözlenen veriden yeni, keşfedilmemiş vakalar türetmektir. İstatistiksel olarak da, verileri analiz etmek ve modeller oluşturmak için kullanılan yöntemler, günümüzde bunu yapmak için bilgisayarları ve programları kullanma imkânı bulunmaktadır. Veri madenciliği, bilinmeyen ve ilk kez keşfedilen ilişkileri ortaya çıkarmak amacıyla kullanıcıya net ve faydalı sonuçlar üretmeyi amaçlayan büyük boyutlu verilerin seçilmesi, araştırılması ve modellenmesi işlemlerinden oluşan bir süreç olarak tanımlanmaktadır. Veri madenciliği, yalnızca bir bilgisayar algoritması veya istatistiksel bir teknik olmayıp, kararları desteklemek için bilgi teknolojileri tarafından sağlanan bir iş zekâsı sürecidir [15].

Bilgi toplama, depolama ve veriyi işleme, giderek artan bir şekilde mevcut verilerin incelenerek anlamlı sonuçlar elde edilmesine imkân sağlamaktadır. Hangi genlerin hangi hastalığa neden olduğunu, hangi müşterilerin kredisini geri ödeyemeyeceğini veya müşterilerin bir sonraki alışverişlerinde neleri alacaklarını tahmin etmek mümkündür. Veri madenciliği, büyük hacimli veri kümelerinden faydalı, uygulanabilir ve anlamlı bilgilerin çıkarılması işlemidir [16].

Veri madenciliği oldukça önemli hale gelmiştir. Büyük verileri analiz etmek, bu analiz sonucunda daha anlamlı bilgi elde etmek ve bu bilgiyi yorumlamak gibi ihtiyaçlar veri madenciliğinin gelişimini etkilemiştir. Bu ihtiyaçların sonucunda otomatik ve akıllı veri tabanı analizi için yeni teknikler ortaya çıkmıştır. Veri madenciliği giderek önemini artıran bir alan haline gelmiştir [17].

Veri madenciliği, bilgi endüstrisinde ve toplumda, çok büyük miktarda verinin bulunması ve bu tür verilerin nitelikli bilgiye dönüştürülmesi için ihtiyaç duyulması sebebiyle son yıllarda toplumu bütün olarak etkilemektedir. Elde edilen nitelikli bilgi, market analizi, dolandırıcılık tespiti, müşteri tutma düzeyinden, üretim kontrolü ve bilim araştırmasına kadar uzanan alanlarda kullanılabilmektedir [18].

Veri madenciliği; Şekil 2.1' de adım adım özetlenmektedir. Veri madenciliği birbirine bağlı adımlardan oluşan geniş bir süreçtir [12]:



Şekil 2.1. Bilgi keşfi sürecinde bir adım olarak veri madenciliği [12]

Şekil 2.1' de bilgi keşfini oluşturan süreç; verinin temizlenmesi, verinin bütünleştirilmesi, veri seçimi, veri dönüştürme, veri madenciliği yöntemlerinin uygulanması, örüntü değerlendirme, sonuçların yorumlanması ve sunum aşamalarından oluşmaktadır. İlk 3 aşama veri önişleme aşamasıdır [12,14]:

-Veri Temizleme: Eksik, hatalı, aykırı değerler analiz aşamasında sıkıntı olmaması açısından bu aşamada bu verilerden temizleme yapılmaktadır.

-Veri Bütünleştirme: Birden fazla veri kaynağının bulunduğu durumda çalışmanın yapılması amacıyla verilerin tek bir düzene getirilmesidir.

-Veri Seçimi: Analiz için gerekli veriler veri tabanından alınıp, ilgili veriler indirgenebilir, değişken sayıları artırıp eksiltilebilmektedir.

-Veri Dönüştürme: Veri azaltma işlemi, orijinal veri bütünlüğüne zarar vermeden daha küçük yapılar haline getirilmesidir. Veri madenciliği için, özet veya toplama işlemi yapılması için, verilerin dönüştürülmesi veya uygun yapılara dönüştürülmesi işlemidir.

-Veri Madenciliği: Veri örüntülerinin ve çeşitli yöntemler kullanılarak elde edilmesi işlemidir. Bu aşamada, veri incelenerek veri yapısına uygun olan sınıflama, kümeleme gibi yöntemlerden hangisinin ele alınacağına karar verilip veri setine uygun olan algoritma kullanılmaktadır.

-Örüntü Değerlendirme: Belirli ölçütlere dayanarak, bilgiyi temsil eden gerçekten ilginç kalıpları tanımlamak amacıyla yapılan değerlendirme işlemidir.

-Sonuçların Yorumlanması ve Sunum: Veri madenciliği algoritmaları uygulandıktan sonra elde edilen nitelikli bilginin düzenlenmesi, yorumlanması işlemidir.

2.2. Veri Madenciliği Tarihçesi

Veri madenciliğinin başlangıcı ilk sayısal bilgisayar olan Elektronik sayısal entegreli hesaplayıcı (ENIAC-Electrical Numerical Integrator And Calculator)' ya kadar gitmektedir. 1946 yılında geliştirilen ilk bilgisayardan günümüze "ilk" bilgisayarın yıllar içerisinde geçirmiş olduğu değişim görülmektedir. İstek ve ihtiyaçlar doğrultusunda geliştirilen ürünler, zamanla şekillenerek son halini almış durumdadır. Donanımın geliştirilmesinin ardından yazılım bu donanıma uyarlanarak kullanıcıya ulaştırılmaktadır. İhtiyaçlar doğrultusunda yazılımda bulunan eksiklikler belirlenip, bu eksiklikler göz önünde bulundurularak yeni yazılımlar geliştirilmektedir. İhtiyaçlar doğrultusunda

şekillenen veri tabanları ve veri modelleme çeşitleri hızla yaygınlaşırken, donanımlar da benzer gelişimi göstermiştir. Gün geçtikçe büyüyen veri tabanlarının düzenlenmesi, kullanılması zorlaştığından veri modelleme kavramı ortaya çıkmıştır. Veri madenciliği kavramı 1960' lı yıllarda, veri analiz problemlerini çözmek için kullanılmaya başlanmıştır. O dönemlerde, bilgisayar yardımıyla, yeterince uzun bir tarama yapıldığında, istenilen verilere ulaşmanın mümkün olacağı görülmektedir. Veri madenciliğine istatistik, makine öğrenimi, veri tabanları gibi disiplinler ve kavramlara dayanan çeşitli yaklaşımlar getirmeye başlanmıştır [19].

1980'lerin ortalarından itibaren veri tabanı teknolojisi, ilişkisel teknolojinin benimsenmesi ve araştırma geliştirme çalışmalarının artması ile daha güçlü bir veri tabanı üzerine nitelendirilmiştir. Bunlar, genişletilmiş ilişkisel, nesne yönelimli, nesne ilişkisel ve tümdengelim modeller gibi gelişmiş veri modellerinin geliştirilmesini teşvik etmektedir. Gün geçtikçe aktif, algılayıcı, bilimsel ve mühendislik veri tabanları, bilgi tabanları ve ofis bilgi tabanları da dâhil olmak üzere uygulama odaklı veri tabanı sistemleri gelişmiştir. Verilerin dağıtımı, çeşitlendirilmesi ve paylaşımı ile ilgili konular kapsamlı olarak incelenmiştir [12].

2.3. Veri Madenciliği Kullanım Alanları

Veri madenciliği, birçok alanda etkili bir alandır. Son yıllarda depolanan anlık verilerin büyük boyutlara ulaşmasıyla, tutulan verileri kullanabilecek, verilerden çıkarımlar yapılabilecek ve yararlı bilgiler sağlayabilecek yöntemlere duyulan ihtiyaç sebebiyle veri madenciliği yöntemlerinin kullanımı çeşitli alanlarda yaygınlaşmıştır. İlk uygulamaları, özellikle pazar sepeti analizi şeklinde, perakendecilik sektöründe olmuştur. Bununla birlikte birçok sektörde, birçok alanda ihtiyaç duyulan bir yöntem olduğundan yaygın bir şekilde kullanım göstermektedir [12,18]:

Perakendecilik sektöründe,

-Ürün konumlaması

-Satış tepkileri

-Çapraz satış hareketleri

-Müşteri sadakat analizi

-Satış kampanya takipleri

Bankacılık sektöründe,

-Müşteri ilişkileri yönetimi

-Kredi hesap kontrolü

-Kredi ödeme tahminleri

-Dolandırıcılık tespiti

-Kredi kart yönetimi

Sigortacılık sektöründe,

-Sigorta dolandırıcılık tespiti

-Riskli müşteri analizi

Telekomünikasyon ve e-ticaret uygulamaları,

-Hileli kalıp analizi

-İlişkili örüntü analizi

-Büyük verilerin analizi

-Olağandışı hareketlerin tespiti

Sağlık, tıp alanında,

-DNA veri analizi

-Hastalıkların teşhis edilmesi

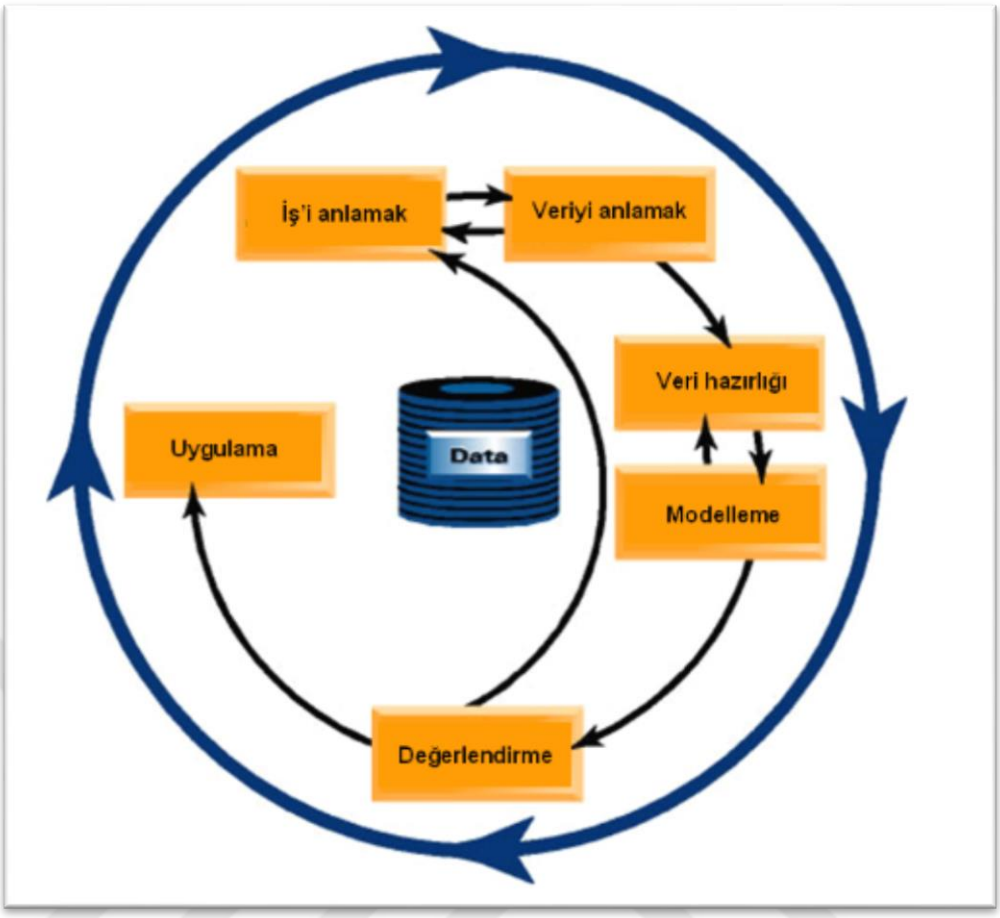
-Gen haritaları analizi

2.4. Veri Madenciliği Süreci

Veri madenciliği analizinde genel bir süreç işlenmektedir. Standart olarak kullanılan 2 süreç bulunmaktadır, biri CRISP olarak adlandırılan genellikle veri madenciliği çalışmasında katılan adımlar dizisinin süreçleri kullanılmaktadır, diğeri ise SEMMA olarak adlandırılan SAS programına ait süreçlerdir. Genellikle kullanılan CRISP döngüsü altı aşamadan oluşmaktadır. Veri araştırması, veri toplama, veri işleme, analiz, yapılan çıkarımlar ve uygulama bu süreçte uygulanan adımlardır [18]:

1. İşin anlaşılması: İş hedeflerinin belirlenmesi, mevcut durumun değerlendirilmesi, veri madenciliği hedeflerinin belirlenmesi ve bir proje planı geliştirilmesi işlemlerini içermektedir. Veri madenciliği araştırmasının temel unsuru, çalışmanın ne için yapıldığının belirlenmesidir.

2. Verinin anlaşılması: İş hedefleri ve proje planı belirlenirken, veri gereksinimleri göz önünde bulundurulmalıdır. İlk adım veri toplama, veri tanımlama, veri arama ve veri kalitesinin doğrulanmasıdır. Bu aşamada veri açıklama, veriyi istatistiksel olarak görme işlemi yapılmaktadır. Veri madenciliği işleminin ilk aşaması, iş görevleri ile ilgili belirli bir bilgiyi doğru bir şekilde tanımlamak için mevcut birçok veri tabanından ilgili verileri ele almaktır.
3. Verinin Hazırlanması: Mevcut veri kaynakları belirlendikten sonra veri seçilmeli, temizlenmeli, istenilen şekle getirilmeli ve biçimlendirilmelidir. Veri modelleme hazırlığında veri temizleme ve veri dönüşümü bu aşamada gerçekleşmektedir. Bu aşamada uygulanan ve ek modeller uygulanması sebebiyle, iş anlayışına dayalı kalıpları görme fırsatı ortaya çıkmaktadır. Verilerin önişleminin yapılmasının amacı, seçilen verilerin temiz ve kaliteli olması için hazırlamaktır.
4. Modelleme: Veri madenciliği yazılım araçlarından görselleştirme (veri ve ilişki kurma) ve kümeleme analizi (hangi değişkenler birlikte daha ilişkili ise) ilk analiz için yararlıdır. Veri kavrandıktan sonra veri tipine uygun daha ayrıntılı model ve kurallar geliştirilmektedir. Verilerin türüne bağlı olarak, çeşitli modeller uygulanabilir.
5. Değerlendirme: Model sonuçları ilk olarak işletme hedefleri göz önünde tutularak, değerlendirilmelidir. Veri yorumlama aşaması oldukça önemlidir.
6. Uygulama: Veri madenciliği hem daha önce düzenlenenleri doğrulamak için beklenmedik hipotezler olarak, hem de bilgi keşfi için kullanılabilmektedir. Önemli durumların tahmini ve tanımlanması dâhil olmak üzere işletme faaliyetlerinde birçok amaç için kullanılmaktadır. İş anlayışını kazanmak, veri madenciliği için, çeşitli görselleştirme sonuçları, istatistiksel ve yapay zekâ araçları kullanıcıya yeni ilişkiler kurmasını sağlayan bir süreçtir.



Şekil 2.2. Veri madenciliği döngüsü [20]

2.5. Veri Madenciliği ve İstatistik

Veri madenciliği yöntemlerinin kullanılmasında işlemlerin uygulama alanı oldukça geniştir. Birçok disiplinle yakından ilişkili çalışma prensiplerine sahiptir. Bu alanlar içerisinde Şekil 2.2' deki gibi, istatistik, veri tabanı sistemleri, veri görselliği, yapay sinir ağları, makine öğrenmesi gibi disiplinlerle ilişkisi bulunmaktadır. Ayrıca kullanılan veri madenciliği yaklaşımına bağlı olarak, sinir ağları, bulanık küme teorisi, mantık programlama veya yüksek performanslı bilgi işlem gibi diğer disiplinler de kullanılmaktadır. Veri türlerine veya verilen veri madenciliği uygulamasına bağlı olarak veri madenciliği sistemleri, aynı zamanda, uzamsal veri analizi, bilgi algılama, model tanıma, görüntü analizi, sinyal işleme, bilgisayar grafikleri, web teknolojisi, ekonomi, işletme, biyoenformatik ve psikoloji gibi teknikleri de bütünleştirmektedir [12].

Şekil 2.3. Veri madenciliği ve disiplinler

Veri madenciliği yöntemlerinde kullanılan yönteme göre ilişkili olduğu alanlar da oldukça çeşitlidir. İstatistik, veri tabanı sistemleri, makine öğrenimi başta olmak üzere zengin bir alt yapıya sahiptir.

Bu disiplinler içinde istatistik biliminin yeri diğerlerinden daha farklıdır. İstatistik tabanlı veri madenciliği yöntemleri ortaya çıkmıştır. Veri madenciliğinde, istatistiksel süreçler uygulanmaktadır. Veri madenciliği ve istatistiğin ortak noktası, karışık verilerden ilginç, yorumlanabilir bilgiler elde etmektir. En büyük farkları, veri madenciliğinin bilgisayar teknolojisi ve birtakım algoritmalar sayesinde büyük verilerle çalışmasıdır [21].

3. VERİ MADENCİLİĞİ YÖNTEMLERİ

3.1. Veri Madenciliği Stratejileri/Fonksiyonları

Veri analizi, verilerin düzenli bir veri tabanında düzenlenmesini gerektirmektedir. Verilerin analiz edilme şekli, verilerin veri tabanında nasıl organize edildiği konusunda yardımcı olmaktadır. Giderek artan boyutlu veriler ve bununla birlikte verileri analiz etme ihtiyacı ortaya çıkmaktadır [15].

Veri madenciliği yapılacak şekilde, birçok türde veri tabanının, veri depolarının bulunduğu bilinmektedir. Veri madenciliği fonksiyonları, verinin türüne, desen türlerine göre belirtmek için kullanılmaktadır. Veri madenciliğinin genel olarak görevleri tanımlayıcı (description) ve tahmin edici (prediction) olmak üzere 2 gruba ayrılmaktadır. Tanımlayıcı özellikteki işlemciler genel özellikleri karakterize etmektedir, tahmin edici işlemciler ise tahmin yapmak için mevcut veriyi kullanmaktadır. Bu sebeple birkaç farklı desen aramak istenebilmektedir. Bu nedenle, veri madenciliği sistemi farklı kullanıcı taleplerini veya uygulamalarını barındıracak şekilde tasarlanmaktadır. Ayrıca, veri madenciliği sistemleri, çeşitli şekillerde desenler keşfedebilecek biçimde kurgulanması gerekmektedir [12].

3.2. Veri Madenciliği Yöntemleri

Veri madenciliği yöntemleri;

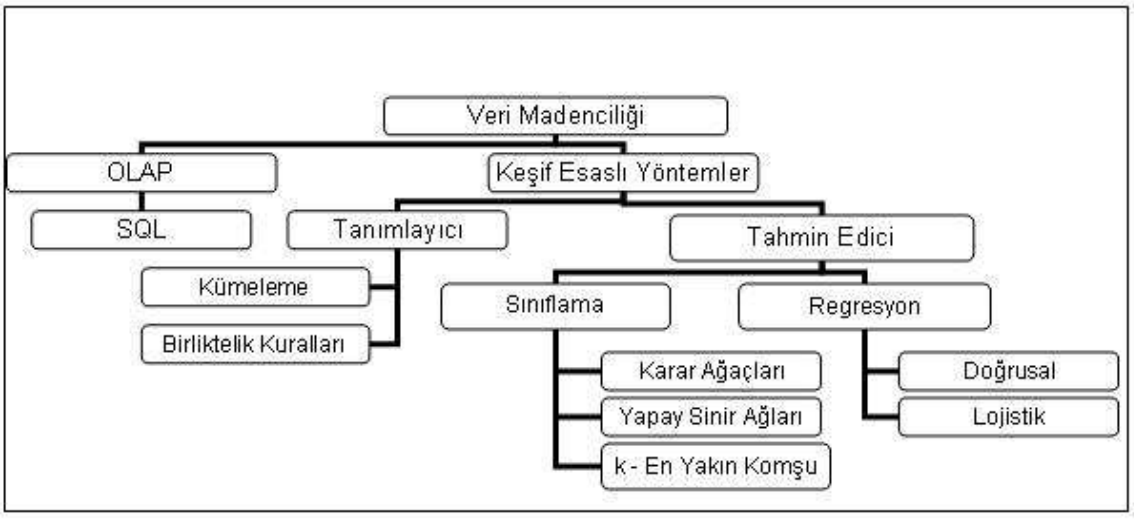
-Sınıflama (Classification) ve Regresyon (Regression) Yöntemleri,

-Kümeleme (Clustering) Yöntemleri

-Birliktelik Kuralları (Association Rules) ve Ardışık Zamanlı Örüntüler (Sequential Patterns) olmak üzere üç ana başlık altında incelenmektedir [22].

Veri madenciliği yöntemleri tanımlayıcı ve tahmin edici olmak üzere, sınıflama ve regresyon modelleri tahmin edici, kümeleme, birliktelik kuralları ve ardışık zamanlı örüntü modelleri tanımlayıcı modelleri olarak ayrılmaktadır [22].

Veri madenciliği yöntemleri Şekil 3.1'de özetlenmektedir.



Şekil 3.1. Veri madenciliği modelleri [23]

Tahmin edici yöntemlerde; sınıflama, veri sınıflarını, modeli tahmin etmek ve sınıf etiketi bilinmeyen nesneleri sınıflamak için tanımlama ve ayırma işlemi yapan modeli bulma işlemlerinden oluşmaktadır. Üretilen model, bir eğitim verisinin (training data) yani sınıf etiketi bilinen veri nesnelerinin analizine dayanmaktadır. Regresyon analizi, istatistiksel bir metodolojidir. Mevcut verilere dayanan dağılım eğilimlerinin tanımlanmasını sağlamaktadır. Kullanılan veri madenciliği yaklaşımına bağlı olarak, sinir ağları, bulanık ve kaba küme teorisi gibi diğer disiplinler uygulanmaktadır. Bilgi gösterimi, tümevarımsal mantık programlama veya yüksek performanslı bilgi işlem disiplinleri de uygulanmaktadır [12].

Sınıflandırma görevi, eğitim kümesi olarak adlandırılan, her kaydın bir öznitelik kümesi ve niteliklerden oluşan kayıt sınıfını belirtmektedir. Hedef, sınıf için bir model bulmaktır. Model daha önce gözlenmemiş kayıtlar için sınıfın niteliğini tanımlamada kullanılmaktadır [24].

Sınıflama ve regresyon modellerinde kullanılan başlıca sınıflandırma teknikleri aşağıdaki gibidir [22]:

· Yapay Sinir Ağları (Artificial Neural Networks),

· Genetik Algoritmalar (Genetic Algorithms),

· K-En Yakın Komşu (K-Nearest Neighboor),

· Bellek Temelli Nedenleme (Memory Based Reasoning),

· Naïve-Bayes,

· Lojistik Regresyon

· Karar Ağaçları (Decision Trees),

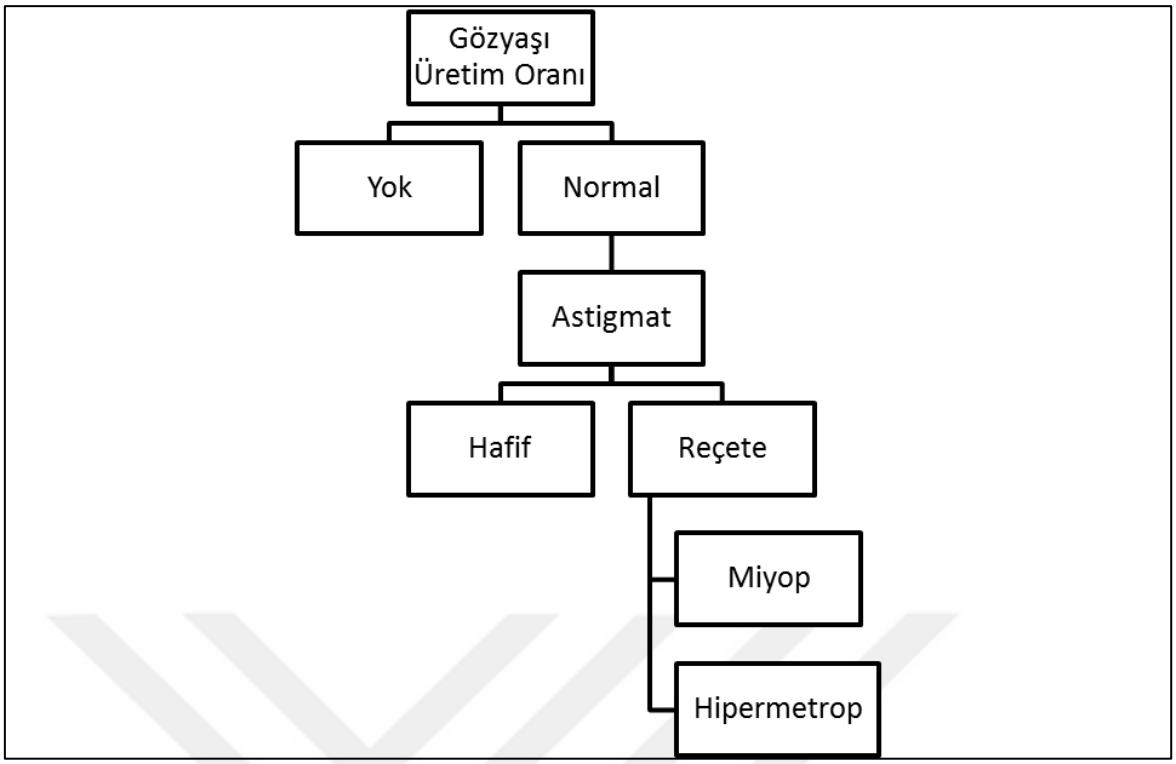
Veri madenciliği yöntemlerinden sınıflama ve regresyon yöntemleri ele alınmış, özellikleri kısaca anlatılmıştır.

3.2.1. Karar Ağaçları

Karar ağaçları, veri madenciliğinde kullanılan temel, önemli yöntemlerden biridir. Bu yöntem, en iyi tahmini yapabilmek için bağımlı ve bağımsız değişkenler arasındaki tüm ilişkilerin araştırılmasını yönetmektedir. En kuvvetli ilişkiye sahip değişken belirlendiğinde veri kümesi bu değişkenin değerlerine göre ayrılmakta, potansiyel bölünmeler bitene kadar sürmektedir. Diğer bir ifadeyle, kök düğümünden başlayıp yukarıdan aşağıya kadar düğümler oluşturarak takip edilen bir yöntemdir [25].

Karar ağaçları, sonuçların yorumlanmasında kolaylıklar sağlaması, veri tabanı sistemleri ile kolayca bütünleştirilebilmesi ve yüksek seviyede güvenilirliklerinin olmasından dolayı veri madenciliğinin sınıflama modelleri arasında sık kullanılan bir algoritmadır. Karar ağaçları, özellikle pazarlamada, bireylerin kredi geçmişlerini kullanarak kredi skorlarının hesaplanması, geçmişte işletmeye kar getiren bireylerin özelliklerini kullanarak işe alımlarda kullanılması, satışları etkileyen değişkenlerin belirlenmesi gibi faaliyetler için önemli ve kolay uygulanabilir bir yöntem olarak uygulanmaktadır [26].

Karar ağacı, gerçekleştirilecek testi belirtmektedir. Ağacın veri kaybetmeden dallara bölünmesini sağlamaktadır. Her düğümde test ve dallara bölünme peş peşe gerçekleşmektedir ve bu bölünme işlemi üst seviyedeki ayrımlara bağlıdır. Dalın ucunda sınıflama olayı gerçekleşmiyorsa, o dalın sonucunda bir karar düğümü oluşmaktadır. Dalın sonunda belirli bir sınıf oluşuyorsa, o dalın sonunda yaprak oluşmaktadır. Karar ağacı işlemi kök düğümünden başlayıp yukarıdan aşağıya doğru yaprağa ulaşana kadar takip eden düğümlerle gerçekleşmektedir [27].



Şekil 3.2. Kontakt lens verisi için karar ağacı örneği [28]

Şekil 3.2' deki kontakt lens verileri için oluşturulmuş bir ağaç yapısını göstermektedir. Ağaç yapısında öncelikle gözyaşı üretimi için deneme yapılmaktadır, 2'ye ayrılan dal yapısı 2 olası sonuca karşılık gelmektedir. Soldaki dal, gözyaşı üretim hızı azalırsa sonuç olarak bir yere gitmediğini göstermektedir. Eğer normalse (sağ dal), ikinci kez astigmatizm üzerine test yapılmaktadır. Sonunda, testlerin sonucu ne olursa olsun, bu durum için kontakt lens önerisini belirleyen bir ağacın yaprağına ulaşılmaktadır.

Karar ağaçlarının oluşturulmasında ağacın kökten sonraki bölünmesinin hangi kritere göre yapılacağının belirlenmesi oldukça önemlidir. Ele alınan kritere göre karar ağacı algoritması da değişiklik göstermektedir. Bu algoritmalar şu şekilde sınıflandırılabilmektedir [14,21]:

- Entropi içeren algoritmalar: ID3, C4.5, C5.0 algoritmaları

- Sınıflandırma ve regresyon ağaçları (CART): Twoing, Gini algoritmaları

- Bellek tabanlı sınıflandırma algoritmaları: En yakın k-komşu algoritması

- İstatistik bazlı algoritmalar: Bayesyen sınıflandırması, CHAID

Karar ağaçlarının dallara bölümlenmesinde kullanılan kriterlerden bilgi kazanımı (information gain) , entropi (entropy), kazanım oranı (gain ratio), gini indeksi (the gini index), ki-kare tablosu istatistiği değerleri kullanımına göre ağaç yapıları farklılaşmaktadır [7].

Bahsi geçen algoritmalar aşağıda kısaca açıklanmıştır:

3.2.1.1. ID3 algoritması

1970' lerin sonu ve 1980'lerin başında, makine öğreniminde araştırmacı olan J. Ross Quinlan tarafından, ID3 (İteratif Dichotomiser) olarak bilinen bir karar ağacı algoritması geliştirilmiştir. E. B. Hunt, J. Marin ve P. T. Stone tarafından Kavram öğrenme sistemleri üzerine yapılan daha önceki çalışmalar genişletilmiştir. Quinlan'ın karar ağacı indüksiyonu araştırması ön plana gelmiştir. ID3 algoritması, bilgi kazanma ölçütü kullanılarak yapılan bir yöntemdir. Bilgi kazanma ölçütlerinin kullanılması, ID3' e yapılan iyileştirmelerden en önemlisidir [12,28].

3.2.1.2. CART Algoritması

Çok sayıda değişken arasından seçim yapabilen parametrik olmayan bir teknik olmakla beraber bunların açıklanacak sonuç değişkeninin belirlenmesinde en önemli tekniklerden biridir. Bağımlı değişken kategorik ise, CART yöntemi sınıflandırma ağacı üretmektedir, bağımlı değişken sürekli ise CART yöntemi regresyon ağacı oluşturmaktadır [29].

CART, Sınıflandırma ve Regresyon Ağacı, her bir iç düğümünde dal bulunan, ikili bir karar ağacı algoritmasıdır. CART algoritması çeşitli kriterlerle çalışmaktadır. CART fonksiyonunu diğer makine öğrenme algoritmalarından ayıran kullandığı budama mekanizmasıdır. CART, budama işlemini hem seçilen ağacın büyüklüğü için hem de doğru sınıflama yaparak doğru tahminlerde bulunmak için yapmaktadır. Budama kriteri olarak Gini katsayısı gibi çeşitli kriterler kullanmaktadır [30].

3.2.1.3. C4.5 Algoritması

Oldukça fazla simge içeren ve ID3 algoritmasına göre daha görsel olan C4.5 algoritması, 1970' lerin sonunda J. Ross Quinlan tarafından geliştirilmiştir.1990' ların başında C4.5' in tam olarak Quinlan' ın geliştirdiği açık kaynak kodlu kitap ile anlaşılabilir hale gelmiştir. ID3 algoritmasının gelişmiş versiyonu olarak ortaya çıkan C4.5 algoritması ID3 ' e kıyasla farklı ve daha yeni öğrenme algoritmalara sahiptir [12].

3.2.1.4. C5.0 Algoritması

C5.0 karar ağacı algoritması popüler olarak kullanılan C4.5 algoritmasının sonrasında uygulamaya konulmuştur. C5.0 algoritması, C4.5 algoritmasının kullandığı karar ağacı indüksiyonuyla aynı gibi görünse de testler bazı farklılıkları ortaya koymaktadır. Bununla birlikte, kural üretme hızı artmasa da farklı bir teknik kullandığı gözlemlenmektedir [28].

C5.0 karar ağacı algoritmasında sınıflama işlemi yapılırken bilgi kazanımı ölçütleri ve entropiler yardımıyla ayırım yapmaktadır. Hesaplanan ölçütler ile her değişken için belirli değerler dallara ayrılmaktadır. Bölümlenecek seviyeye gelen kadar işlem sürüp, bölünecek yeni bir düğüm kalmayana kadar devam etmektedir [7].

Bu ölçüye göre belirlenen değişkenin her bir değeri dallara dönüşmekte ve devam eden süreçte kalan değişkenlerin dikkate alınması ile aynı işlem sürdürülmektedir.

C5.0 karar ağacı algoritması tarafından oluşturulan ağaçlar CART algoritması tarafından oluşturulanlara benzemektedir CART' den farkı, C5.0 algoritması kategorik değişkenler üzerinde çoklu bölünmeler yapılmasıdır. CART gibi, C5.0 algoritması da ilk önce bir budanmamış ağacı büyütmesi ve daha kararlı bir model oluşturulması için onu geri atmaktadır. Fakat C5.0 algoritmasının budama stratejisi oldukça farklıdır. C5, alt seçimler arasından seçim yapmak için bir doğrulama setinden yararlanmamaktadır; Ağacın büyütülmesi için kullanılan veriler aynı zamanda ağacın nasıl budanması gerektiğine karar vermek için de kullanılmaktadır. C5.0, bir yaprakta görülmesi muhtemel olan en kötü hata oranını tahmin etmek için istatistiksel örnekleme metodolojisi kullanılmaktadır. Yaprakta bulunan veriler, her biri iki olası sonuçtan birine sahip olabilecek bir dizi denemenin sonuçlarını temsil edecek şekilde çalışmaktadır [31].

3.2.1.5. CHAID (Chi-Squared Automatic Interaction Detector) Algoritması

Bağımlı değişken ile bir ya da birden fazla sayıdaki bağımsız değişkenler arasındaki ilişkinin incelenmesi, bağımlı değişkene ait değerlerin en iyi şekilde öngörülmesi için kullanılan yöntemlerden biridir. Bir olayı bağımsız değişkenlerle, olası alt gruplara ayırarak bu grupların ayrıntılı biçimde incelenmesiyle aralarındaki ilişkiye ilişkin doğru yorumlamalar yapılmasını sağlamaktadır. Yöntem olarak CHAID yöntemi, daha avantajlı ve daha fazla kolaylık sağladığından CART ve QUEST yöntemlerine göre daha fazla yaygınlaşmıştır [32].

CHAID yönteminde; bağımsız değişkenlerle bağımlı değişkenler için çapraz tablo oluşturularak bağımsız değişkeni ifade eden anlamlılığı en düşük olan kategoriler bulunarak birleşmeleri anlamlı bulunan iki kategori gruplanmaktadır. Bu adım bağımsız

değişkenin kendi içindeki birleşmeleri zayıflayıncaya kadar devam etmektedir. Üç veya daha fazla sayıda kategoriye sahip olan kategorilerin her biri için iki bölünme bulunmaktadır. Anlamlılığın kritik değerden düşük kaldığı durumda, ikinci adım tekrarlanmaktadır [33].

İstatistiksel olarak, değişken değerleri kategorik olduğunda ve bağımsız değişkenler ile kategorik olarak ölçülebilen sonuçlar arasında ilişki gerektiren durumlarda oldukça yararlı sonuçlar üreten CHAID yöntemi, ki-kare parametrik olmayan istatistik yöntemini kullandığından CHAID, kesimleri önemli bir ki-kare tarafından gösterilen bir bağımlılık ilişkisinin bir sonucu olarak yapılandıran değişkenlerin geri kalanıyla ilişkili bir kriter değişkeni oluşturan tahmini bir analiz oluşturmaktadır [34].

Chaid yöntemi, kullandığı ki-kare p değerinden dolayı algoritma adını almaktadır. Chaid yönteminde etkili olan çeşitli nitelikler algoritmanın öne çıkmasını sağlamaktadır. Her düğümde optimal bölünmeyi sağlayacak tahmin ediciler belirlemekte ve p değerleri ile Bonferroni değerini ayırma kriteri olarak kullanmaktadır [35].

3.2.2. Bayesyen Sınıflandırması

Bayes Sınıflandırıcısı, Bayes teoremini temel alan bir olasılıksal sınıflandırma yöntemidir. Sınıflanmış verileri kullanarak yeni bir verinin elimizdeki sınıflardan herhangi birine ait olma ihtimalini hesaplayan bir yöntemdir. Bu sınıflandırma yönteminde nitelikler birbirinden bağımsız olarak kabul edilmektedir. Her biri n adet nitelikten oluşan ve m adet sınıftan herhangi birine dâhil olan bir veri setinin olduğu bir durumda, hangi sınıfa ait olduğu bilinmeyen yeni bir X örneği için, Bayes eşitlikleri kullanılarak o sınıfa ait olma olasılığı hesaplanmaktadır. En yüksek olasılığa sahip olan sınıf örneğin ait olduğu sınıf olarak kabul görmektedir [36].

Sınıflandırmada yaygın olarak kullanılmakta olan uygulanabilirliği ve performansı ile kolaylık sağlayan bir yöntemdir. Bayes yönteminin uygulanmasında çeşitli kabuller yapılmaktadır. Bunlardan en önemlisi niteliklerin birbirinden bağımsız olması durumudur. Eğer nitelikler birbirini etkiliyorsa burada olasılık hesaplamak zorlaşmaktadır [37].

Bayes teoremini oluşturan olasılık fonksiyonu;

$$P(\theta/X) = \frac{P(X/\theta)P(\theta)}{P(X)} \tag{3.1}$$

olarak tanımlanmaktadır [38].

P(X/θ); gözlemlerin bağımsız ve eşit olarak dağılmış olduğu varsayımından türetilen bir koşullu fonksiyon, $P(\theta)$ önceki dağılımı, $P(X)$ ise önsel (marjinal) dağılım olarak adlandırılmıştır. P(X/θ)' nın gözlemlerin birbirinden bağımsız olduğu varsayımından dolayı koşullu fonksiyon aşağıdaki şekilde hesaplanmaktadır [38]:

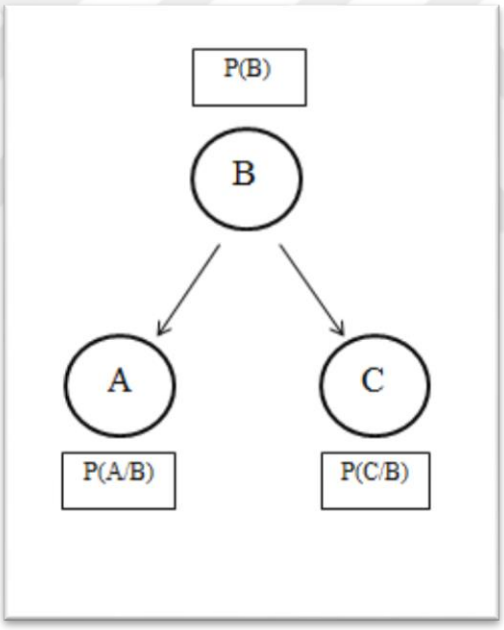
$$P(X/\theta)=\prod_{i=1}^{n} f(x/\theta) \quad (3.2)$$

Ayrıca 2' den fazla değişkeni bulunan modeller için aşağıdaki şekilde hesaplanmaktadır [39]:

$$P(\theta/X,Y) = \frac{P(\theta)\, P(X/\theta)\, P(Y/\theta,X)}{P(X)P(Y/X)} \quad (3.3)$$

Bayes ağları ise, grafiksel olarak bayes ağı kurularak sınıflama yapılabilmektedir.

Şekil 3.3' te A ve C' nin B' den koşullu olarak bağımsız olduğu varsayılan basit bir bayes ağı gösterilmektedir [14]:



Şekil 3.3. Bayes ağ yapısı

Bayes ağları, yönlendirilmemiş kenarlı ağları (Markov ağları) içeren grafiksel modeller olarak adlandırılan istatistiksel modellerin daha geniş bir sınıfının özel bir örneğidir. Bayes ağları ile bir öğrenme algoritması kurmak için; verilere dayalı olarak belirli bir ağın değerlendirilmesi için bir işlev ve olası ağların alanını araştırmak için bir yöntem olmak üzere iki bileşeni tanımlanması gerekmektedir. Oluşturulan ağın kalitesi; ağa verilen ağ

olasılıklarıyla ölçülmektedir. Ağın her örneğe uyum sağlama olasılığı hesaplanmaktadır. Ortaya çıkan olasılıkların logaritmaları kullanılmaktadır. Ortaya çıkan nicelik, verilerin logaritmik olasılığıdır. Ağ yapısında ortaya çıkan düğümler, her değişken için 1 tane olmak üzere, önceden belirlenmektedir. Ağ yapısını öğrenmek, her küme için olasılık tablolarını tahmin etmek ve olasılıklarını hesaplamakla işlemektedir. Bayesyen ağ yapılarını öğrenme algoritmaları, ağ yapılarının aranmasında çeşitli farklılıklar göstermektedir. 2 tip Bayes ağları sınıflayıcısı bulunmaktadır. İlki Markov ağ yapısıdır, bu ağ tüm ebeveynleri, çocukları ve çocukların ebeveynlerini içeren bir yapıya sahiptir. Bir düğümün, Markov battaniyesindeki düğümler için verilen değerlerle diğer tüm düğümlerden koşullu olarak bağımsız olduğu gösterilebilmektedir. Markov ağ yapısında, sınıflamayla ilgisi olmadığı düşünülen, ilgisiz ilişkiler model dışında kalmaktadır. Bir diğer öğrenme algoritması ise Ağaç Artırılmış Naive Bayes (TAN) olarak adlandırılan ağ yapısıdır. Bayes ağının her bir düğümü tek bir ebeveyndir. Sınıf düğümü ve ikinci bir ebeveyn eklenmediği, yalnızca 1 düğüm olduğu varsayılırsa ortaya çıkan sınıfın ebeveyni olmayan bir düğümde kökleştiği bir ağaç yapısı oluşmaktadır. Sınıflandırmada, en üste çıkarılmak istenen, sınıfın diğer özellik değerlerinin aldığı koşullu olasılıktır [28].

3.2.3. Lojistik Regresyon Yöntemi

Değişkenler arasındaki ilişkilerin çalışılmasında regresyon yöntemleri sıklıkla kullanılmaktadır. Regresyon analizinde öncelikle, bağımlı ve bağımsız değişkenlerin niteliklerinin belirlenmesi gerekmektedir. Bağımlı değişkenin ölçülebilir nitelikte olmadığı durumlarda lojistik regresyon yöntemleri kullanılmaktadır. Bu yöntem uygun ve kabul edilebilir en uyumlu modelin belirlenmesi amacıyla kullanılmaktadır [40].

Doğrusal regresyon denkleminin matematiksel gösterimi aşağıdaki şekilde yapılmaktadır [41,42]:

$$Y_i = \sum_{k=0}^{p} \beta_k x_{ik} + \varepsilon_i \quad (3.4)$$

Burada,

Y: bağımlı değişken,

β: regresyon katsayıları,

X: bağımsız değişkenler,

ε: hata terimini göstermektedir.

Diğer bir ifadeyle; k bağımsız değişken ve N gözlem olduğunda doğrusal regresyon modelinin genel formu i. gözlem için,

$$y_i=\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+.....+\beta_k x_{ki}+\varepsilon_i \quad (3.5)$$

şeklinde yazılmaktadır. Bağımlı değişkenin düzeylerinin 0-1 arasında değerler aldığında ve bu değerlerin sonsuz değer olabilen bağımsız değişkenlerden dolayı yukarıdaki eşitliğin sağlanamadığı durumlarda olasılık değerinin dönüşümler yapılarak $-\infty,\infty$ arasında tanımlı duruma gelmesi sağlanmaktadır. Söz konusu durum için lojit (logit) adıyla bilinen dönüşüm sonucunda olasılık değerinin bağımsız değişken ve bağımlı değişken arasında eğrisel bir ilişkiyi sağlayan model oluşturulmaktadır. β_1'in işaretine göre S veya ters S şeklinde olan eğrileri sağlayan,

$$E(y_i) = P(y_i) = \ln\left(\frac{Pi}{1-Pi}\right)=\frac{e^{(\beta_0+\beta_1 x_{1i}1i+\beta_2 x_{2i}+.....+\beta_k x_{1ki}}}{1+e^{(\beta_0+\beta_1 x_{1i}1i+\beta_2 x_{2i}+.....+\beta_k x_{1ki}}} \quad (3.6)$$

şeklinde formülize olan bu fonksiyona "Lojistik Fonksiyon" adı verilmektedir. Lojit dönüşümü sayesinde oluşturulan lojistik regresyonla elde edilen modeldeki P_i olasılık değeri,

$$P(y) =\frac{1}{(1+e^{-y})} \quad (3.7)$$

olarak yazılmaktadır. Bu değer lojistik modele ait olan olasılık değeridir [40,42,43]. Doğrusal regresyon analizinde yer alan katsayıların anlamlı olup olmadığının belirlenmesi için kullanılan F testi, lojistik regresyonda yerini çeşitli ki-kare, olabilirlik oran testi gibi uyum iyiliği kriterlerine bırakmaktadır [44].

Lojistik regresyon modelinde, Wald testi olarak adlandırılan parametrelerin maksimum olabilirlik oranlarının karşılaştırılmasıyla Ho: $\beta_1 = 0$ hipotezi altında, değişken anlamlılığının test edilmesinde kullanılan bir diğer yöntemlerden biridir. Katsayıların anlamlılığı ve etkisinin ölçüldüğü bir diğer kriterdir. Katsayıların tahmin değerlerinin standart hatalarına oranı olup, Z test değeri ile karşılaştırılmaktadır. Wald kriterinin 2' den büyük olması değişkenlerin model için önemli olduğu sonucunu vermektedir [44,45].

Modele dahil edilen bağımsız değişkenlerin -2Log(L) istatistiği olarak hesaplanan ve bağımlı değişkendeki varyans değişimine sebep olan, doğrusal regresyon modelin bulunan hata kareler toplamı mantığıyla benzer biçimde işlemektedir. Olabilirlik oranı 1 ise, – 2 Log(L) istatistiği sıfıra eşit olmaktadır. Model, verileri tamamıyla açıkladığı durumda olabilirlik oranı 1 ve dolayısıyla – 2 Log(L)'nin sıfır olması anlamına gelir. Bu sebeple – 2 Log(L) istatistiğinin küçük olması, daha iyi bir model olduğunu göstermektedir. Ayrıca

AIC, BIC gibi tutarlı sonuçlar veren ve en iyi modelin belirlenmesinde kullanılan kriterler de kullanılmaktadır. [46,47]

Lojistik regresyon fonksiyonunda tahmin edilen regresyon katsayılarının yorumlanması doğrusal regresyon modeline göre farklılık göstermektedir. Lojistik regresyon modelinde x değişkenindeki bir birimlik artışın bağımlı değişkene etkisini yorumlamak oldukça güçtür. β_1 katsayısı yorumlanırken bağımlı değişkende görülen bir birimlik artış için $\pi_i/(1-\pi_i)$ odds tahmini ile $\exp(\beta_1)$ çarpılarak elde edilen lojistik regresyon fonksiyonu kullanılmaktadır. Lojistik regresyon modelinde etkiler odds değerleri ile yorumlanmaktadır. Bağımsız değişkenin bir değerinde kestirilen odds değerinin diğer değerinde kestirilen odds değerine oranı olarak yorumlanmaktadır. Bağımsız değişken yani x değeri 1 olan bireylerin x değeri 0 olan bireylere göre bağımlı değişkenin kaç kat daha fazla 1 olarak görüldüğü olasılığını vermektedir [48].

3.3. Veri Madenciliği Yöntemlerinin Karşılaştırılması için Kriterler

Yapılan çalışmalar sonucunda, farklı yöntemlerin birlikte kullanılmasının tahminlerin geliştirilmesinde oldukça etkili olduğu düşünülmektedir [49].

Veri madenciliği yöntemleri ile büyük veri yığınlarından gizli, bilinmeyen bilgiler çıkarılması, kullanıcıya kolaylık sağlayan, ileriye yönelik tahminler yapılmasına imkân tanıyan yöntemler bütünü olsa da, kullanılan veri kümesine ait elde edilen bilgi kesinleşmiş bir bilgi değildir. Model başarısını öncelikle etkilediği düşünülen aşamalar bulunmaktadır. Bu aşamalarda başarı kullanıcıya göre değişiklik gösterebilmektedir. Söz konusu aşamalar aşağıda kısaca özetlenmiştir [50]:

Veri önişleme aşaması; Veri madenciliği yöntemler uygulanmadan önce yapılan veri temizleme, dönüştürme, indirgeme işlemleri model başarısını oldukça etkileyen aşamalardandır. İlgili konu hakkında çalışacak bireyin bilgi ve tecrübesine göre değişiklik göstermektedir.

Parametre seçimi aşaması; Veri madenciliğinde yer alan algoritmaların farklı parametreleri bulunabilmektedir. Yapay sinir ağlarında bulunan nöron sayısı, karar ağaçlarında durdurma kriteri gibi farklılıklar algoritmaların başarı sonuçlarını etkilemektedir.

Test kümesinin seçimi; Model kurulmadan önce veri setinin hem eğitilmesi hem de test edilmesi amacıyla yapılan bölümleme (Partition) işlemi veri setini test verisi ve öğrenme verisi olarak ayırmaktadır. Bu bölümlemenin hangi oranda, hangi metotla yapılacağı gibi faktörler bölünmeyi değiştireceğinden model başarıları da bu aşamadan etkilenmektedir.

Model başarısının değerlendirilmesinde literatürde en çok kullanılan temel ölçütlerden (Verinin Bölümlenmesi)Hold-out yöntemi, doğruluk, kesinlik, duyarlılık ve F-ölçütüdür [50,51].

-Hold-out yöntemi; Veri setinin en az 2 kümeye bölümlenmesiyle uygulanmaktadır. Modelin eğitildiği veri seti olarak eğitim (training) verisi, sonuçlarının test edildiği test verisi olmak üzere bölünmesiyle doğru sınıflama oranları test verisi üzerinden yorumlanmaktadır. Bu yöntem veri setinin yeterli büyüklükte olduğu kümelerde kullanılmaktadır [51,52].

Modellerin başarıları değerlendirilirken, modelin doğruluk oranı, hata oranı, kesinlik, duyarlılık ve F-ölçütü gibi kriterler aşağıdaki tabloda belirtilen sınıfların tahmin düzeyine göre aşağıdaki gibi hesaplanmaktadır [50].

Çizelge 3.1.Sınıflama matrisi tablosu

		Tahmini Sınıf	
Doğru Sınıf		x=1	x=0
	x=1	a	b
	x=0	c	d

a: Doğru Pozitif

b:Yanlış Pozitif

c:Yanlış Negatif

d:Doğru Negatif

-Doğruluk Oranı; Model karşılaştırmada, en sık kullanılan kriterlerden biridir. Oluşturulmuş modelin yaptığı doğru örnek sayısının, örnek sayılarının tamamının toplamına bölünmesiyle elde edilen orandır [53].

$$\text{Doğruluk oranı} = \frac{a+d}{a+b+c+d} \tag{3.8}$$

$$\text{Hata Oranı} = \frac{b+c}{a+b+c+d} \tag{3.9}$$

Hata oranı da doğruluk oranından yola çıkılarak yanlış tahmin edilen örnek sayısının toplam örnek sayısına oranlanmasıyla elde edilmektedir [50].

- Kesinlik ölçütü; sınıfı 1 olarak tahminlenmiş örneklerin, gerçek sınıfı 1 olan tüm örnek sayısına oranlanmasıyla elde edilmektedir [53].

$$\text{Kesinlik oranı} = \frac{a}{a+b} \tag{3.10}$$

- Duyarlılık (Anma) Ölçütü; sınıfı ve tahmini 1 olan (Gerçek pozitif) örneklerin sınıfı 1 olarak tahmin edilmiş (Pozitif) örneklerin tamamına oranlanmasıyla elde edilmektedir [53].

$$\text{Duyarlılık oranı} = \frac{a}{a+c} \tag{3.11}$$

-F Ölçütü; Duyarlılık ve kesinlik oranlarının yalnızca pozitif örnek ve tahminler üzerinden hesaplama yapmasına karşın, F-ölçütü her iki oranın harmonik ortalaması ile hesaplanmaktadır [53].

$$\text{F ölçütü} = \frac{(2 * \text{Kesinlik} * \text{Duyarlılık})}{(\text{Kesinlik} + \text{Duyarlılık})} \tag{3.12}$$

-Lift Ölçütü: Sınıflandırma modellerinin performansını karşılaştırmanın en yaygın yollarından biri de lift ölçütüdür. Tüm beklentileri, modelin tahmin ettiği gibi cevap verme olasılıklarına göre sıralayarak oluşturulur ve her parametre bir değer alır. Modelin test setindeki ölçüm yüksekliği doğru modelin seçilmesine yardımcı olmaktadır. Lift değeri pozitif olarak sınıflandırılan verilerin gerçekte pozitiflere oranının, tüm veri setindeki pozitif oranına oranlanmasıyla hesaplanmaktadır [21,31].

Diğer bir ifadeyle gerçekleşen durumların model tarafından tahmin edildiği kadarının model ele alınmadan rasgele olarak alınan örneklerde gerçekleşen durumlara oranı olarak açıklanmaktadır [7].

Başarılı bir model için lift değerinin 1 değerinden daha fazla olması beklenmektedir. Lift Grafiği gösteriminde ise, lift değerinin 1'den daha da yukarda başlaması model başarı göstergesidir, grafik paralel seyir göstererek birden 1 değerine ulaşması da modelin başarılı olduğuna işaret etmektedir [6,54].

4. UYGULAMA

Bu bölümde, 1988 yılından itibaren Uluslararası Çalışma Örgütü tarafından belirlenen standartlarda toplanan, 2014 yılından itibaren ise Avrupa Birliği (AB) yönetmeliklerine tam uyum sağlanması kapsamında çeşitli değişkenlerde yapılan düzenleme ve değişikliklerle hali hazırda uygulanmakta olan, Hanehalkı İşgücü Araştırmaları (HİA)' na ilişkin veriler çalışma kapsamında ele alınmıştır. HİA' nda veri seti formatında 2014 yılından itibaren çeşitli değişiklikler olduğundan dolayı 2014, 2015 ve 2016 yılına ait Hanehalkı İşgücü Araştırmaları (HİA) verileri kullanılmıştır. TÜİK tarafından uygulanan ve yürütülen HİA' na ait detaylı bilgiler ışığında işgücü durumlarının belirlenmesi için hanede yaşayan 15 yaş ve üzeri bireylere sorulan formlara ait veri setleri TÜİK' ten temin edilerek çalışma kapsamına alınmıştır.

4.1. Kullanılan Veri Setine İlişkin Tanım ve Kavramlar

Çalışmada kapsamında kullanılan HİA' na ait veri setiyle ilgili olarak çeşitli kavramların tanımlamaları aşağıda verilmiştir [55]:

Hanehalkı: Aralarında akrabalık olup olmamasına bakılmaksızın aynı evde yaşayan, temel ihtiyaçlarını birlikte karşılayan ve hanehalkı yönetimine katılan kişilerin topluluktur.

Kurumsal Olmayan Nüfus: Üniversite yurtları, huzurevi, hapishane, kışla gibi yerlerde ikamet edenler dışındaki topluluktur.

Kurumsal Olmayan Çalışma Çağındaki Nüfus: Kurumsal olmayan nüfus içerisinde 15 yaş ve üzerinde bulunanların oluşturduğu topluluktur.

İşgücü: Ekonomik mal ve hizmet üretimi için, emek veren çalışma çağında bulunan nüfustur. İşgücü, istihdamda bulunanlar ile işsizlerin toplamından oluşmaktadır.

İstihdam: İşbaşında olanlar ve işbaşında olmayanlar grubuna dâhil olan kurumsal olmayan çalışma çağındaki nüfus, istihdam edilen nüfustur.

İşsiz: Referans dönemi içinde istihdam halinde olmayan kişilerden iş aramak için son 4 hafta içinde iş arama yollarına başvuran ve 15 gün içinde işbaşı yapabilecek durumda olan kurumsal olmayan çalışma çağındaki tüm kişilerdir.

İşgücüne Dâhil Olmayanlar: İşsiz veya istihdamda olmayan 15 ve daha yukarı yaştaki nüfustan oluşmaktadır.

- İş aramayıp çalışmaya hazır olanlar: Çeşitli nedenlerle iş aramayan, ancak 2 hafta içinde işbaşı yapmaya hazır olduğunu belirten kişilerdir. İki alt başlıkta ele alınmaktadır:

a) İş bulma ümidi olmayanlar: Daha önce iş aradığı halde bulamayan veya kendi vasıflarına uygun bir iş bulamayacağını düşündüğünden iş aramayan ancak işbaşı yapmaya hazır olan kişilerdir.

b) Diğer: Mevsimlik çalışma, ev kadını olma, öğrencilik, emeklilik ve çalışamaz halde olma gibi nedenlerle iş aramayıp ancak işbaşı yapmaya hazır olduğunu belirten kişilerdir.

4.2. Kullanılan Veri Seti ve Yöntemler

Türkiye'de aktif nüfus yapısı hakkındaki bilgiler, beş yılda bir yapılan Genel Nüfus Sayımlarından ve belirli dönemlerde uygulanan Hanehalkı İşgücü Araştırmalarından (HİA) çalışmalarıyla başlamıştır. Özellikle işgücü piyasasının durumunun incelenmesinde elde edilen veriler, değişkenler ve sınıflandırmalardaki farklılıklar sebebiyle kıyaslama yapmak güçleşmiştir. Bu sebeplerden ötürü, Birleşmiş Milletler Kalkınma Programı (UNDP) koordinatörlüğünde bulunan İşgücü Piyasası Enformasyon Sistemi (İPES) Projesi ve Dünya Bankası İstihdam ve Eğitim Projesi kapsamında işgücü piyasası göstergelerinin daha doğru ölçülmesi ve yorumlanması amacıyla tekrar yapılanması hedeflenmiştir. Söz konusu araştırma, ülkedeki işgücünün piyasasının belirlenmesi, takip edilmesi, istihdam edilen kişilerin mesleki durumu, çalışma süreleri, iş arama süreleri ve benzer özellikleri hakkında çeşitli bilgilerin derlenmesi amacıyla yapılmaktadır. Belirli örnekleme yöntemleriyle seçilen hanelerde yaşayan bireylere uygulanan araştırma, Türkiye geneline ait işgücü durumu hakkında oldukça detaylı bilgiler içermektedir [56].

Bu bağlamda, Birleşmiş Milletler Kalkınma Programı (UNDP) koordinatörlüğünde yürütülen İşgücü Piyasası Enformasyon Sistemi (İPES) Projesi kapsamında ve daha sonra Dünya Bankası İstihdam ve Eğitim Projesi çerçevesinde işgücü piyasası göstergelerinin daha doğru ve zamanında ölçülmesi amacına yönelik olarak Hanehalkı İşgücü Araştırmasının yeniden yapılandırılması hedeflenmiştir. 1988 yılına kadar, işgücü anketlerinden elde edilen veriler, coğrafi bölge, değişken ve sınıflamada ortaya çıkan birçok farklılıktan dolayı kıyaslama yapmaya uygun olmaması sebebiyle belirli normlara uygun hale getirilmesi planlanmıştır. TÜİK tarafından yürütülen HİA 1988 yılından beri, Uluslararası Çalışma Örgütü tarafından belirlenen standartlarda gerçekleştirilmektedir. HİA' da 2014 yılında Avrupa Birliği (AB) yönetmeliklerine tam uyum sağlanması kapsamında örneklem tasarımı, işsizlik kriterindeki iş arama süresi, idari bölünmelerdeki yeni yapılar gibi çeşitli değişiklikler yapılmıştır. Çalışmanın ilk bölümünde hane içinde bulunan fertlerin cinsiyet, yaş, eğitim düzeyi gibi demografik özellikler; diğer bölümde ise 15 yaş ve üstü bireylerin işgücü durumunun tespiti yapılmaktadır. HİA' nın

yürütülmesinde, üniversite yurtları, huzurevi, hapishane, kışla, orduevleri gibi kurumsal yerlerde ikamet edenler dışında Türkiye'de bulunan tüm hanelerde yaşayan bireyler çalışma kapsamına alınmıştır. Hanehalkı işgücü araştırmasında iki aşamalı, tabakalı küme örneklemesi yöntemi kullanılmıştır. Çalışmanın tasarımında her dönemde uygulama yapılacak örneklem genişliği haftalara eşit olarak dağıtılmıştır [56].

4.3. Veri Setinin Düzenlenmesi

Veri seti 15 yaş ve üstü olan bireylere ilişkin işgücü bilgilerini içermektedir. Bağımlı değişken olan işgücü durumu kategorik bir değişkendir. İşgücü durum değişkeni TÜİK' ten elde edilen veri setinden yola çıkılarak bağımlı değişken 2 kategoriden oluşmaktadır. Bağımlı değişkenin kategorileri; İstihdam / İstihdam dışı (işsiz ve işgücüne dâhil olmayanlar) olmak üzere algoritmalar uygulanmıştır.

TÜİK tarafından yürütülen HİA 2014, 2015 ve 2016 yıllarına ait toplam 1,163,566 bireye ilişkin işgücü verileri çalışma kapsamına alınmıştır. HİA çalışması veri setlerine ilişkin değişkenler ve değişken açıklamaları aşağıdaki şekilde düzenlenmiştir:

Yaş Grubu Değişkeni; HİA 2013 yılı çalışmasında, ele alınan yaş grupları değişkeninde bulunan gruplar 2014 yılından itibaren değişikliğe uğramıştır. 2014 yılı HİA' da ele alınan gruplarda sınıflama ve yorumlama açısından kategoriler arası birleştirme işlemi yapılmıştır. Veri setinde bulunan yaş grupları eşit aralıklı olmadığından ve oldukça fazla olması sebebiyle yorumda sıkıntı yaratabileceği düşünüldüğünden bitirilen yaş değişkeninden yola çıkılarak yeni gruplar aşağıdaki şekilde oluşturulmuştur:

15-24 yaş arası, 25-34 yaş arası, 35-44 yaş arası, 45-54 yaş arası, 55-64 yaş arası, 65 yaş ve üzeri olmak üzere yaş grubu değişkeni 6 kategoriden oluşmaktadır.

Cinsiyet değişkeni; 1-Erkek, 2-Kadın olmak üzere modele alınmış 2 kategorili bir değişkendir.

Medeni Hal değişkeni; HİA çalışması kapsamında medeni hal değişkeni;

1- Hiç evlenmedi, 2- Evli, 3- Boşandı, 4- Eşi öldü olarak düzenlenmiştir, kategorilerin dağılımının orantılı olması açısından, "Boşandı" ve "Eşi öldü" düzeyleri birleştirilerek değişken 3 kategorili bir hale getirilmiştir.

1-Hiç evlenmedi, 2-Evli, 3-Dul olmak üzere medeni hal değişkeni 3 kategoriden oluşmaktadır.

İBBS-1 (İstatistiki Bölge Birimleri Sınıflaması) Değişkeni; Türkiye'nin Avrupa Birliği'ne aday olarak kabulü ile hazırlanan Ulusal Program kapsamında gerekli olan bölge sınıflamaları çalışmaları yapılmıştır. Devlet Planlama Teşkilatı (DPT), Devlet İstatistik Enstitüsü (DİE) ve İçişleri Bakanlığı tarafından yapılan çalışmalar sonucunda AB ülkelerindeki yapıya benzer olarak 3 düzeyden oluşan sınıflandırmalar yapılmıştır. Nüfus dışında kültürel durum, gelişmişlik düzeyi gibi faktörler de ele alınarak hazırlanmıştır. Şekil 4.1' de yerleşme merkezlerinin kademelenmesini coğrafi koşulları, istatistik toplama ve plan yapma amacına uygunluğu da dikkate alarak Türkiye'de de üç düzeyden oluşan İstatistiki Bölge Birimleri Sınıflandırması (İBBS) yapılmıştır [57].

Şekil 4.1. İBBS1 bölge dağılım [57]

Çizelge 4.1' de İBBS sınıflamalarından İBBS-1 düzeyi için oluşturulan 12 düzey TÜİK' te belirtildiği şekildedir:

Çizelge 4.1. İBBS-1 düzeyleri

İBBS-1 DÜZEYİ	DÜZEY ADI
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	Kuzeydoğu Anadolu
TRB	Ortadoğu Anadolu
TRC	Güneydoğu Anadolu

İBBS-2 (İstatistiki Bölge Birimleri Sınıflaması) Değişkeni; Türkiye'de de üç düzeyden oluşan İstatistiki Bölge Birimleri Sınıflandırması (İBBS) düzeylerinden İBBS-2 düzeyi Şekil 4.2' deki gibi bölümlendirilmiştir:

Şekil 4.2. İBBS2 bölge dağılım [57]

Çalışma kapsamında anket uygulanan bireylerin yaşadığı illere göre İBBS sınıflamalarından İBBS-2 düzeyi için oluşturulan 26 düzey TÜİK' te belirtildiği gibi aşağıdaki şekildedir:

Çizelge 4.2. İBBS-2 düzeyleri

İBBS-2 DÜZEYİ	DÜZEY ADI
TR10	İstanbul
TR21	Edirne, Tekirdağ, Kırklareli
TR22	Balıkesir, Çanakkale
TR31	İzmir
TR32	Denizli, Aydın, Muğla
TR33	Manisa, Afyonkarahisar, Kütah
TR41	Bursa, Eskişehir, Bilecik
TR42	Kocaeli, Sakarya, Düzce, Bolu
TR51	Ankara
TR52	Konya, Karaman
TR61	Antalya, Isparta, Burdur
TR62	Adana, Mersin

TR63	Hatay, Kahramanmaraş, Osmaniye
TR71	Nevşehir, Aksaray, Niğde, Kırıkkale,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	Kars, Ağrı, Iğdır, Ardahan
TRB1	Malatya, Elazığ, Bingöl, Tunceli
TRB2	Van, Muş, Bitlis, Hakkari
TRC1	Gaziantep, Adıyaman, Kilis
TRC2	Diyarbakır, Şanlıurfa
TRC3	Siirt, Mardin, Batman, Şırnak

Okuryazarlık Değişkeni; 1-evet 2-hayır olmak üzere 2 kategoriden oluşmaktadır. En son bitirilen okul sorusuna "hiç" cevabı verenlere sorulmuştur, veri setinde boş gelen satırlar eğitim alan kişiler olduğundan dolayı okuryazarlık sütunu evet olarak doldurulmuştur.

Bitirilen Okul Değişkeni; Veri setinde bulunan Bitirilen okul değişkenindeki ilkokul ve ortaokul düzeyleri birleştirilerek ilköğretim olarak ele alınmıştır. Elde edilen bitirilen okul değişkeni düzeyleri; 1-Hiçbir Okul Bitirmeyen, 2-İlköğretim, 3-Genel Lise, 4-Mesleki Lise, 5- 2 veya 3 yıllık yüksekokul, 4 yıllık yüksekokul veya fakülte ve daha üzeri (yüksek lisans, doktora) olmak üzere 5 kategoriden oluşmaktadır.

Bitirilen Bölüm Değişkeni; En son bitirilen okuldan mezun olunan bölüm kodu olarak TÜİK tarafından genel eğitim, mesleki eğitim ile ilgili tüm eğitim programlarını sınıflandırmak amacıyla kullanılan Eğitim ve Öğretim Alanları Sınıflaması (FOET99) sınıflamasına göre bölümler ele alınmıştır.

Bu değişken mesleki lise ve üzeri mezuniyeti olan kişilere sorulmuştur bu sebeple genel lise, İlköğretim ve hiç bir okul bitirmeyen kişilerde boş olduğundan bu durumda olanlar "0" ile doldurulmuştur. FOET-99 sınıflamasına göre bölümler aşağıdaki gibi sınıflandırılmıştır:

Çizelge 4.3. FOET sınıflamasına göre bölümler

FOET-99 KODU	BÖLÜM ADI
0	Bölüm Yok
1	Öğretmen eğitimi ve eğitim bilimleri
2	Sanat
3	Beşeri bilimler
4	Sosyal bilimler ve davranış bilimleri
5	Gazetecilik ve enformasyon
6	İş ve yönetim
7	Hukuk
8	Yaşam bilimleri
9	Fizik bilimleri
10	Matematik ve istatistik
11	Bilgisayar
12	Mühendislik ve işleri
13	İmalat ve işleme
14	Mimarlık ve inşaat
15	Tarım, ormancılık ve balıkçılık
16	Veterinerlik
17	Sağlık
18	Sosyal hizmetler
19	Kişisel hizmetler
20	Ulaştırma hizmetleri ve çevre koruma
21	Güvenlik hizmetleri

Eğitime Devam Değişkeni; Çalışmadaki bireylerin açık öğretim dâhil örgün eğitime devam edip etmemesine göre 1-Evet 2-Hayır olarak sınıflandırılmıştır.

4.4. Tanımlayıcı İstatistikler

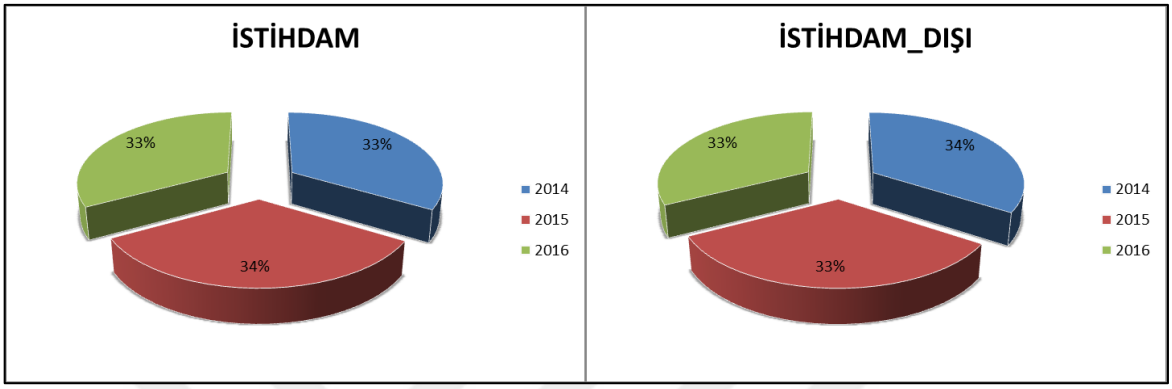
Veri setinde yer alan bağımlı ve bağımsız değişkenlere ait tanımlayıcı istatistikler elde edilmiştir. Bu bölümde her değişken için tanımlayıcı istatistiklere ait tablolar elde edilmiş, bağımlı değişkene göre bağımsız değişkenlerin kategorileri arasındaki dağılım görselleştirilmiştir.

Çizelge 4.4. İşgücü durum değişkeninin kategorilere göre dağılımı

İŞGÜCÜ DURUM

	Sıklık	Yüzde(%)	Kümülatif Yüzde(%)
İSTİHDAM	520141	44,7	44,7
İSTİHDAM DIŞI	643425	55,3	100,0

Çizelge 4.4' te işgücü durumu değişkeni değerlendirildiğinde; istihdam durumu kategorisi tüm veri setinin yaklaşık %45' ini, istihdam dışı kategorisi ise kalan %55' lik bölümü kapsadığı görülmüştür.



Şekil 4.3. İşgücü durumu değişkeninin yıllara göre dağılımı

Şekil 4.3' te 2014, 2015, 2016 yıllarında Hanehalkı İşgücü Araştırması'nın uygulandığı birey sayıları birbirine oldukça yakındır. İşgücü durumu kategorilerinden; hem istihdam hem istihdam dışı için her yıl eşit oranlarda bir dağılım olduğu görülmüştür. İşgücü durumu için her 2 kategoride (İstihdam, istihdam dışı) yıllık dağılım yaklaşık %33 civarında olmuştur.

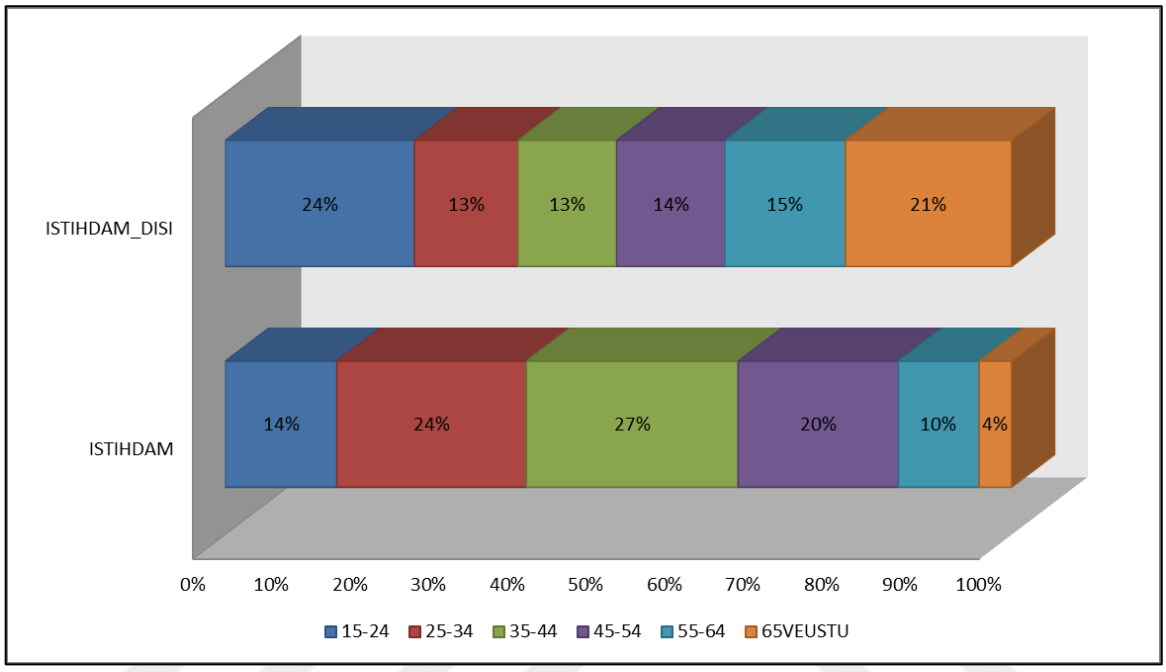
-İşgücü Durumuna Göre Yaş Gruplarının Dağılımı:

Çalışma veri seti, istihdam verisi olduğundan 15 yaş ve üstü bireylere ilişkin bilgileri kapsamaktadır.

Çizelge 4.5. Yaş grubu değişkeninin kategorilere göre dağılımı

YAŞ GRUBU			
	Sıklık	**Yüzde(%)**	**Kümülatif Yüzde(%)**
15-24	228692	19,7	19,7
25-34	210097	18,1	37,7
35-44	220813	19,0	56,7
45-54	194611	16,7	73,4
55-64	151999	13,1	86,5
65 VE ÜSTÜ	157354	13,5	100,0
TOPLAM	1163566	100,0	

Çizelge 4.5’ te HİA çalışmasına katılan 15 yaş ve üzeri tüm bireylerin dağılımı incelendiğinde; 55-64 ile 65 ve üstü yaş grubunda olan bireylerin çalışma kapsamında yaklaşık %13’lük oranlarda olduğu, diğer yaş gruplarının ise yaklaşık %16-19 aralığında dağılım gösterdiği görülmektedir.



Şekil 4.4. İşgücü durumu değişkeninin yaş gruplarına göre dağılımı

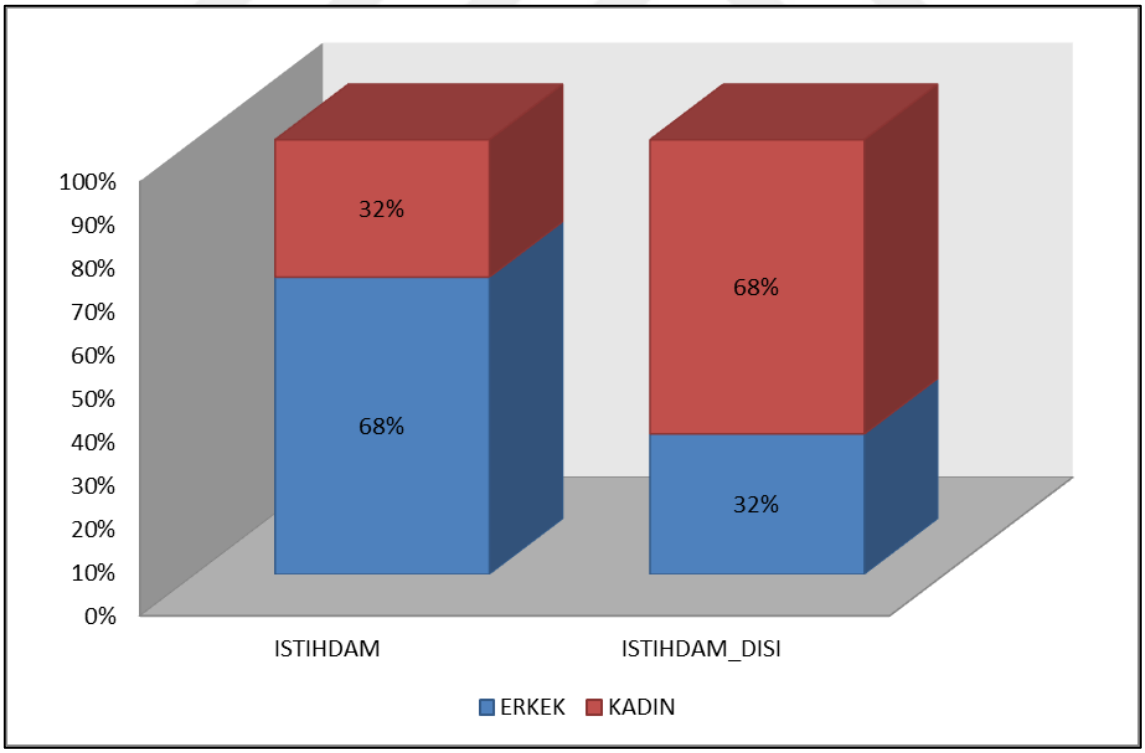
Şekil 4.4’ te istihdam durumunda olan bireyler incelendiğinde; bu bireylerin %14’ ü 15-24 yaş aralığındaki grupta yer almaktadır. 35-44 yaş aralığındaki bireyler en yüksek orana sahip olup istihdamdaki bireylerin %27’ lik kısmını, %4’ lük kısmını ise 65 yaş ve üzerindeki bireylerin oluşturduğu görülmektedir. İstihdamda olan bireylerin yaklaşık %71’lik kısmını 25 ile 54 yaş aralığındaki bireyler oluşturmaktadır.

İstihdam dışında olan bireyler incelendiğinde; en düşük istihdam dışı oranının yaklaşık %13 olarak 25-34 ile 35-44 yaş aralığındaki bireylerden oluştuğu görülmektedir. İstihdam dışındaki bireylerin yaklaşık %24’lük kısmı 15-24 yaş aralığındaki bireylerden oluşmaktadır. 65 yaş ve üzeri olan bireyler ise tüm istihdam dışındaki bireylerin yaklaşık %21’ ini oluşturmaktadır. İstihdam dışındaki bireylerin yaklaşık %45’ i 15-24 ve 65 yaş üstü bireyleri kapsamaktadır.

Çizelge 4.6. Cinsiyet değişkeninin kategorilere göre dağılımı

CİNSİYET			
	Sıklık	**Yüzde(%)**	**Kümülatif Yüzde(%)**
ERKEK	563553	48,4	48,4
KADIN	600013	51,6	100,0
TOPLAM	1163566	100,0	

Çizelge 4.6' da HİA çalışmasına katılan bireylerin cinsiyete göre işgücü durumu ele alınmıştır. Çalışma kapsamındaki bireylerin %51.6' sı kadın, %48.4' sının erkek olduğu görülmektedir. Cinsiyet değişkeni için dağılım birbirine oldukça yakındır.



Şekil 4.5. İşgücü durumu değişkeninin cinsiyete göre dağılımı

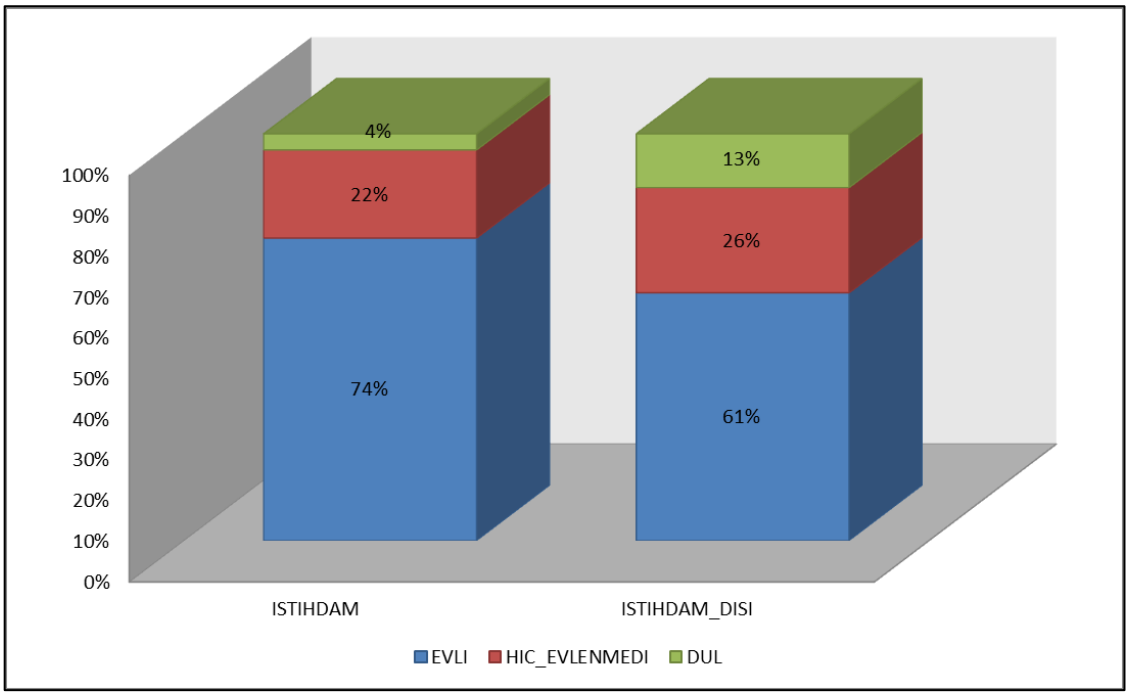
Şekil 4.5’ te istihdamda olan bireylerin %32’sinin kadın olduğu; istihdam dışında olan bireylerin ise %68’ nin kadın olduğu görülmektedir. İstihdam dışında kalan bireylerin çoğunluğunun kadın olması dikkat çekmektedir.

İstihdamda olan bireylerin %68’ nin erkek olduğu; istihdam dışında olan bireylerin %32’sinin erkek olduğu görülmektedir. Erkeklerin istihdamda bulunma oranı, kadınların istihdamda bulunma oranının 2 katından yüksek olduğu görülmektedir.

Çizelge 4.7. Medeni hal değişkeninin kategorilere göre dağılımı

MEDENİ HAL			
	Sıklık	**Yüzde(%)**	**Kümülatif Yüzde(%)**
DUL	105534	9,1	9,1
EVLİ	778828	66,9	76,0
HİÇ EVLENMEDİ	279204	24,0	100,0
TOPLAM	1163566	100,0	

Çizelge 4.7’ de HİA çalışmasına katılan bireylerin medeni durumlarına ilişkin değişken değerleri ele alınmıştır. Veri setinin yaklaşık %24’ünü bekârların, yaklaşık %66’ sını evli bireylerin kalan %9’luk kısmı ise dul bireylerin oluşturduğu görülmektedir.



Şekil 4.6. İşgücü durumu değişkeninin medeni hal durumuna göre dağılımı

Şekil 4.6’ da istihdamda olan bireylerin yaklaşık %74’ ünün evli olduğu görülmüştür. İstihdamdaki bireylerin yaklaşık %22’lik kısmını bekârlar, yaklaşık %4’lük kısmını dul bireyler oluşturmaktadır. İstihdam dışında olan bireyler incelendiğinde; yaklaşık %61’ inin evli olduğu görülmektedir. Yaklaşık %26’ sını bekârların oluşturduğu, yaklaşık %13’ lük oranı dul bireylerden oluştuğu görülmüştür. İstihdamda çoğunlukla evli bireylerin yer aldığını söylemek mümkündür.

-İşgücü Durumuna Göre İbbs-1 Dağılımı:

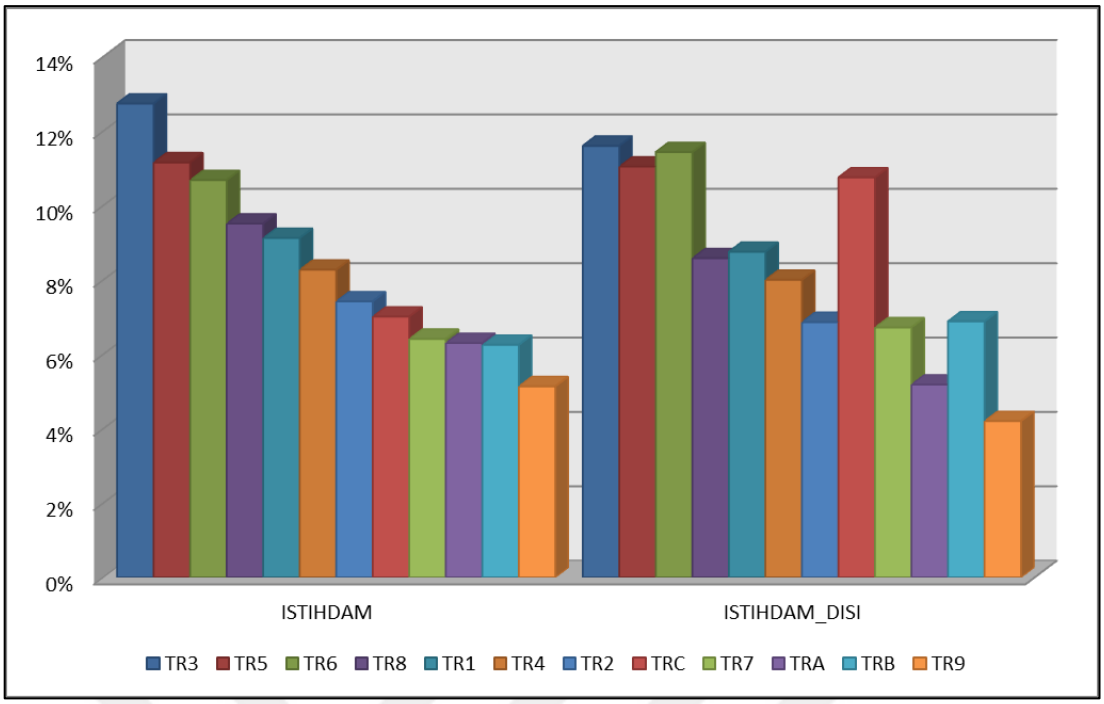
Şekil 4.8’ de yer aldığı gibi İBBS-1 düzeyinde dağılım 12 bölgeden oluşmaktadır.

Çizelge 4.8. İBBS-1 bölge düzeyi değişkeninin kategorilere göre dağılımı

İBBS_1			
	Sıklık	**Yüzde(%)**	**Kümülatif Yüzde(%)**
TR1	103748	8,9	8,9
TR2	82740	7,1	16,0
TR3	140897	12,1	28,1
TR4	94481	8,1	36,3
TR5	129074	11,1	47,3
TR6	129105	11,1	58,4
TR7	76583	6,6	65,0
TR8	104672	9,0	74,0
TR9	53744	4,6	78,6
TRA	66112	5,7	84,3
TRB	76783	6,6	90,9
TRC	105627	9,1	100,0
TOPLAM	1163566	100,0	

*(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 Kuzeydoğu Anadolu, TRB Ortadoğu Anadolu, TRC Güneydoğu Anadolu)

Çizelge 4.8’ de İBBS-1 düzeylerine göre istihdam verilerinde en fazla sıklığın görüldüğü bölge %12.1 oranı ile TR3(Ege) bölgesi olmuştur, sırasıyla TR5(Batı Anadolu) ve TR6(Akdeniz) yaklaşık %11.1 oranında takip etmektedir.



Şekil 4.7. İşgücü durumu değişkeninin İbbs-1 düzeylerine göre dağılımı

Şekil 4.7' de İBBS-1 düzeylerine göre istihdam durumu değişkeni incelendiğinde; istihdamdaki nüfusun yaklaşık %13'lük kısmı TR3(Ege) bölgesinde, yaklaşık %11'lik oranla TR5(Batı Anadolu) bölgesinde olduğu görülmüştür. İstihdam dışında olan grup incelendiğinde; bu grupta en yüksek TR3(Ege) bölgesinde yaklaşık %12 oranıyla, %4'lük oranla istihdam dışı TR9(Doğu Karadeniz) bölgesinde olduğu görülmüştür.

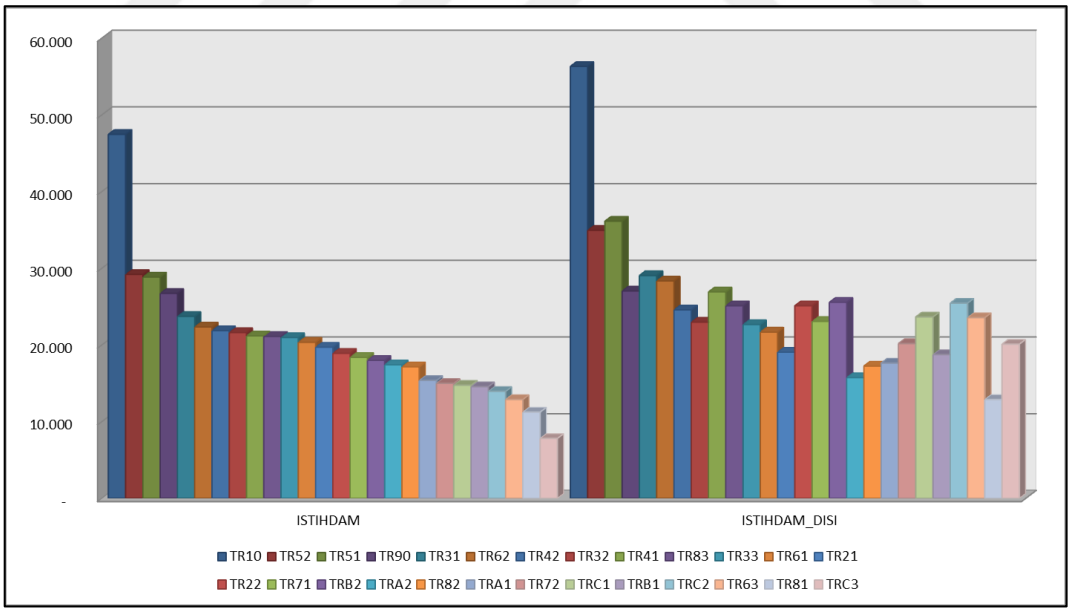
-İşgücü Durumuna Göre İbbs-2 Dağılımı:

Çizelge 4.9' da yer aldığı gibi İBBS-2 düzeyinde dağılım 26 bölgeden oluşmaktadır.

Çizelge 4.9. İBBS-2 bölge düzeyi değişkeninin kategorilere göre dağılımı

İBBS_2			
	Sıklık	**Yüzde(%)**	**Kümülatif Yüzde(%)**
TR10	103748	8,9	8,9
TR21	38741	3,3	12,2
TR22	43999	3,8	16,0
TR31	52746	4,5	20,6
TR32	44553	3,8	24,4
TR33	43598	3,7	28,1
TR41	48063	4,1	32,3
TR42	46418	4,0	36,3

TR51	64966	5,6	41,8
TR52	64108	5,5	47,3
TR61	42006	3,6	51,0
TR62	50666	4,4	55,3
TR63	36433	3,1	58,4
TR71	41439	3,6	62,0
TR72	35144	3,0	65,0
TR81	24107	2,1	67,1
TR82	34365	3,0	70,1
TR83	46200	4,0	74,0
TR90	53744	4,6	78,6
TRA1	32979	2,8	81,5
TRA2	33133	2,8	84,3
TRB1	33267	2,9	87,2
TRB2	43516	3,7	90,9
TRC1	38373	3,3	94,2
TRC2	39368	3,4	97,6
TRC3	27886	2,4	100,0
TOPLAM	1163566	100,0	



Şekil 4.8. İşgücü durumu değişkeninin İbbs-2 düzeylerine göre dağılımı

Şekil 4.8' de İBBS-2 düzeylerine göre istihdam verilerinde en yüksek sıklığın görüldüğü bölge TR10(İstanbul) olmuştur, en düşük sıklık ise %2.1 oranında TR81(Zonguldak, Karabük, Bartın) bölgesinde görülmüştür.

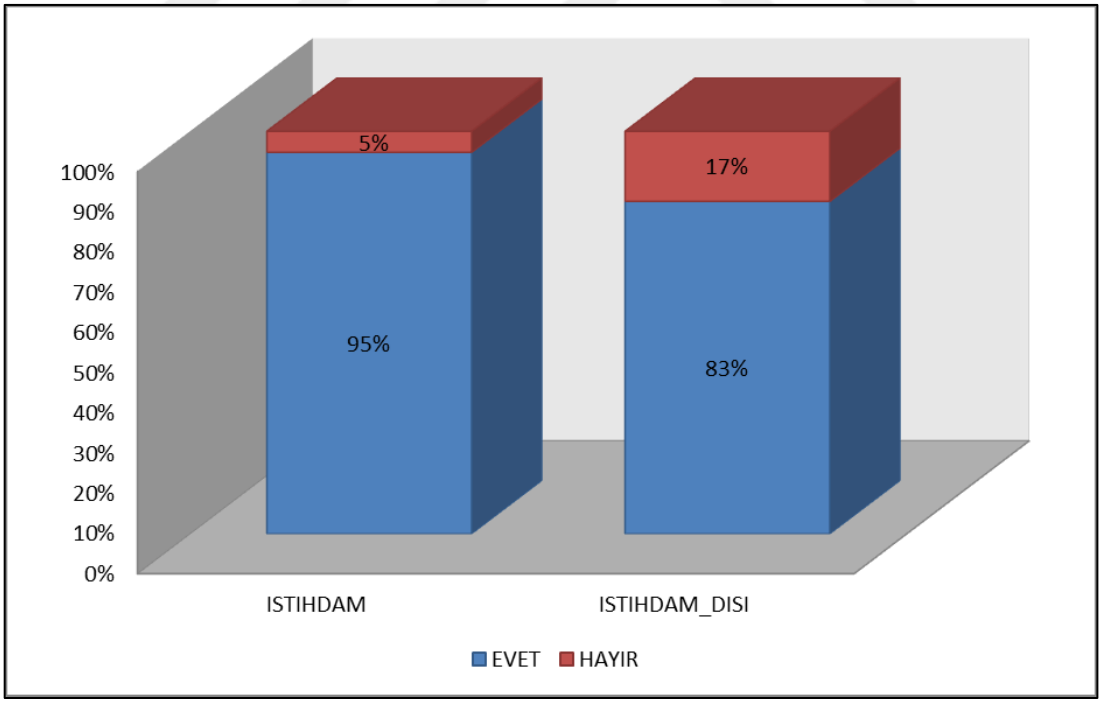
İBBS-2 düzeylerine göre işgücü durumu değişkeni incelendiğinde; istihdamdaki nüfusun TR10 (İstanbul) bölgesinde en yüksek, TRC3 (Siirt, Mardin, Batman, Şırnak) bölgesinde en düşük olduğu görülmüştür. İstihdam dışında ise, en yüksek yoğunluk TR10 (İstanbul) bölgesinde görülmüş, TR81 (Zonguldak, Karabük, Bartın) bölgesinde en düşük oranda istihdam dışı olduğu görülmüştür.

-İşgücü Durumuna Göre Okuryazarlık Durumu Dağılım:

Çizelge 4.10'da HİA çalışmasına katılan bireylerin okuryazarlıklarının olup olmamasına göre oluşturulan okuryazarlık değişkeni ele alınmıştır.

Çizelge 4.10. Okuryazarlık durumu değişkeninin kategorilere göre dağılımı

OKURYAZARLIK			
	Sıklık	**Yüzde(%)**	**Kümülatif Yüzde(%)**
EVET	1024471	88,0	88,0
HAYIR	139095	12,0	100,0
TOPLAM	1163566	100,0	



Şekil 4.9. İşgücü durumu değişkeninin okuryazarlık durumuna göre dağılımı

Şekil 4.9’ da istihdamda olan bireylerin oldukça yüksek oranda okuryazarlık durumu evet olan yani okuryazarlığı olan bireylerden oluşmaktadır. Yalnızca %5’lik kısmının okuryazarlığının olmadığı görülmektedir.

İstihdam dışında olan bireylerde ise okuryazarlığı olmayanların oranı %17 iken; okuryazarlığı olanlar yaklaşık %83 oranlarında olduğu görülmektedir.

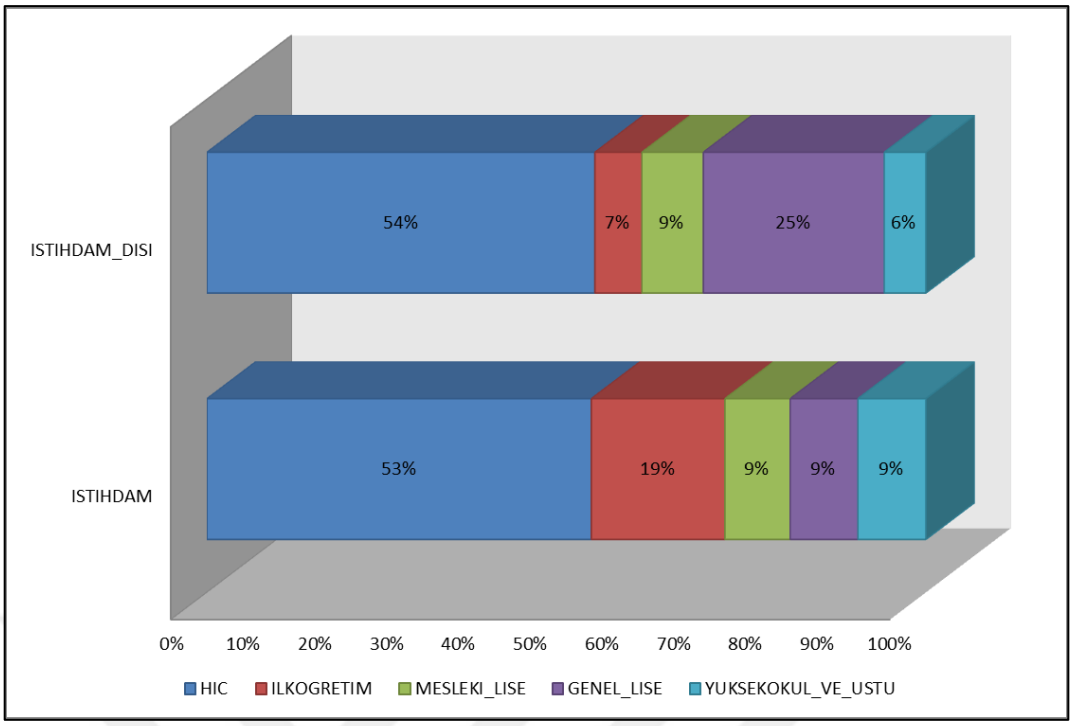
-İşgücü Durumuna Göre Bitirilen Okul Dağılımı:

Çizelge 4.11’ de işgücü durumu ile eğitim düzeyi “Bitirilen okul” arasındaki ilişki ele alınmıştır.

Çizelge 4.11. Bitirilen okul değişkeninin kategorilere göre dağılımı

BİTİRİLEN OKUL			
	Sıklık	**Yüzde(%)**	**Kümülatif Yüzde(%)**
GENEL LİSE	102226	8,8	8,8
HİÇ	210872	18,1	26,9
İLKÖĞRETİM	625340	53,7	80,7
MESLEKİ LİSE	86549	7,4	88,1
Y.OKUL VE ÜSTÜ	138579	11,9	100,0
TOPLAM	1163566	100,0	

Çizelge 4.11’ de görüldüğü üzere; HİA’ ya katılan bireylerin %53.7’ si ilköğretim mezunudur. Bunu sırasıyla %18.1 oranla hiçbir okul bitirmeyenler, %11.9’ luk oranla yüksekokul ve üzeri mezuniyeti olan bireyler, %8.8’ lik oranla genel liseden mezun olan bireyler ve %7.4’lük kısmını mesleki liseden mezun olan bireyler oluşturmaktadır.



Şekil 4.10. İşgücü durumu değişkeninin bitirilen eğitim düzeyine göre dağılımı

Şekil 4.10' da istihdamda olan bireylerde; mesleki lise, genel lise, yüksekokul ve üzeri mezuniyeti bulunanların sayısının birbirine yakın olduğu, yaklaşık %9'luk oranlarda bir dağılım gösterdiği görülmektedir. Bireylerin yarısından fazlasının hiçbir okul bitirmeyen grupta olduğu dikkat çekmektedir. Yaklaşık %19' luk kısmının ise; ilköğretim düzeyinde mezuniyet durumunda oldukları göze çarpmaktadır.

İstihdam dışında olan bireylerde ise; istihdamdaki bireylerle benzer şekilde %54' ünü hiçbir okul bitirmeyenler oluşturmaktadır. Bu grupta en düşük oran yüksekokul ve üzeri bireyler olarak görülmektedir. İstihdam dışındaki bireylerin yaklaşık %25' inin genel lise mezunu olduğu görülmektedir.

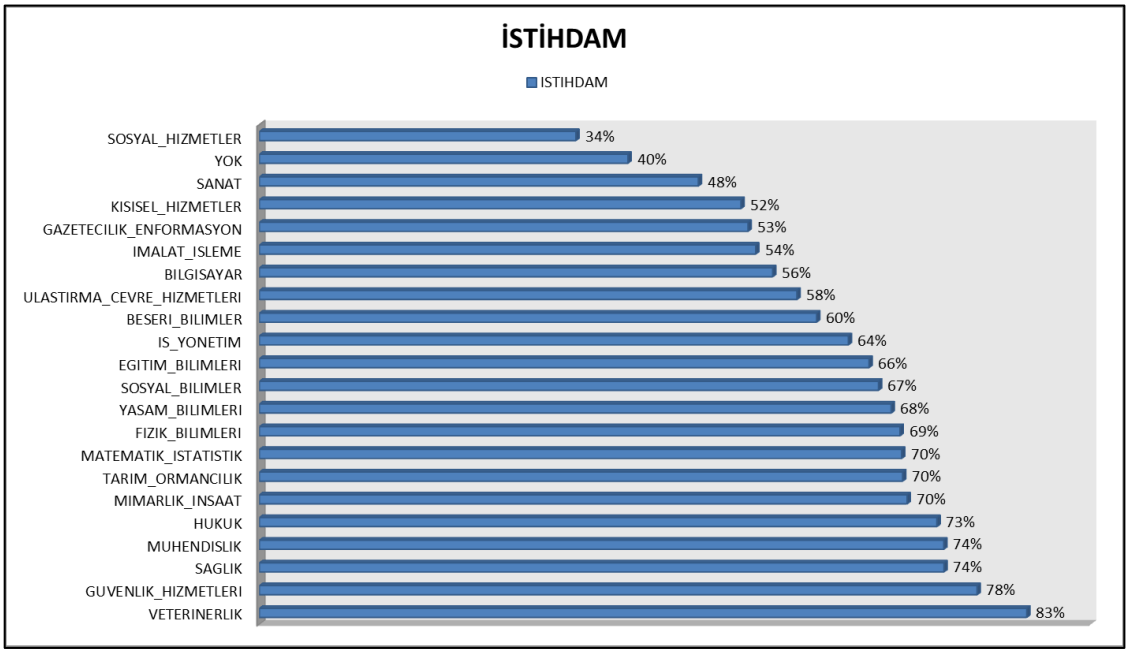
-İstihdam Durumuna Göre Bitirilen Bölüm Dağılımı:

Çizelge 4.12. Bitirilen bölüm değişkeninin kategorilere göre dağılımı

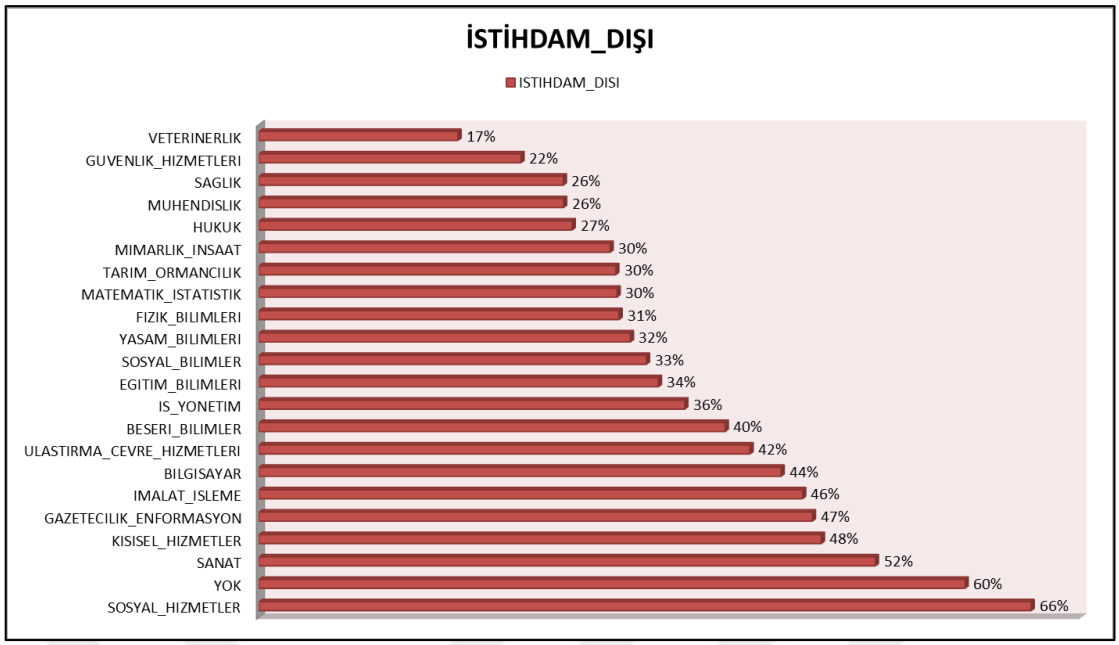
BİTİRİLEN BÖLÜM			
	Sıklık	**Yüzde(%)**	**Kümülatif Yüzde(%)**
BEŞERİ BİLİMLER	19073	1,6	1,6
BİLGİSAYAR	5654	,5	2,1
EĞİTİM BİLİMLERİ	24502	2,1	4,2
FİZİK BİLİMLERİ	3387	,3	4,5

GAZETECİLİK ENFORMASYON	537	,0	4,6
GÜVENLİK HİZMETLERİ	3022	,3	4,8
HUKUK	2870	,2	5,1
İMALAT İŞLEME	11690	1,0	6,1
İŞ YÖNETİM	53996	4,6	10,7
KİŞİSEL HİZMETLER	6260	,5	11,3
MATEMATİK İSTATİSTİK	1926	,2	11,4
MİMARLIK İNŞAAT	7548	,6	12,1
MÜHENDİSLİK	41164	3,5	15,6
SAĞLIK	11509	1,0	16,6
SANAT	6519	,6	17,2
SOSYAL BİLİMLER	11805	1,0	18,2
SOSYAL HİZMETLER	6465	,6	18,7
TARIM ORMANCILIK	3678	,3	19,0
ULAŞT. ÇEVRE HİZMETLERİ	641	,1	19,1
VETERİNERLİK	1216	,1	19,2
YAŞAM BİLİMLERİ	1664	,1	19,3
YOK	938440	80,7	100,0
TOPLAM	1163566	100,0	

Çizelge 4.12’ de HİA çalışmasına katılan bireylerin bitirdikleri bölüme göre işgücü durumu ele alınmıştır. Bitirilen bölümü olmayan bireyler ise “YOK” kategorisine dâhil edilmiştir. Çalışma kapsamındaki bireylerin büyük çoğunluğu bölümü olmayan “YOK” kategorisinde, %4.6’sı “İş ve Yönetim”, %3.5’i “Mühendislik” bölümünü bitirmiştir.



Şekil 4.11. İstihdam kategorisinin bitirilen bölümlere göre dağılımı



Şekil 4.12. İstihdam dışı kategorisinin bitirilen bölümlere göre dağılımı

Şekil 4.11' de istihdam kategorisinde olanlarda en yüksek oran %83 oranında veterinerlik bölümünde görülmüştür. Bu bölümü sırasıyla güvenlik hizmetleri %78 oranında, sağlık ve mühendislik bölümleri %74 takip etmektedir. Matematik ve istatistik bölümünü bitiren bireylerin ise yaklaşık %70' inin istihdamda olduğu görülmüştür.

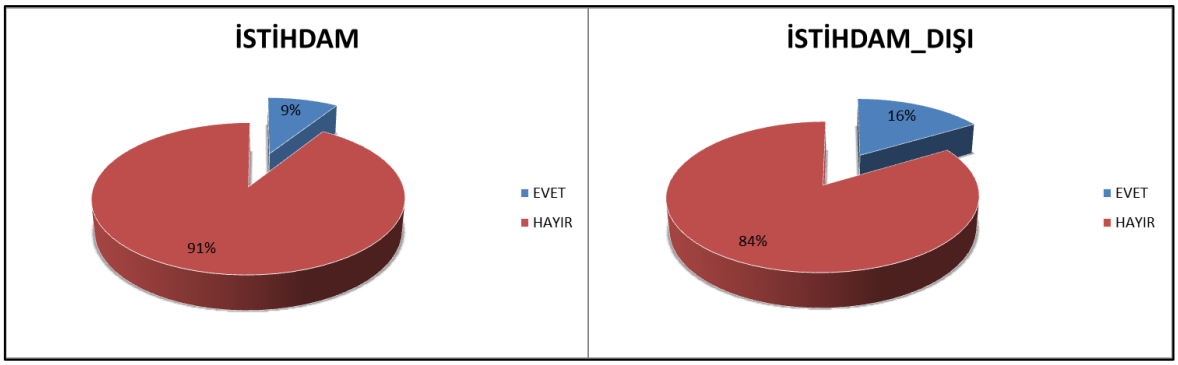
İstihdam oranının en düşük olduğu bölüm ise %34 ile "Sosyal hizmetler" olmuştur, bölümü olmayan "Yok" kategorisindeki bireylerin ise %40'ı istihdam durumunda olduğu görülmüştür.

-İşgücü Durumuna Göre Okula Devam Etme Durumu Dağılımı:

Çizelge 4.13. Eğitime devam etme durumu değişkeninin kategorilere göre dağılımı

EĞİTİME DEVAM DURUMU			
	Sıklık	**Yüzde(%)**	**Kümülatif Yüzde(%)**
EVET	152000	13,1	13,1
HAYIR	1011566	86,9	100,0
TOPLAM	1163566	100,0	

Çizelge 4.13' te HİA çalışmasına katılan bireylerin okula devam edip etmeme durumuna göre işgücü durumu ele alınmıştır.



Şekil 4.13. İşgücü durumu değişkeninin eğitime devam etme durumuna göre dağılımı

Şekil 4.13' te istihdamda olan bireylerin yaklaşık %91'lik kısmı eğitim hayatına devam etmemekte olup kalan %9'luk kısım eğitime devam etmektedir. İstihdam dışında kalan bireylerin yaklaşık %84'ü eğitim hayatına devam etmemekte olup, kalan %16'lık kısım eğitime devam etmek olduğu görülmüştür.

4.5. Veri Madenciliği Yöntemlerinin Uygulanması

HİA çalışması veri setine uygun olan çeşitli sınıflama yöntemleri CRISP-DM döngüsü adımları ile uygulanmış olup, döngüye uygun biçimde yorumlanmıştır.

4.5.1. İşi Anlama

1988 yılından itibaren belirli standartlar çerçevesinde yapılan HİA' nın sağladığı Hanehalkı işgücü profilinin değerlendirilmesi, işgücü durumunun çeşitli faktörler ışığında ele alınması ve ülkemizde işgücü düzeyinin çeşitli yöntemler kullanılarak değerlendirilmesi, sınıflandırılması ve çeşitli modeller yardımıyla yorumlanması amaçlanmaktadır.

Çeşitli örnekleme yöntemleri kullanılarak seçilen hanelerde yaşayan fertlerin tamamına uygulanmak üzere, yalnızca işgücü sorularının kapsamını oluşturan 15 ve daha yukarı yaştaki fertlere ait bilgileri içermektedir. İşgücü verileri incelendiğinden 15 yaş ve üzeri bireyler çalışma kapsamına alınmaktadır. Veri madenciliği yöntemleri kullanılarak, ülkemizde işgücü durum değerlendirmesi yapılabilecek, çeşitli sınıflama ve regresyon yöntemleriyle bireylerin işgücü durumuna ilişkin değerlendirmeler ve tahminler uygulanabilecektir.

4.5.2. Veriyi Anlama

Veriyi anlama aşamasında öncelikle veri setinin toplanması gerekmektedir. HİA çalışmasına ilişkin veri setinin oluşturulmasında, TÜİK' ten talep edilen mikro veri seti

2014, 2015 ve 2016 yıllarını kapsamaktadır. Veri seti 1,163,566 bireye ilişkin işgücü bilgilerini içermektedir. Yaş grubu, medeni hal, cinsiyet, İBBS-1 bölge düzeyi, İBBS-2 bölge düzeyi, okuryazarlık durumu, bitirilen okul, bitirilen bölüm, eğitime devam durumu olmak üzere 9 bağımsız değişken içermektedir.

4.5.3. Veriyi Hazırlama

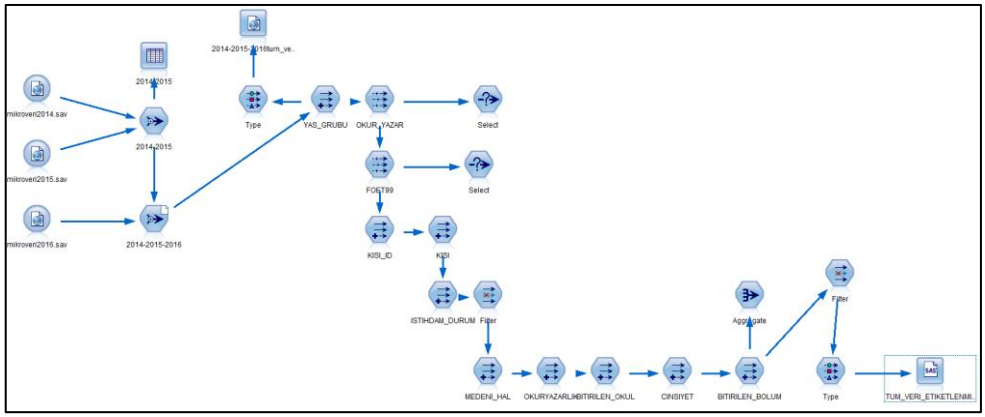
Çalışmanın içerdiği sorulara bağlı olarak veri seti birçok değişken içermektedir. Bağımlı değişkenle etki içinde olduğu düşünülen değişkenler analiz kapsamına alınmıştır. İstihdama katılmayanlar kategorisi çıkarıldığında işsizler ve istihdamdakiler arasında sayısal anlamda büyük bir fark oluştuğundan bu grup işsizlerle birleştirilmiştir ve istihdam dışı olarak adlandırılmıştır.

Field	Measurement	Values	Missing	Check	Role
REFERANS_YIL	Continuous	[2014.0,2016.0]		None	Input
IBBS_1	Nominal	TR1,TR2,TR3,TR4,TR5,TR6,TR7,TR8,TR9,TRA,TRB,...		None	Input
IBBS_2	Nominal	TR10,TR21,TR22,TR31,TR32,TR33,TR41,TR42,TR5...		None	Input
YAS_GRUBU	Nominal	"15-24","25-34","35-44","45-54","55-64","65VEUSTU"		None	Input
ISTIHDAM_DURUM	Flag	ISTIHDAM_DISI/ISTIHDAM		None	Input
MEDENI_HAL	Nominal	DUL,EVLI,HIC_EVLENMEDI		None	Input
OKURYAZARLIK	Nominal	EVET,HAYIR		None	Input
BITIRILEN_OKUL	Nominal	GENEL_LISE,HIC,ILKOGRETIM,MESLEKI_LISE,YUK...		None	Input
CINSIYET	Nominal	E,K		None	Input
BITIRILEN_BOLUM	Nominal	BESERI_BILIMLER,BILGISAYAR,EGITIM_BILIMLERI,...		None	Input
EGITIM_DEVAM	Nominal	EVET,HAYIR		None	Input
PARTITION	Nominal	"1_Training","2_Testing"		None	Partition

Şekil 4.14. Değişkenlerin etiketlenme aşaması

Şekil 4.14' te tip (type) düğümüyle modele girecek bağımlı ve bağımsız değişkenlerin düzeylerine göre nominal, sürekli olmak üzere rol tanımlaması yapılmıştır. İşgücü durum değişkeni bağımlı değişken (target) olarak diğer değişkenler ise bağımsız değişken olarak etiketlenmiştir.

Veri Temizlenmesi aşamasında ekran görüntüsü aşağıdaki şekildedir:



Şekil 4.15. Veri temizlenmesi ekran görüntüsü

Şekil 4.15'te veri setinde yapılan düzenleme, etiketleme, temizleme ve indirgeme işlemleri görsel olarak elde edilmiştir. Medeni hal, yaş grubu, bitirilen okul düzeyi değişkenleri için kategoriler arasında düzenleme yapılmış, diğer bağımsız değişken için ise etiketleme yapılmıştır.

4.5.4. Modelleme

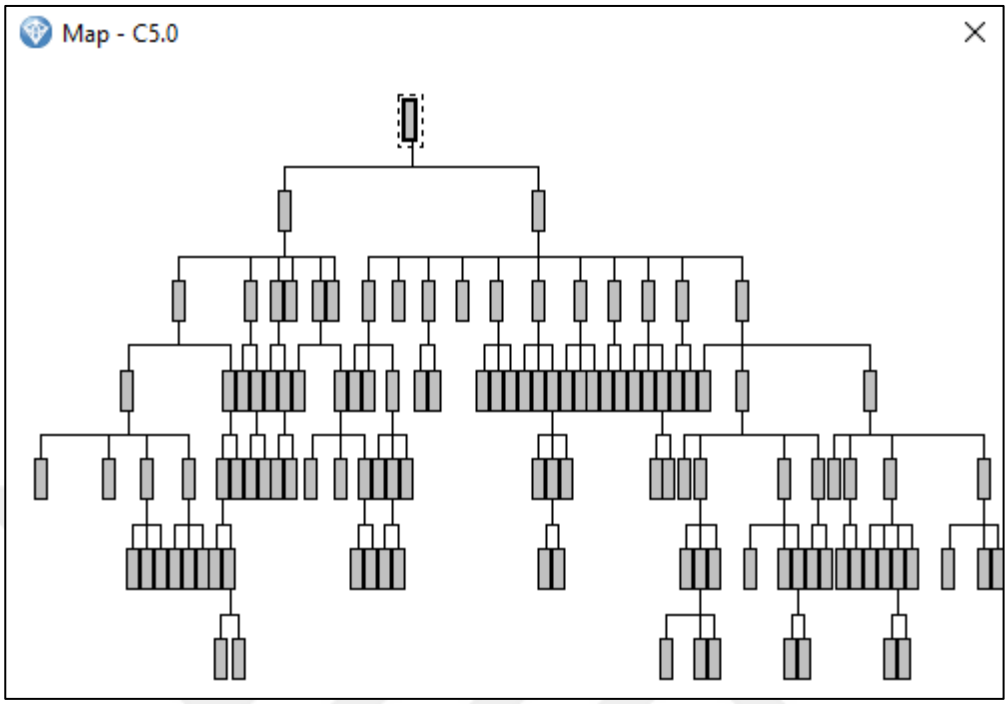
Modelleme aşamasında; öncelikle veri setinin eğitim (training), test (testing) ve geçerlilik (validation) olmak üzere 3'e bölünmesi işlemleri ya da yalnızca eğitim ve test olmak üzere 2 gruba ayırma işlemleri yapılmaktadır. Bu işlem manuel yapılabileceği gibi; paket programlarda bulunan bölümleme (partition) düğümü ile de yapılmaktadır. Çoğunlukla veri setinin 2'ye bölünmesi tercih edilmektedir, rasgele bölünmüş veri setinden, bölümleme oranlarının toplamda %100 olması gerekmektedir. Modelin eğitiminde eğitim verisi kullanır daha sonra test kümesi ile modelin başarısı test edilir [58].

Verilerin bölümlenmesi, çoğunlukla (%80-%20, %70-%30) oranlarında yapılmaktadır. Yapılan çalışmada, veri seti (%80-%20) oranında eğitim ve test kümesi olarak ayrılmıştır.

4.5.4.1. C5.0 Karar Ağacı Yönteminin Uygulanması

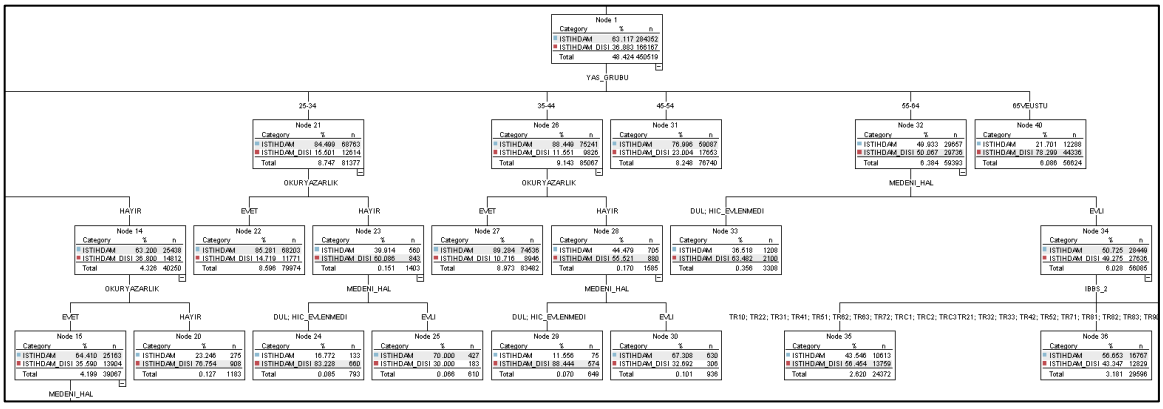
Veri setine uygulanan C5.0 algoritmasına ait çeşitli kombinasyonlarda modeller üretilmiş olup, en uygun model belirlenmesi amaçlanmaktadır. Veri setine uygun olan modelin, budama şiddetinin %50 oranında, oluşturulacak her bir düğüm/yaprak için en az 500 gözlem olacak şekilde bir model kurgulandığı durumda oluştuğu model test edilmiştir. Uygulanan algoritma 6 adımdan oluşmaktadır.

Uygulanan C5.0 algoritması ile bağımsız değişkenlerin bağımlı değişkenin üzerindeki etkisi ve değişkenlerin önem düzeyleri elde edilmiştir. C5.0 algoritmasına göre, bağımsız değişkenlerin bağımlı değişkene etkisinin oldukça az olduğu görülmüştür. İşgücü durumunu etkileyen en önemli değişken cinsiyet ve yaş grubu olmuştur. İBBS-1 ve İBBS-2 değişkeninin etkilerinin oldukça az olduğu görülmüştür.

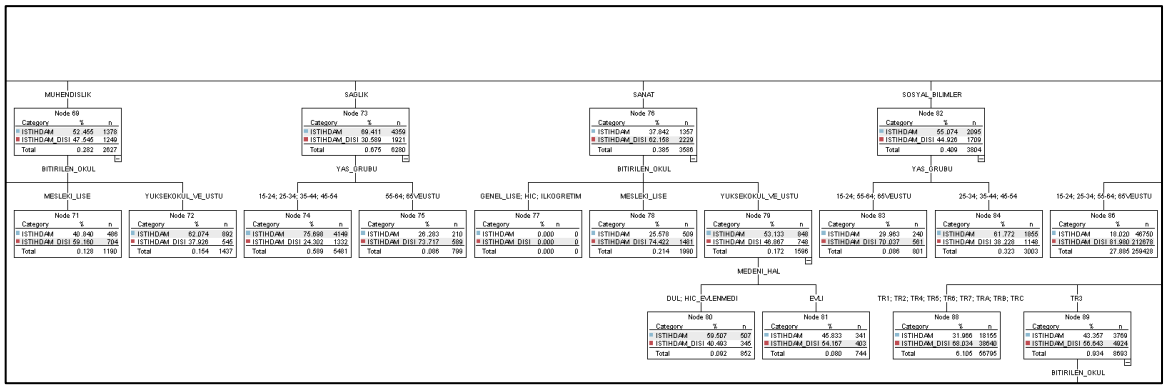


Şekil 4.16. C5.0 karar ağacı haritası

Algoritmadan elde edilen kurallar bütünü model 6 daldan oluşacak şekilde kurgulanmış olup, toplamda 120 adet kural elde edilmiştir. Karar ağacının tamamının görüntüsü alınamadığından kısmi olarak karar ağacının görüntüsü aşağıdaki şekildedir:



Şekil 4.17. C5.0 karar ağacı örneği-1



Şekil 4.18. C5.0 karar ağacı örneği-2

Bağımlı değişkeni etkileyen en önemli değişken cinsiyet olarak belirlendiğinden ağacın ilk kırılımı cinsiyet değişkenine göre olmuştur. Erkek ve kadın bireyler için oluşan düğümlerden sonra, erkeklerde yaş grubuna göre, kadınlarda ise bireylerin bitirdiği bölüme göre kırılımlar oluşmuştur. İlk düğümde bağımlı değişkene göre veri setinin yaklaşık %45'inin istihdam kategorisinde, yaklaşık %55' inin istihdam dışı kategorisinde dağıldığı görülmüştür.

C5.0 karar ağacı algoritması ile elde edilen kurallardan bazıları aşağıdaki şekildedir:

Kural 1 (Düğüm 12): Cinsiyeti "Erkek" olanlardan,

"15-24" yaş aralığında ve

"Okuryazar" lığı olan bireyler %85,3 olasılıkla istihdam kapsamındadır.

Kural 2 (Düğüm 19): Cinsiyeti "Erkek" olanlardan,

15-24 yaş aralığında ise,

"Eğitime Devam" etmiyorsa,

"Okuryazarlığı" varsa,

"Hiç evlenmemiş" ise ve

"Bitirilen okul" değişkeni, "Genel Lise" dışında bir düzey olan bireyler %66,4 olasılıkla istihdam kapsamındadır.

Kural 3 (Düğüm 29): Cinsiyeti "Erkek" olanlardan,

"35-44" yaş aralığında olup,

"okuryazarlığı" olmayan ve

Evli olmayan (Dul, hiç evlenmedi) bireyler %11.5 olasılıkla istihdam kapsamındadır.

Kural 4 (Düğüm 36): Cinsiyeti "Erkek" olanlardan,

"55-64" yaş aralığında olup,

(İBBS-2 düzeyi TR21 Edirne, Tekirdağ, Kırklareli,

TR32 Denizli, Aydın, Muğla,

TR33 Manisa, Afyonkarahisar, Kütahya, Uşak

TR42 Kocaeli, Sakarya, Düzce, Bolu, Yalova

TR52 Konya, Karaman

TR71 Nevşehir, Aksaray, Niğde, Kırıkkale, Kırşehir

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 Kars, Ağrı, Iğdır, Ardahan

TRB1 Malatya, Elazığ, Bingöl, Tunceli

TRB2 Van, Muş, Bitlis, Hakkari) bölgelerinde yaşayanların %56.6'sı istihdam kapsamındadır.

Kural 5 (Düğüm 51): Cinsiyeti "Kadın" olanlardan,

Bitirilen bölümü (Bilgisayar, Gazetecilik-enformasyon, Kişisel hizmetler, Sosyal Hizmetler, Ulaştırma ve Çevre Hizmetleri) olanların

yaklaşık %35.7' si istihdam kapsamındadır.

Kural 6 (Düğüm 54): Cinsiyeti "Kadın" olanlardan,

Bitirilen bölümü Eğitim bilimleri olanlardan,

Yaş grubu "25-54" aralığında olanların

%79' u istihdam kapsamındadır.

Kural 7 (Düğüm 55): Cinsiyeti “Kadın” olanlardan,

Bitirilen bölümü (Fizik bilimleri, Güvenlik hizmetleri, hukuk, matematik ve istatistik, mimarlık-inşaat, tarım-ormancılık, veterinerlik ya da yaşam bilimleri) olanların yaklaşık %60.8’ i istihdam kapsamındadır.

Kural 8 (Düğüm 68): Cinsiyeti “Kadın” olanlardan,

Mezun olunan eğitim düzeyi Yüksekokul ve üzeri olanlardan,

Bitirilen bölümü İş ve Yönetim ise ve

“35-44”,”45-54” yaş aralığında olanların

yaklaşık %61’ i istihdam kapsamındadır.

Kural 9 (Düğüm 66): Cinsiyeti “Kadın” olanlardan,

Mezun olunan eğitim düzeyi Yüksekokul ve üzeri olanlardan,

Bitirilen bölümü İş ve Yönetim ise ve

“25-34” yaş aralığında olanlardan,

Medeni hali “evli” olmayanların %64.4’ ü istihdam kapsamındadır.

Kural 10 (Düğüm 72): Cinsiyeti “Kadın” olanlardan,

Mezun olunan eğitim düzeyi Yüksekokul ve üzeri olanlardan,

Mühendislik bölümünü bitirenlerin yaklaşık %62’si istihdamdadır.

Kural 11 (Düğüm 75): Cinsiyeti “Kadın” olanlardan,

Sağlık bölümü mezunu olanlardan,

“55-64”,”65 ve üzeri” yaş grubunda olanların %26.2’si istihdam kapsamındadır.

Kural 12 (Düğüm 93): Cinsiyeti “Kadın” olanlardan,

Bitirilen bölümü olmayan,

“35-44” yaş aralığında,

İBBS-1 düzeyi TR3 (Ege Bölgesi) olan,

İBBS-2 düzeyi (TR31) İzmir ya da

TR33 (Manisa, Afyonkarahisar, Kütahya, Uşak) olan,

İlköğretim düzeyinde okul bitirmiş bireylerin %57.1’ i istihdam dışındadır.

Kural 13 (Düğüm 98): Cinsiyeti “Kadın” olanlardan,

Bitirilen bölümü olmayan,

“35-44” yaş aralığında olan,

İBBS-1 düzeyine göre TR8 (Batı Karadeniz Bölgesinde) olan,

İBBS-2 düzeyine göre (TR81 Zonguldak, Karabük, Bartın

TR83 Samsun, Tokat, Çorum, Amasya) bölgelerinde olanların yaklaşık %57.2’ si istihdam dışındadır.

Kural 14 (Düğüm 111): Cinsiyeti “Kadın” olanlardan,

Bitirilen bölümü olmayan,

“45-54” yaş aralığında olan,

İBBS-1 düzeyine göre TR9 (Doğu Karadeniz) bölgesinde olanlardan,

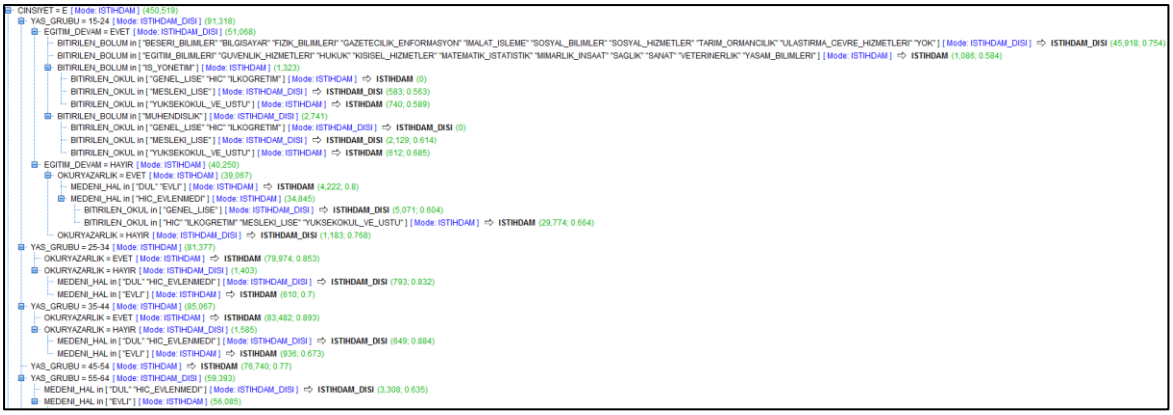
Genel lise düzeyinde eğitim tamamlamış bireylerin %62.3’ü istihdam dışındadır.

Kural 15 (Düğüm 25): Cinsiyeti “Erkek” olanlardan,

“25-34” yaş aralığında olan,

Okuryazarlığı olmayan,

Evli bireylerin %30’u istihdam dışıdır.



Şekil 4.19. C5.0 karar ağacı kural setleri-1

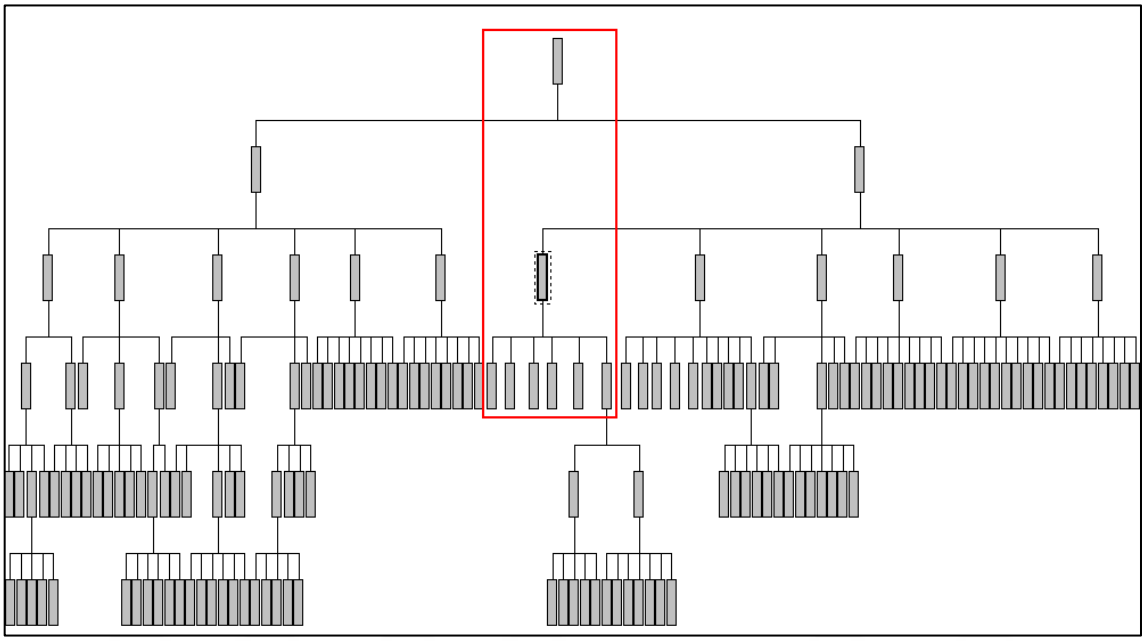
Şekil 4.20. C5.0 karar ağacı kural setleri-2

Şekil 4.19 ve Şekil 4.20' de C5.0 karar ağacı yöntemiyle kurgulanmış olan karar ağacı yapısına ait kural setleri şeklinde görüntüsü elde edilmiştir. Kural setlerinde toplam 120 adet kural bulunmaktadır. İşgücü durumunu etkileyen en önemli değişkenler cinsiyet ve yaş grubu olarak belirlendiğinden ilk düğüm cinsiyete göre olmuştur. C5.0 karar ağacından elde edilen ağaç yapısında erkek bireyler için yaş grubu değişkeni, kadın bireyler için ise bitirilen bölüm değişkeni önemli bulunduğundan cinsiyet değişkeninden sonraki kırılımlar bu değişkenlere göre olmuştur.

4.5.4.2. CHAID Karar Ağacı Yönteminin Uygulanması

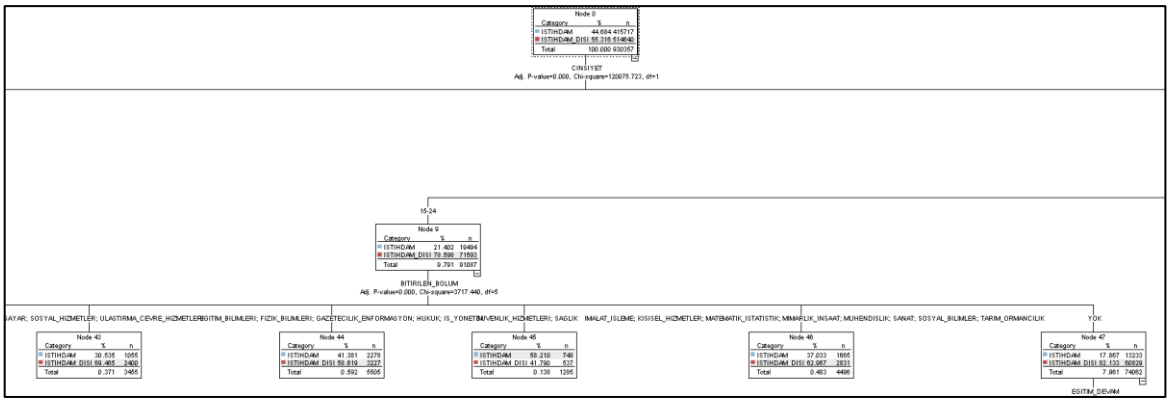
Veri setine uygulanan CHAID karar ağacı yöntemine ait çeşitli kombinasyonlarda kurallar üretilmiş olup, en uygun karar ağacı yapısının belirlenmesi amaçlanmaktadır. Veri setine uygun olan yapının, her düğümde en az %3 oranında, her yaprakta en az %0,75 oranında kayıt olacak şekilde bir model kurgulandığı durumda oluşan model test edilmiştir. Karar ağacının ağaç derinliği maksimum 6 düğüm oluşturacak şekilde düzenlenmiştir. Uygulanan

CHAID yöntemi ile bağımsız değişkenlerin bağımlı değişkenin üzerindeki etkisi ve değişkenlerin önem düzeyleri elde edilmiştir. CHAID yöntemine göre C5.0 yöntemiyle benzer şekilde, en önemli değişken cinsiyet ve yaş grubu olarak elde edilmiştir. Okuryazarlık olup olmaması değişkeni bağımlı değişkene etkisi en az olan değişken olarak belirlenmiştir. İBBS-2 bölge düzeyi değişkeni karar ağacında etkili bir değişken olarak görülmemiştir.

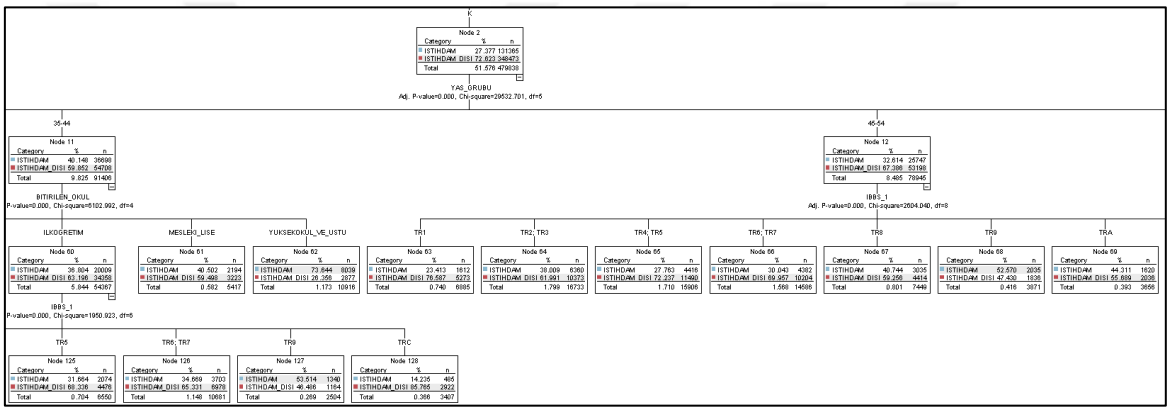


Şekil 4.21. CHAID karar ağacı haritası

Şekil 4.21' de görüldüğü üzere karar ağacından elde edilen kurallar bütünü model 6 daldan oluşacak şekilde kurgulanmış olup, toplam 162 adet kural elde edilmiştir. Karar ağacının tamamının görüntüsü alınamadığından kısmi olarak ağaç görüntüleri Şekil 4.22 ve Şekil 4.23' te olduğu gibi elde edilmiştir:



Şekil 4.22. CHAID karar ağacı örneği-1



Şekil 4.23. CHAID karar ağacı örneği-2

Modelde en önemli değişken cinsiyet olarak belirlendiğinden ağacın ilk kırılımı cinsiyet değişkenine göre olmuştur. İlk düğümde bağımlı değişkene göre veri setinin yaklaşık %45'i istihdam kategorisinde, yaklaşık %55' i istihdam dışı kategorisinde dağılmaktadır.

CHAID karar ağacı algoritması ile elde edilen kurallardan bazıları aşağıdaki şekildedir:

Kural 1 (Düğüm 96): Cinsiyeti "Erkek" olan,

"15-24" yaş aralığında olanlardan,

Eğitime devam etmeyen,

Bitirilen eğitim düzeyi " İlköğretim" olanların

yaklaşık %71.4' ü istihdam kapsamındadır.

Kural 2 (Düğüm 91): Cinsiyeti “Erkek” olan,

“15-24” yaş aralığında olanlardan,

Eğitime devam eden ve bitirilen eğitim düzeyi bulunmayanlar ile Yüksekokul ve üstü olanların %64.2’si istihdam kapsamındadır.

Kural 3 (Düğüm 139): Cinsiyeti “Erkek” olan,

“25-34” yaş aralığında olanlardan,

Medeni hali “Hiç evlenmemiş” olanlar ve “Okuryazarlığı” olan bireylerden,

İBBS-1 bölge düzeyi (TRB Ortadoğu Anadolu, TRC Güneydoğu Anadolu) olanların %63.8’ i istihdam kapsamındadır.

Kural 4 (Düğüm 109): Cinsiyeti “Erkek” olan,

“35-44” yaş aralığında olanlardan,

Medeni hali “Evli” olan,

Bitirilen eğitim düzeyi Yüksekokul ve üstü olanların %97’ si istihdam kapsamındadır.

Kural 5 (Düğüm 25): Cinsiyeti “Erkek” olan,

“45-54” yaş aralığında olanlardan,

Medeni hali “Hiç evlenmemiş” olanların %54.1’i istihdam dışıdır.

Kural 6 (Düğüm 150): Cinsiyeti “Erkek” olan,

“45-54” yaş aralığında olanlardan

Medeni hali “Evli” olanlar,

Bitirilen eğitim düzeyi “İlköğretim, Genel lise” olanlardan,

İBBS-1 bölge düzeyi (TR9 Doğu Karadeniz, TRA Kuzeydoğu Anadolu) olanların %83.1’ istihdam kapsamındadır.

Kural 7 (Düğüm 39): Cinsiyeti “Erkek” olan,

“65 ve üstü” yaş aralığında olanlardan,

İBBS-1 bölge düzeyi TR9 Doğu Karadeniz olanların %65.8’ i istihdam dışı kapsamındadır.

Kural 9 (Düğüm 47): Cinsiyeti “Kadın” olan,

“15-24” yaş aralığında olanlardan,

Bitirilen bölümü olmayanların %82.1’ i istihdam dışındadır.

Kural 10 (Düğüm 52): Cinsiyeti “Kadın” olan,

“25-34” yaş aralığında olanlardan,

Bitirilen bölümü “Güvenlik Hizmetleri, Hukuk, Matematik-İstatistik” olanların %71.1’i istihdam kapsamındadır.

Kural 11 (Düğüm 55): Cinsiyeti “Kadın” olan,

“25-34” yaş aralığında olanlardan,

Bitirilen bölümü “Sağlık, Veterinerlik” olanların %81.9’ u istihdamdadır.

Kural 12 (Düğüm 121): Cinsiyeti “Kadın” olan,

“25-34” yaş aralığında olanlardan,

Bitirilen bölümü olmayan,

İBBS-1 bölge düzeyi TRC Güneydoğu Anadolu olanların %88.1’ i istihdam dışındadır.

Kural 13 (Düğüm 122): Cinsiyeti “Kadın” olan,

“35-44” yaş aralığında olanlardan,

Bitirilen eğitim düzeyi “İlköğretim” olan,

İBBS-1 bölge düzeyi (TR1 İstanbul, TRB Ortadoğu Anadolu) olanların %71.8’ i istihdam dışındadır.

Kural 14 (Düğüm 127): Cinsiyeti “Kadın” olan,

“35-44” yaş aralığında olanlardan,

Bitirilen eğitim düzeyi “İlköğretim” olan,

İBBS-1 bölge düzeyi TR9 Doğu Karadeniz olanların %53.5’ i istihdam kapsamındadır.

Kural 15 (Düğüm 68): Cinsiyeti “Kadın” olan,

“45-54” yaş aralığında olanlardan,

İBBS-1 bölge düzeyi TR9 Doğu Karadeniz olanların %52.5' i istihdam kapsamındadır.

Kural 16 (Düğüm 42): Cinsiyeti "Kadın" olan,

"15-24" yaş aralığında olanlardan,

Bitirilen bölümü "Beşeri Bilimler, Veterinerlik, Yaşam Bilimleri" olan bireylerin %77.4' ü istihdam dışındadır.

Kural 17 (Düğüm 56): Cinsiyeti "Kadın" olan,

"25-34" yaş aralığında olanlardan,

Bitirilen bölümü "Sosyal Hizmetler" olan bireylerin %63'ü istihdam dışındadır.

Kural 18 (Düğüm 104): Cinsiyeti "Kadın" olan,

"25-34" yaş aralığında olanlardan,

Medeni hali "Hiç evlenmemiş" olanlar,

ve "Okuryazarlığı" olmayanların %83.7'si istihdam dışındadır.

Genel olarak C5.0 ve CHAID karar ağaçları değerlendirildiğinde, ağaç yapısındaki dallanmalar her iki karar ağacı için de 2' den fazla gruplar halinde elde edilmektedir. Karar ağaçları farklı ölçütler kullanarak kırılımlar oluşturduğundan ağaç yapıları aynı biçimde oluşmamıştır. Fakat her 2 karar ağacı sonuçlarından veri setine ilişkin önemli değişkenlerin tahmini benzer şekilde yapılmıştır.

4.5.4.3. Lojistik Regresyon Yönteminin Uygulanması

İşgücü durumuna etki eden değişkenlerin modellenmesi için en uygun yöntemi belirlerken, bağımlı değişkenin 2 kategorili olması göz önünde bulundurularak Lojistik regresyon yöntemi kullanılmıştır. Lojistik regresyon ile işgücü durumunu etkileyen ve ele alınan değişkenlerin anlamlı olup olmadığı yorumlanmış ve değişkenlerin kategorileri arasında odds oranları elde edilerek yorumlamalar yapılmıştır. Veri setine uygun modelin belirlenmesi için yapılan uygulamalardan sonra, değişkenlerin modele alınması için en uygun metodun adımsal olarak seçim yapan "stepwise" metodu seçilmiş ve yalnızca temel etkilerin görüldüğü "main effects" durumlarının seçildiği model uygun bulunmuştur. Adımsal (stepwise) metodunda veri setindeki değişkenlerin adımsal olarak modele girmesi ve modelden çıkarılmasında olabilirlik oran kriteri kullanılmış olup, Modele alınacak değişkenler için ($\alpha = 0,05$) olarak, modelden çıkarılacak değişkenler için ($\alpha = 0,10$) olarak

alınmıştır. Bağımlı değişken için baz alınan kategori model kurulurken değiştirilebilmektedir, işgücü durumu olan y bağımlı değişkeni için "istihdam dışı" durumu referans kategori olarak alınmıştır.(y=0 istihdam dışı, y=1 istihdam) Lojistik regresyon yöntemi ile test edilecek hipotezler kurularak yorumlanmıştır. Kurulan lojistik regresyon modelinde, en önemli değişkenin sırasıyla bitirilen yaş grubu ve cinsiyet değişkeni olduğu görülmüştür. Önem düzeyi en düşük olan değişken İBBS-1 bölge düzeyi olarak belirlenmiştir. Aşağıdaki tabloda bağımlı ve bağımsız değişkenlerin sıklıklarına ilişkin tablo aşağıdaki şekilde elde edilmiştir.

Çizelge 4.14. Durum özet tablosu

		N	YÜZDE(%)
İŞGÜCÜ DURUMU	**İSTİHDAM**	415,717	44,70%
	İSTİHDAM DIŞI	514,64	55,30%
İBBS1	**TR1**	82,904	8,90%
	TR2	66,18	7,10%
	TR3	112,793	12,10%
	TR4	75,5	8,10%
	TR5	103,072	11,10%
	TR6	103,2	11,10%
	TR7	61,318	6,60%
	TR8	83,608	9,00%
	TR9	43,022	4,60%
	TRA	52,699	5,70%
	TRB	61,54	6,60%
	TRC	84,521	9,10%
İBBS2	**TR10**	82,904	8,90%
	TR21	30,994	3,30%
	TR22	35,186	3,80%
	TR31	42,13	4,50%
	TR32	35,702	3,80%
	TR33	34,961	3,80%
	TR41	38,396	4,10%
	TR42	37,104	4,00%
	TR51	51,812	5,60%
	TR52	51,26	5,50%
	TR61	33,584	3,60%
	TR62	40,508	4,40%
	TR63	29,108	3,10%
	TR71	33,214	3,60%
	TR72	28,104	3,00%
	TR81	19,241	2,10%
	TR82	27,407	2,90%
	TR83	36,96	4,00%
	TR90	43,022	4,60%
	TRA1	26,306	2,80%
	TRA2	26,393	2,80%
	TRB1	26,605	2,90%
	TRB2	34,935	3,80%
	TRC1	30,761	3,30%
	TRC2	31,472	3,40%
	TRC3	22,288	2,40%

CİNSİYET	**ERKEK**	450,519	48,40%
	KADIN	479,838	51,60%
YAŞ GRUBU	**15-24**	182,405	19,60%
	25-34	168,216	18,10%
	35-44	176,473	19,00%
	45-54	155,685	16,70%
	55-64	121,595	13,10%
	65VEÜSTÜ	125,983	13,50%
MEDENİ HAL	**DUL**	84,581	9,10%
	EVLİ	623,024	67,00%
	HİÇ EVLENMEDİ	222,752	23,90%
OKURYAZARLIK	**EVET**	819,162	88,00%
	HAYIR	111,195	12,00%
BİTİRİLEN OKUL	**GENEL LİSE**	81,882	8,80%
	HİÇ	168,629	18,10%
	İLKOĞRETİM	500,13	53,80%
	MESLEKİ LİSE	69,085	7,40%
	YÜKSEKOKUL VE ÜSTÜ	110,631	11,90%
BİTİRİLEN BÖLÜM	**BEŞERİ BİLİMLER**	15,273	1,60%
	BİLGİSAYAR	4,545	0,50%
	EĞİTİM BİLİMLERİ	19,676	2,10%
	FİZİK BİLİMLERİ	2,675	0,30%
	GAZETECİLİK ENFORMASYON	421	0,00%
	GÜVENLİK HİZMETLERİ	2,422	0,30%
	HUKUK	2,275	0,20%
	İMALAT İŞLEME	9,348	1,00%
	İŞ YÖNETİM	42,913	4,60%
	KİŞİSEL HİZMETLER	4,996	0,50%
	MATEMATİK İSTATİSTİK	1,545	0,20%
	MİMARLIK İNŞAAT	6,034	0,60%
	MÜHENDİSLİK	32,812	3,50%
	SAĞLIK	9,239	1,00%
	SANAT	5,176	0,60%
	SOSYAL BİLİMLER	9,44	1,00%
	SOSYAL HİZMETLER	5,191	0,60%
	TARIM ORMANCILIK	2,941	0,30%
	ULAŞTIRMA ÇEVRE HIZMETLERİ	494	0,10%
	VETERİNERLIK	980	0,10%
	YAŞAM BİLİMLERİ	1,319	0,10%
	YOK	750,642	80,70%
EĞİTİM DEVAM	**EVET**	121,149	13,00%
	HAYIR	809,208	87,00%
TOPLAM		930,357	

Çizelge 4.14’ te her kategori için sıklıklar yer almaktadır. Veri setinin yaklaşık %45’i istihdam, yaklaşık %55’ i istihdam dışı kapsamında yer almaktadır.

Çizelge 4.15. Model anlamlılık tablosu

MODEL	MODEL UYGUNLUK KRİTERİ	OLABİLİRLİK ORAN TESTİ		
	- 2LOG (L)	Kİ-KARE	SERBESTLİK DERECESİ	P OLASILIK DEĞERİ
SABİT TERİMLİ MODEL	433281.650			
SON MODEL	135153.902	298127.748	60	0.000

İşgücü durumu bağımlı değişken olmak üzere (y değişkeni),

H_0: $\beta_1 = \beta_2 = \ldots = \beta_9 = 0$

H_S: En az bir β_i 0' dan farklıdır

Hipotez test edilirken Çizelge 4.15' teki P olasılık değeri α değeri ile kıyaslanmaktadır. P(Sig.) değeri < α = 0,05 olduğundan hipotez reddedilmektedir. Buna göre, kurulan model anlamlıdır. Bir başka ifadeyle, en az bir β_i' nin 0' dan farklı olduğu görülmüştür.

Çizelge 4.16. Olabilirlik oran testleri tablosu

ETKİ	MODEL UYGUNLUK KRİTERİ	OLABİLİRLİK ORAN TESTLERİ		
	- 2LOG (L)	Kİ-KARE	SERBESTLİK DERECESİ	P DEĞERİ
SABİT	135.153.902	0	0	.
İBBS2	147501,538	12347,637	25	0
CİNSİYET	255815,700	120661,799	1	0
YAŞ GRUBU	227227,394	92073,493	5	0
MEDENİ HAL	136917,111	1763,210	2	0
OKURYAZARLIK	135211,424	57,522	1	0
BİTİRİLEN OKUL	137820,436	2666,535	4	0
BİTİRİLEN BÖLÜM	137179,677	2025,776	21	0
EĞİTİM DEVAM	142737,160	7583,258	1	0

H_o: $\beta_i = 0$, i=1,2,…,8 (Lojistik regresyon katsayısı sıfırdır.)

H_s: $\beta_i \neq 0$ (Lojistik regresyon katsayısı sıfırdan farklıdır.)

Çizelge 4.16' da Olabilirlik oran testi tablosunda modeldeki değişkenlere ilişkin -2Log (L) gibi model belirleme kriteri, ki-kare değeri ve bağımsız değişkenlerin P değerleri yer

almaktadır. İBBS-2 bölge düzeyi, cinsiyet, yaş grubu, medeni hal, okuryazarlık durumu, bitirilen okul düzeyi, bitirilen bölüm, eğitime devam etme durumu değişkenleri için ki-kare değerine ilişkin P(Sig.)değeri < α=0,05 durumunu sağladığından H_o hipotezi reddedilmektedir. İBBS-1 bölge düzeyi değişkenin modele etkisi anlamlı bulunmadığından bu değişken modelden çıkarılmıştır. P(Sig.)değeri=0,00 < α=0,05 olan bağımsız değişken katsayıları sıfırdan farklıdır, bağımlı değişkenlerin modele etkisi anlamlıdır. Ayrıca tabloda hangi değişkenin modelde daha fazla etki yaratacağı da model belirleme kriterleri ile yorumlanmaktadır. Her değişken için elde edilen, bağımlı değişkendeki varyans değişim etkisini gösteren, -2Log Likelihood(-2LL) değeri 255815,700 olan cinsiyet değişkeni modele en fazla etkisi olan değişkendir. Cinsiyet değişkeninden sonra yaş grubunun etkisinin yüksek olduğu belirlenmiştir.

Çizelge 4.17. Parametre tahmin tablosu

PARAMETRE TAHMİNİ								
İŞGÜCÜ DURUMU(A)	**B**	**STD. HATA**	**WALD DEĞERİ**	**SD**	**P**	**ODDS ORANI (EXP(B))**	**95% GÜVEN ARALIĞI(ODDS)**	
							ALT SINIR	**ÜST SINIR**
SABİT	-17,351	,070	61431,249	1	0			
[EĞİTİMDEVAM=EVET]	-,794	,009	7516,820	1	0	,452	,444	,460
[EĞİTİMDEVAM=HAYIR]	0			0				
[CİNSİYET=ERKEK]	1,807	,006	105831,18	1	0	6,094	6,028	6,161
[CİNSİYET=KADIN]	0			0				
[OKURYAZARLIK=EVET]	,102	,013	57,612	1	0	1,107	1,078	1,137
[OKURYAZARLIK=HAYIR]	0			0				
[YAŞ GRUBU =15-24]	1,818	,014	17154,347	1	0	6,162	5,997	6,332
[YAŞ GRUBU=25-34]	2,567	,012	49169,554	1	0	13,020	12,728	13,319
[YAŞ GRUBU =35-44]	2,745	,011	59792,786	1	0	15,570	15,231	15,916
[YAŞ GRUBU =45-54]	2,277	,011	42744,678	1	0	9,746	9,538	9,958
[YAŞ GRUBU =55-64]	1,287	,011	13116,401	1	0	3,623	3,544	3,703
[YAŞGRUBU=65VE ÜSTÜ]	0			0				

[IBBS2=TR10]	,774	,019	1642,045	1	0	2,167	2,088	2,250
[IBBS2=TR21]	1,195	,022	2965,005	1	0	3,304	3,165	3,449
[IBBS2=TR22]	,896	,022	1725,845	1	0	2,451	2,349	2,557
[IBBS2=TR31]	,891	,021	1832,314	1	0	2,439	2,341	2,540
[IBBS2=TR32]	1,146	,021	2870,537	1	0	3,144	3,015	3,279
[IBBS2=TR33]	1,141	,021	2820,703	1	0	3,129	30	3,263
[IBBS2=TR41]	,780	,021	1361,759	1	0	2,182	2,093	2,274
[IBBS2=TR42]	,986	,021	2161,879	1	0	2,681	2,572	2,795
[IBBS2=TR51]	,719	,020	1262,382	1	0	2,053	1,973	2,136
[IBBS2=TR52]	,945	,020	2191,689	1	0	2,574	2,474	2,678
[IBBS2=TR61]	1,025	,022	2255,767	1	0	2,788	2,672	2,908
[IBBS2=TR62]	,827	,021	1575,549	1	0	2,285	2,194	2,381
[IBBS2=TR63]	,436	,022	381,828	1	0	1,546	1,480	1,616
[IBBS2=TR71]	,930	,022	1850,680	1	0	2,534	2,429	2,643
[IBBS2=TR72]	,814	,022	1326,374	1	0	2,257	2,161	2,359
[IBBS2=TR81]	1,149	,024	2233,466	1	0	3,156	3,009	3,310
[IBBS2=TR82]	1,403	,023	3861,679	1	0	4,069	3,893	4,253
[IBBS2=TR83]	1,071	,021	2544,895	1	0	2,918	2,799	3,041
[IBBS2=TR90]	1,340	,021	4166,950	1	0	3,820	3,668	3,979
[IBBS2=TRA1]	1,094	,023	2351,995	1	0	2,985	2,856	3,120
[IBBS2=TRA2]	1,355	,022	3655,532	1	0	3,877	3,710	4,051
[IBBS2=TRB1]	,901	,023	1594,223	1	0	2,462	2,355	2,573
[IBBS2=TRB2]	,720	,021	1147,850	1	0	2,054	1,970	2,141
[IBBS2=TRC1]	,514	,022	550,978	1	0	1,672	1,602	1,746
[IBBS2=TRC2]	,414	,022	360,265	1	0	1,513	1,450	1,579
[IBBS2=TRC3]	0			0				
[MEDENİ HAL=DUL]	,052	,014	13,901	1	0	1,053	1,025	1,082
[MEDENİ HAL=EVLİ]	,317	,009	1242,553	1	0	1,372	1,349	1,397
[MEDENİ HAL=HİÇ EVLENMEDİ]	0			0				

[BİTİRİLENOKUL= GENEL LİSE]	12,964	,066	38783,870	1	0	426885,21	375211,97	485674,76
[BİTİRİLENOKUL=Hİ Ç]	13,039	,066	38771,894	1	0	459800,58	403836,29	523520,49
[BİTİRİLEN OKUL=İLKÖĞRETİM]	13,027	,065	39619,888	1	0	454706,36	399964,04	516941,15
[BİTİRİLENOKUL=ME SLEKİ LİSE]	-,700	,014	2569,552	1	0	,497	,483	,510
[BİTİRİLENOKUL=YU KSEKOKULVEÜSTÜ]	0			0				
[BİTİRİLENBÖLÜM= BEŞERİ BİLİMLER]	13,904	,069	41146,766	1	0	1092813,0	955431,75	1249948,3
[BİTİRILENBÖLÜM= BİLGİSAYAR]	13,921	,074	35212,428	1	0	1110785,6	960472,38	1284622,9
[BİTİRİLENBÖLÜM= EĞİTİM BİLİMLERİ]	14,025	,067	43185,666	1	0	1233371,6	1080553,1	1407802,7
[BİTİRİLENBÖLÜM= FİZİK BİLİMLERİ]	13,880	,080	29925,288	1	0	1066618,1	911405,82	1248263,2
[BİTİRİLENBÖLÜM=G AZETECİLİK ENFORMASYON]	13,588	,128	11252,351	1	0	796843,60	619918,61	1024263,0
[BİTİRİLENBÖLÜM=G ÜVENLİK HİZMETL	13,844	,085	26671,134	1	0	1028435,4	871010,23	1214313,5
[BİTİRİLENBÖLÜM=H UKUK]	14,436	,084	29430,043	1	0	1859477,4	1576753,1	2192896,4
[BİTİRİLENBÖLÜM= İMALAT İŞLEME]	13,870	,070	38978,603	1	0	1056470,6	920568,84	1212435,3
[BİTİRİLENBÖLÜM= İŞ YÖNETİM]	13,903	,066	43763,795	1	0	1091088,9	957837,35	1242878,0
[BİTİRİLENBÖLÜM=K İŞİSEL HİZMETLER]	13,771	,073	35109,776	1	0	956635,33	828299,63	1104855,2
[BİTİRİLENBÖLÜM= MAT. İSTATİSTİK]	14,040	,090	24399,339	1	0	1251109,4	1049033,0	1492112,0
[BİTİRİLENBÖLÜM= MİMARLIK İNŞAAT]	13,932	,073	36726,513	1	0	1123312,3	974138,38	1295330,0
[BİTİRİLENBÖLÜM= MÜHENDİSLİK]	14,050	,068	43288,048	1	0	1264778,0	1107979,1	1443766,7
[BİTİRİLENBÖLÜM= SAĞLIK]	14,985	,071	44949,388	1	0	3218770,2	2802394,3	3697010,6
[BİTİRİLENBÖLÜM=S ANAT]	13,789	,073	35748,551	1	0	973364,75	843723,32	1122926,1

[BİTİRİLENBÖLÜM= SOSYAL BİLİMLER]	13,763	,070	38877,393	1	0	948936,29	827601,98	1088059,3
[BİTİRİLENBÖLÜM= SOSYAL HİZMETLER]	14,016	,073	36564,576	1	0	1222270,8	1058704,7	1411107,3
[BİTİRİLENBÖLÜM= TARIM ORMANCILIK]	13,798	,079	30165,037	1	0	982419,16	840764,94	1147939,6
[BİTİRİLENBÖLÜM= ULAŞTIRMA ÇEVRE HİZMETLERİ]	13,691	,122	12537,785	1	0	883339,46	695098,78	1122557,8
[BİTİRİLENBÖLÜM= VETERİNERLİK]	14,708	,114	16629,412	1	0	2440539,8	1951657,9	3051884,6
[BİTİRİLENBÖLÜM= YAŞAM BİLİMLERİ]	14,162	0		1		1414532,2	1414532,2	1414532,2
[BİTİRİLENBÖLÜM= YOK]	0			0				
A. Referans alınan kategori: istihdam dışı								

Çizelge 4.17' de kurulan modele ilişkin β parametreleri, parametrelere ilişkin Wald değerleri, serbestlik dereceleri, önem düzeyleri ve Odds Oranları (Exp(β)) ve Odds Oranı için güven aralıkları yukarıdaki tabloda elde edilmiştir.

H_o: $\beta_i = 0$, i=1,2,…,8 (Lojistik regresyon katsayısı sıfırdır.)

H_s: $\beta_i \neq 0$ (Lojistik regresyon katsayısı sıfırdan farklıdır.)

Modeldeki katsayıların anlamlılığının test edilmesinde bir başka yöntem olan Wald istatistiği değerleri ile de yorumlanabilmektedir. Wald istatistiğinin P değerine göre, bağımsız değişkenlerin P(Sig.)değeri=0,00 < α=0,05 olduğundan H_0 hipotezi reddedilmektedir, Modele alınan İBBS-1 değişkeni dışındaki tüm bağımsız değişkenler 0' dan farklıdır, modele alınan katsayıların istatistiksel olarak anlamlı olduğu görülmüştür.

İşgücü durumu değişkeni için, bireyin istihdamda olması ile istihdam dışında olması olasılığına oranı olan odds oranları elde edilmiştir. Bu değerler de Exp(β) olarak tanımlanan sütundaki değerlerdir. Bağımsız değişken yani x değeri 1 olan bireylerin x değeri 0 olan bireylere göre bağımlı değişkenin kaç kat daha fazla 1 olarak görüldüğü sonucunu vermektedir. Referans kategori düzeyi İstihdam dışı olmak üzere İstihdam düzeyi için,

Yaş grubu değişkeninde, 65 yaş ve üstü bireylere göre 55-64 yaş grubundaki bireylerin istihdamda olma olasılığı 3,62 kat daha fazladır. 65 yaş ve üstü bireylere göre 15-24 yaş grubundaki bireylerin istihdamda olma olasılığı 6,16 kat daha fazladır.

Medeni hal değişkeninde, "Hiç Evlenmeyenler" kategorisi referans kategori düzeyi olarak ele alınmış olup; medeni hali "Dul" olan bireylerin hiç evlenmeyenlere göre istihdam kategorisinde olma olasılığı 1,05 kat daha fazla iken, medeni hali "Evli" olan bireylerin hiç evlenmeyenlere göre istihdam dışı kategorisinde olma olasılığı 1,37 kat daha fazladır.

Okuryazarlığı olmayan bireyler referans alındığında, okuryazarlığı olanların olmayanlara göre istihdamda olma olasılığı 1,1 kat daha fazladır.

Eğitime devam etme durumu değişkenine ait β katsayı negatif olarak elde edilmiştir. Bağımlı değişkene etkisi negatif olduğundan (β katsayısı negatif) odds oranları (1 – odds)alınarak yorumlanmaktadır [43]. Eğitime devam etmeyenler kategorisi referans alındığında, eğitime devam edenlerin etmeyenlere göre istihdamda olma olasılığı (1-odds)=0,54 kat daha azdır.

Cinsiyet değişkeni detayında, "Kadın" kategorisi referans kategori düzeyi olarak ele alınmış olup; Erkek bireylerin kadın bireylere göre istihdam kategorisinde olma olasılığı yaklaşık 6,09 kat daha fazladır.

Lojistik regresyon modeli ile kurulan modelde işgücü durumu (istihdam / istihdam dışı) bağımlı değişken olmak üzere (Y değişkeni),

y = 1 durumu ele alınmak üzere, $P(y_i)$ fonksiyonunda bireylerin istihdamda olma olasılığının aşağıdaki şekilde modellenmesi mümkündür [42,43]:

$$E(y_i) = P(y_i) = \ln\left(\frac{Pi}{1-Pi}\right) = \frac{e^{(\beta_0+\beta_1 x_{1i}1\mathrm{i}+\beta_2 x_{2i}+\ldots\ldots\ldots+\beta_k x_{1ki}}}{1+e^{(\beta_0+\beta_1 x_{1i}1\mathrm{i}+\beta_2 x_{2i}+\ldots\ldots\ldots+\beta_k x_{1ki}}} \quad (4.1)$$

Veri setine ilişkin lojistik regresyon modeli aşağıdaki gibidir :

$e^{(\beta_0+\beta_1 x_{1i}1\mathrm{i}+\beta_2 x_{2i}+\ldots\ldots\ldots+\beta_k x_{1ki})}$ = exp((-17,35 + 0,77*TR10 + 1,19*TR21 + 0,89*TR22 + 0,89*TR31 + 1,14*TR33 + 0,78*TR41 + 0,98*TR42 + 0,71*TR51 + 0,94*TR52 + 1,02*TR61 + 0,82*TR62 + 0,43*TR63 + 0,92*TR71 + 0,81*TR72 + 1,14*TR81 + 1,403*TR82 + 1,07*TR83 + 1,34*TR90 + 1,09*TRA1 + 1,35*TRA2 + 0,9*TRB1 + 0,71*TRB2 + 0,51*TRC1 + 0,41*TRC2 + 1,8*ERKEK + 1,81*"15-24" + 2,56*"25-34" + 2,74*"35-44" + 2,27*"45-54" + 1,28*"55-64" + 0,05*"DUL" + 0,31*"EVLİ" + 0,10*"OKURYAZAR" + 12,96*"GENEL_LİSE" + 13,04*"BITIRILEN_OKUL_YOK" + 13,03*"İLKÖĞRETİM" - 0,69*"MESLEKİ_LİSE" + 13,9*"BEŞERİ_BİLİMLER" + 13,92*"BİLGİSAYAR" + 14,03*"EĞİTİM_BİLİMLERİ" + 13,88*"FİZİK_BİLİMLERİ" + 13,59*"GAZETECİLİK" + 13,84*"GÜVENLİK" + 14,44*"HUKUK" +

13,87*"İMALAT" + 13,9*"İŞ_YÖNETİMİ" + 13,77*"KİŞİSEL_HİZMET" + 14,04*"MATEMATİK_İSTATİSTİK" + *13,93*"MİMARLIK" + 14,05*"MÜHENDİSLİK" + 14,98*"SAĞLIK" + 13,79*"SANAT" + 13,76*"SOSYAL_BİLİMLER" + 14,02*"SOSYAL_HİZMETLER" + 13,8*"TARIM_ORMAN" + 13,69*"ULAŞTIRMA" + 14,71*"VETERİNERLİK" + 14,16*"YAŞAM_BİLİMLERİ"-0,79*"EĞİTİME_DEVAM_EDEN")) olmak üzere model kurulmaktadır.

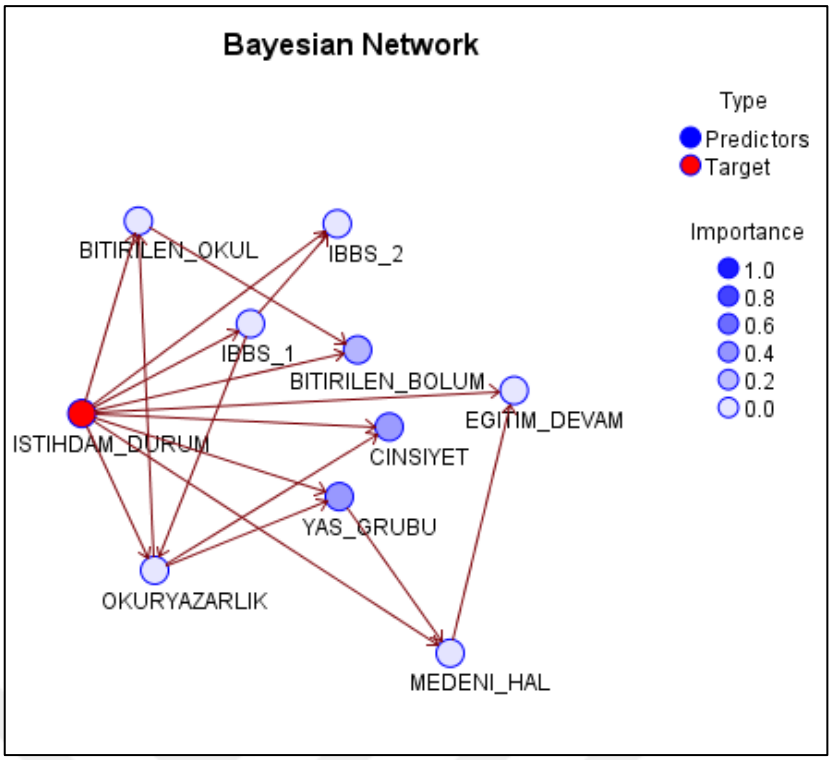
Çizelge 4.18. Sınıflama tablosu

SINIFLAMA			
GERÇEK	**TAHMİN**		
	İSTİHDAM	**İSTİHDAM DIŞI**	**DOĞRU SINIFLAMA (%)**
İSTİHDAM	277,111	138,606	66,7%
İSTİHDAM DIŞI	92,175	422,465	82,1%
TOPLAM(%)	39,7%	60,3%	75,2%

Çizelge 4.18' de veri setindeki gözlemler ve tahmin edilen gözlemlere ilişkin oranlar verilmiştir. İşgücü durumu üzerine kurulan lojistik regresyon yöntemi sonucunda elde edilen model istihdam kategorisinde gerçek durumda 415,717 bireyden 277,111 bireyi istihdamda olarak tahmin etmiştir. İstihdam kategorisi için %66,7 oranında doğru sınıflamada bulunmuş, istihdam dışı kategorisi için ise %82,1 oranında doğru sınıflama yapmıştır. Modelin her iki kategori için sınıflamada genel doğruluk oranı %75,2 olarak elde edilmiştir.

4.5.4.4. Bayes Ağları Yönteminin Uygulanması

İstihdam durumuna etki eden değişkenlerin modellenmesi için en uygun yöntemi belirlerken, çeşitli sınıflama ve regresyon yöntemler kullanılmış olup, son olarak Bayes ağları algoritmasından Bayes Ağları(Bayes Net) düğümü yardımıyla algoritma uygulanmıştır. Veri madenciliğinde uygulaması olan Bayes ağları algoritmalarında bulunan sınıflayıcılardan Markov ağ yapısı ve Ağaç Artırılmış Naive Bayes (TAN) olarak adlandırılan ağ yapıları olmak üzere 2 ağ yapısı bulunmaktadır. Bağımlı değişken işgücü durumu olmak üzere bağımsız değişkenler Bayes algoritmasında modele dahil edilmiştir. Diğer algoritmalarda uygulandığı gibi Bölümlenmiş (Partition) veri seti kullanılmıştır. Hem Markov ağ yapısı hem de Ağaç Artırılmış Naive Bayes (TAN) ağ yapısı ile Bayes ağları veri setine uygulanmıştır. İlişkisel olarak veri setine en uygun model için TAN ağ yapısıyla elde edilen sonuçlar yorumlanmıştır.



Şekil 4.24. Model değişkenlerinin önem düzeyi

Bayes ağlarında, bağımlı ve bağımsız değişkenler arasındaki ilişkisel durum grafiksel olarak gösterilmektedir. İstihdam durumunun bağımsız değişkenlerle olasılıksal olarak ilişkisi ok geçişleriyle gösterilmektedir. Grafikte oklarla ilişkilendirilmeyen değişkenler arasında ilişki de bulunmamaktadır. Şekil 4.24' te değişkenlerin önem düzeyi hakkında bağımlı değişkeni etkileyen en önemli bağımsız değişkenlerin yaş grubu ve cinsiyet olduğu görülmüştür. Bağımsız değişkenler arasında ise İBBS-1 bölge düzeyi İBBS-2 bölge düzeyinin ebeveyn değişkeni (parent) olarak; okuryazarlık durumu değişkeni bitirilen okul değişkeninin ebeveyni olarak ilişkisel olduğu grafikten yorumlanabilmektedir. Bayes ağları uygulamasından birbiriyle ilişkili bulunan bağımlı ve bağımsız değişkenler için elde edilen koşullu olasılıklar elde edilmiş ve aşağıdaki şekildedir:

Çizelge 4.19. Koşullu olasılık tablosu-1

EBEVEYNLER		YAŞ GRUBU İÇİN KOŞULLU OLASILIKLAR					
İŞGÜCÜ DURUMU	**OKURYAZARLIK**	**15-24**	**25-34**	**35-44**	**45-54**	**55-64**	**65VEÜSTÜ**
İSTİHDAM	**EVET**	0.15	0.25	0.28	0.2	0.09	0.03
İSTİHDAM	**HAYIR**	0.04	0.08	0.16	0.25	0.28	0.2
İSTİHDAM DIŞI	**EVET**	0.28	0.15	0.13	0.14	0.15	0.15
İSTİHDAM_DIŞI	**HAYIR**	0.04	0.06	0.08	0.12	0.18	0.52

Bayes ağlarında modele alınan bağımsız değişkenler için elde edilen Şekil 4.24' te yer alan her bir ilişkisel gösterimin olasılıksal bir değeri bulunmaktadır. Şekil 4.24' te birbirine oklarla yönlendirilen bağımsız değişkenler arasındaki ilişki koşullu olasılıklar ile elde edilmiştir. Söz konusu ilişkilerin tamamının yorumlanması güç olduğundan yaş grubu ile okuryazarlık durumu ve cinsiyet ile okuryazarlık durumu arasındaki ilişkiye ait koşullu olasılıklar yorumlanmıştır.

Yaş grubuna ait koşullu olasılıklar incelendiğinde, İstihdam durumunda iken ve okuryazarlığı olan kişilerin %28'i 35-44 yaş aralığındadır. İstihdam dışında iken okuryazarlığı olmayan bireylerin %52'si 65 yaş ve üstü grupta bulunmaktadır. İstihdam dışında iken okuryazarlığı olan bireylerin %28' i 15-24 yaş aralığındadır.

Çizelge 4.20. Koşullu olasılık tablosu-2

EBEVEYNLER		**MEDENİ HAL İÇİN KOŞULLU OLASILIKLAR**		
YAŞ GRUBU	**İŞGÜCÜ DURUMU**	**DUL**	**EVLİ**	**HİÇ_EVLENMEDİ**
15-24	**İSTİHDAM**	0.00	0.13	0.87
15-24	**İSTİHDAM DIŞI**	0.00	0.13	0.87
25-34	**İSTİHDAM**	0.02	0.68	0.30
25-34	**İSTİHDAM DIŞI**	0.02	0.75	0.22
35-44	**İSTİHDAM**	0.04	0.90	0.06
35-44	**İSTİHDAM DIŞI**	0.05	0.87	0.08
45-54	**İSTİHDAM**	0.05	0.93	0.02
45-54	**İSTİHDAM DIŞI**	0.10	0.86	0.04
55-64	**İSTİHDAM**	0.06	0.93	0.01
55-64	**İSTİHDAM DIŞI**	0.16	0.82	0.02
65VEÜSTÜ	**İSTİHDAM**	0.13	0.86	0.00
65VEÜSTÜ	**İSTİHDAM_DIŞI**	0.40	0.59	0.01

Medeni hal değişkenine ait koşullu olasılıklar incelendiğinde, 15-24 yaş aralığında iken istihdam durumunda olan kişilerin %87' si hiç evlenmemiş kişilerdir. 35-44 yaş aralığında iken ve istihdam dışında kişilerin %87' si evli kişilerdir.

Çizelge 4.21. Koşullu olasılık tablosu-3

EBEVEYNLER		CİNSİYET DEĞİŞKENİ İÇİN KOŞULLU OLASILIKLAR	
İŞGÜCÜ DURUMU	**OKURYAZARLIK**	**ERKEK**	**KADIN**
İSTİHDAM	**EVET**	0.71	0.29
İSTİHDAM	**HAYIR**	0.24	0.76
İSTİHDAM DIŞI	**EVET**	0.36	0.64
İSTİHDAM DIŞI	**HAYIR**	0.15	0.85

Cinsiyet değişkenine ait koşullu olasılıklar incelendiğinde, İstihdam durumunda iken okuryazarlığı olan bireylerin %71'i erkek %29' u kadındır. İstihdam durumunda iken okuryazarlığı olmayan bireylerin %76' sı kadındır. İstihdam dışında iken okuryazarlığı olmayan bireylerin %36' sı erkek, %64' ü kadındır. Ayrıca tablo olarak eklenmeyip elde edilen koşullu olasılıklarından bazıları yorumlanmıştır; istihdamda olup yüksekokul ve üzeri mezun olanların %17' si eğitim bilimleri bölümü mezunudur. İstihdam dışı olup yüksekokul ve üzeri mezun olanların %3' ü matematik-istatistik bölümü mezunudur. İstihdamda olup okuryazarlığı olanların %56' sı ilköğretim düzeyinde okul bitirmiştir. İstihdam dışında olup okuryazarlığı olanların %65' i ilköğretim düzeyinde okul bitirmiştir. İBBS-1 bölge düzeyi TR7(Orta Anadolu) olan bireylerden istihdam dışında olanların %19' unun okuryazarlığının bulunmadığı görülmüştür.

4.5.5. Değerlendirme ve Uygulama

Modellerin uygulama aşamasından sonra işlenen yöntemlerin değerlendirilmesi ve oluşturulan modellerin Türkiye'de işgücü durumu bilinmeyen bir bireyin durumunun doğru tahmin edilebilmesi için iyileştirilmesi yapılmaktadır. Değerlendirme aşamasında hangi modelin daha başarılı olduğunu yorumlamak için çeşitli model başarı kriterleri kullanılarak model karşılaştırmaları yapılabilmektedir. Yapılan karşılaştırmalar sonucunda en başarılı bulunan model belirlenmektedir.

4.6. Yöntemlerin Karşılaştırılması

Modellerin başarıları değerlendirilirken, Hold-out yöntemi, doğruluk oranı, hata oranı, kesinlik, duyarlılık ve F-ölçütü gibi birçok kriter kullanılmaktadır. Bu kriterler yardımıyla, uygulanan 4 yönteme ait model başarıları kıyaslanmıştır.

Hold-out yönteminin uygulanması açısından veri seti (eğitim verisi-test verisi) %80-%20 ve %70-%30 olmak üzere bölümlenmiş ve veri setine ait sınıflama oranları elde edilmiştir. İki farklı bölünmeden elde edilen sınıflama sonuçları birbirine oldukça yakın olarak elde edilmiştir. Bu sebeple yalnızca bölünmenin %80-%20 oranlarında yapıldığı modelin test verileri üzerinden sınıflama oranları yorumlanmıştır.

Çizelge 4.22. C5.0 karar ağacı gerçek ve tahmini veri seti çapraz tablosu

İŞGÜCÜ DURUMU	**TAHMiN**		
GERÇEK		**İSTİHDAM**	**İSTİHDAM DIŞI**
	İSTİHDAM	68,265	36,159
	YÜZDE(%)	**%65.37**	%34.63
	İSTİHDAM DIŞI	20,060	108,725
	YÜZDE(%)	%15.5	**%84.5**

Çizelge 4.23. CHAID karar ağacı gerçek ve tahmini veri seti çapraz tablosu

İŞGÜCÜ DURUMU	**TAHMiN**		
GERÇEK		**İSTİHDAM**	**İSTİHDAM DIŞI**
	İSTİHDAM	66,778	37,646
	YÜZDE(%)	**%63.95**	%36,05
	İSTİHDAM DIŞI	19,612	109,173
	YÜZDE(%)	%15.22	**%84.78**

Çizelge 4.24. Lojistik regresyon yöntemi gerçek ve tahmini veri seti çapraz tablosu

İŞGÜCÜ DURUMU	**TAHMiN**		
GERÇEK		**İSTİHDAM**	**İSTİHDAM DIŞI**
	İSTİHDAM	69,415	35,009
	YÜZDE(%)	**%66.47**	%33.53
	İSTİHDAM_DIŞI	23,037	105,748
	YÜZDE(%)	%17.89	**%82.11**

Çizelge 4.25. Bayes ağları yöntemi gerçek ve tahmini veri seti çapraz tablosu

İŞGÜCÜ DURUMU	TAHMiN		
GERÇEK		İSTİHDAM	İSTİHDAM DIŞI
	İSTİHDAM	69,733	34,691
	YÜZDE (%)	**%66.78**	%33.22
	İSTİHDAM DIŞI	24,804	103,981
	YÜZDE (%)	%19.26	**%80.74**

Programda matris düğümü yardımıyla gerçek veri seti ile modelin tahmin ettiği veri seti arasındaki doğruluk oranları elde edilmiştir.

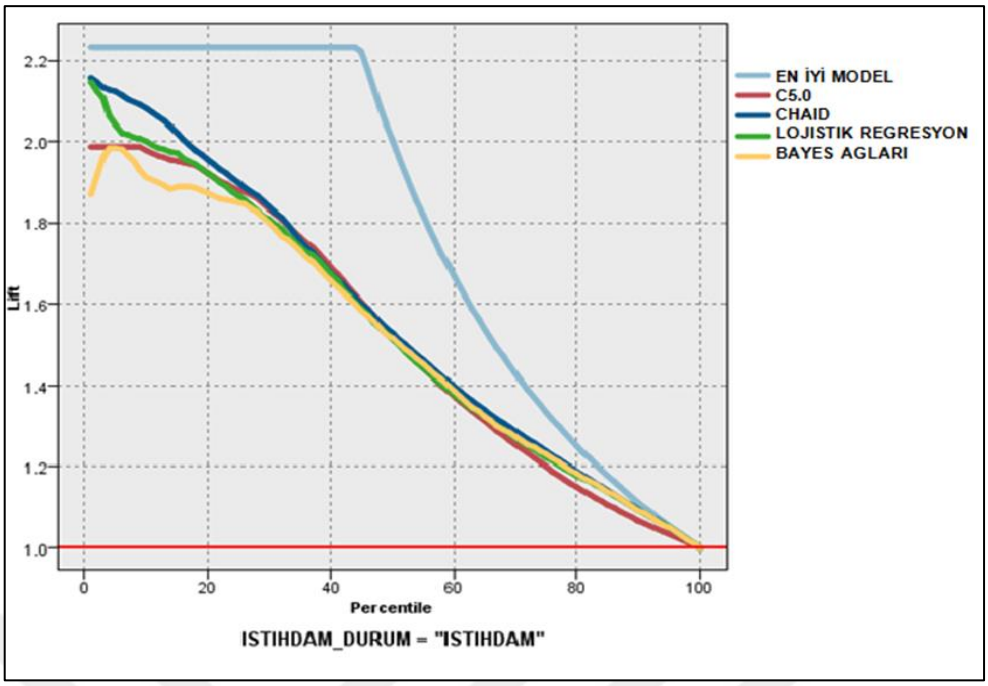
Test verisi üzerinden modelin daha önceden görmediği veri seti üzerinden yaptığı tahminlerde Çizelge 4.22' de C5.0 algoritması gerçekte istihdam durumunda olan bireyleri %65,3 oranında istihdam olarak; gerçekte istihdam dışı durumunda olan bireyleri %84.4 oranında istihdam dışı olarak tahmin ettiği görülmüştür.

Çizelge 4.23' te CHAID yöntemi gerçekte istihdam durumunda olan bireyleri %63,9 oranında istihdam olarak; gerçekte istihdam dışı durumunda olan bireyleri %84.7 oranında istihdam dışı olarak tahmin ettiği görülmüştür.

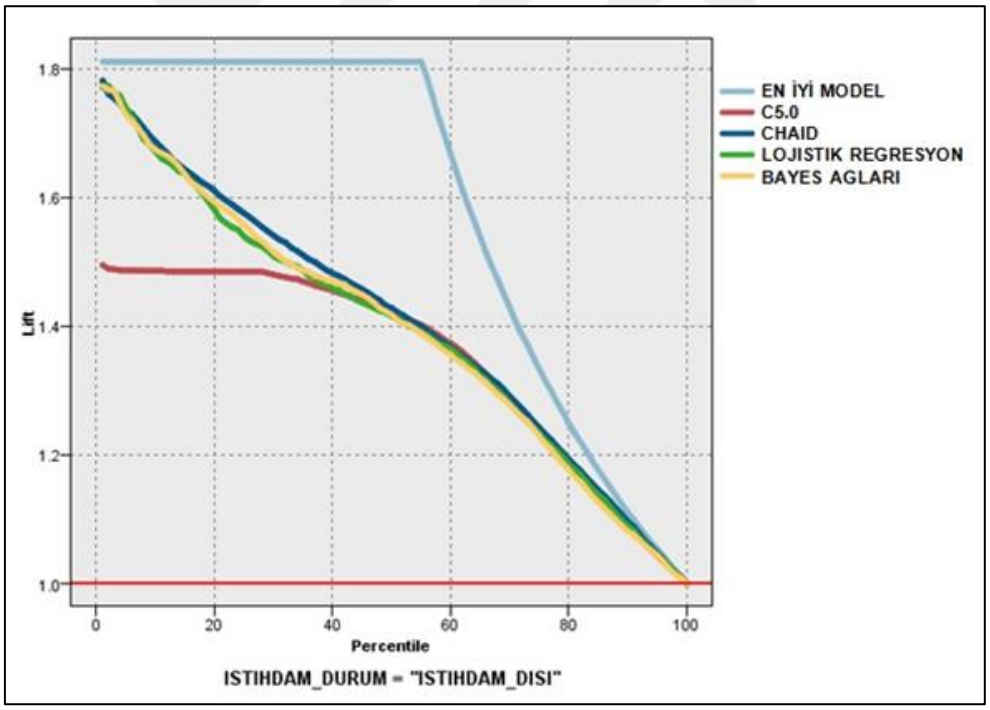
Çizelge 4.24' te Lojistik Regresyon yöntemi gerçekte istihdam durumunda olan bireyleri %66,4 oranında istihdam olarak; gerçekte istihdam dışı durumunda olan bireyleri %82,1 oranında istihdam dışı olarak tahmin ettiği görülmüştür.

Çizelge 4.25' te Bayes ağları yöntemi gerçekte istihdam durumunda olan bireyleri %66,7 oranında istihdam olarak; gerçekte istihdam dışı durumunda olan bireyleri %80,7 oranında istihdam dışı olarak tahmin ettiği görülmüştür.

Programdaki değerlendirme düğümü yardımıyla gerçek veri seti ile modelin tahmin ettiği veri seti arasındaki başarı ele alınmaktadır. Lift, Gains, ROC gibi birden fazla değerlendirme kriteri içeren değerlendirme düğümü kullanılmıştır. Model performansının ölçülmesi amacıyla Lift ölçütü test veri setine uygulanmıştır ve Lift ölçütüne göre hangi modelin optimale yakın olduğunun grafiksel olarak gösterimleri elde edilmiştir.



Şekil 4.25. Modellerin lift grafikleri-1



Şekil 4.26. Modellerin lift grafikleri-2

Şekil 4.25 ve Şekil 4.26' da tahminler için en iyi çizgi olarak tanımlanan en üstteki model mavi çizgi ile, kırmızı olan çizgi temel düzey, arada oluşanlar ise modellerin yaptığı tahmin doğrusunu vermektedir. Hem istihdam hem istihdam dışı kategorileri için yapılan

tahminler Lift değerine göre grafiklendirilmiştir. Lift tablosunda tahminlerin doğruluğuna göre yüzde olarak gösterimi yorumlanmaktadır. Test verisi üzerinden modelin daha önceden görmediği veri seti üzerinden yaptığı tahminlerde model sonuçları birbirine yakın gibi elde edilmesine rağmen, en iyi tahmin çizgisine en yakın olan model her iki kategori için de CHAID karar ağacı yöntemi olarak belirlenmiştir. Grafiklerde lift değeri istihdam kategorisi tahminleri için 2' den başlayarak; istihdam dışı kategorisi için 1,8' den başlayarak azalmaktadır. Model başarısı için lift değerinin 1' den fazla olması beklendiğinden, bu değerler model başarısı için geçerli bir göstergedir.

Modellere ait çeşitli karşılaştırma kriterlerinin test verileri üzerinden hesaplanması yapılmıştır. Modellerin doğruluk, kesinlik, duyarlılık gibi kriterleri hesaplanmıştır, ayrıca duyarlılık kesinlik kriterleri ile bu hesaplamalar yardımıyla ortaya çıkan F-ölçütü değeri elde edilmiştir. Kriterlerin birlikte yorumlanması ve değerlendirilmesinin doğru olduğu düşünüldüğünden veri seti için C5.0 algoritmasının daha doğru ve başarılı tahminlerde bulunduğu düşünülmektedir.

Çizelge 4.26. Modellerin karşılaştırma kriterleri sonuçları

KRİTERLER	C5.0	CHAID	LOJİSTİK REGRESYON	BAYES AĞLARI
DOĞRULUK	75,9%	75,4%	75,1%	74,5%
HATA	24,1%	24,6%	24,9%	25,5%
KESİNLİK	65,4%	63,9%	66,5%	66,8%
DUYARLILIK	77,3%	77,3%	75,1%	73,8%
F-ÖLÇÜTÜ	70,8%	69.9%	70.5%	70.1%

Veri setine C5.0 karar ağacı, CHAID karar ağacı, Lojistik regresyon ve Bayes ağları yöntemleri uygulanmış ve bu yöntemler sonucunda işgücü durumu için test verisinde elde edilen tahmin verilerine ilişkin doğruluk oranı, hata oranı, kesinlik, duyarlılık ve F-ölçütü değerleri hesaplanmıştır. Sınıflamada en yüksek doğruluk oranı C5.0 karar ağacı yöntemi olarak görülmüştür. Buradan çıkarımla hatalı sınıflama oranı da en düşük yöntem C5.0 karar ağacı olmuştur. Kesinlik ve Duyarlılık ölçütleri ile hesaplanan F-ölçütü ise, kesinlik ve duyarlılık kriterlerinin tek başına yorumlanmasından daha etkili olduğundan, F-ölçütü olarak %70,8 oranı yine C5.0 karar ağacı yönteminin çalışma konusu veri seti için en uygun tahminler yapan model olduğuna işaret etmektedir.

5. SONUÇ

Günümüz teknolojik yaşantısında "bilgi"; mihenk taşı haline gelmiştir. Özellikle bilgisayarların her alanda hayatımıza girmesiyle birlikte; bilgi daha depolanabilir, paylaşılabilir ve üzerinde analizler yapılabilir hale gelmiştir. Bilgisayar teknolojisi ve bilgideki söz konusu gelişmelere, artan ihtiyaçların karşılığı olarak her geçen gün artış gösteren kaynaklara bağlı olarak daha güçlü, daha doğru ve daha hızlı istatistiksel veri analiz süreçleri de geliştirilmiştir. Dolayısıyla bu süreçte bilgi teknolojisine bağlı olarak büyüyen veri hacimlerinin tahlil, tetkik ve düzenlemelerinde kolaylık sağlayan veri madenciliği yöntemleri ortaya çıkmıştır. Öyle ki; insan toplumunun attığı adımdan dahi veri türeten günümüz teknolojisine eş değer olarak veri madenciliği yöntemlerine olan gereksinim de gün geçtikçe artmaktadır. Bu sebeple hemen hemen her disiplin tarafından çeşitli amaçlar için kullanılan veri madenciliği alanındaki gelişmeler, akademisyen ve araştırmacı bilim çevreleri tarafından, uluslararası bilimsel dergiler ve kongreler aracılığıyla takip edilmektedir.

Bu çalışma kapsamında; TÜİK tarafından uygulanan ve yürütülen Hanehalkı İşgücü Araştırmaları (HİA)' nın 2014, 2015 ve 2016 yıllarına ait Hanehalkı İşgücü Araştırmaları (HİA) verileri kullanılmıştır. Söz konusu 3 yıl için 1,163,566 bireye ilişkin demografik, coğrafi özellikler ve işgücü verileri TÜİK tarafından sağlanmış olup bireylerin işgücü verileri çalışma kapsamına alınmıştır. HİA verilerinin kullanılmasındaki amaç, Türkiye' de işgücü durumunun detaylı olarak ele alınması, veri madenciliği yöntemlerinden veri setine uygun olan sınıflama yöntemlerinden kullanılarak, modelin tanımadığı işgücü durumu bilinmeyen bir bireyin model sonucunda işgücü durumunun doğru tahmin edilmesini sağlamaktır. Modellerin sonucunda işgücü durumu değişkeni için çeşitli model başarı kriterlerine göre en uygun model belirlenmiştir.

Çalışma, veri madenciliği süreçlerinden CRISP-DM döngüsü adımlarına göre yürütülmüştür. HİA çalışması veri setinde sınıflama yapılması amacıyla, test verisi ve öğrenme verisi olarak bölümlenen veri setine, veri madenciliği sınıflama yöntemlerinden C5.0 karar ağaçları, CHAID karar ağaçları, Lojistik Regresyon ve Bayes Ağları yöntemleri uygulanmıştır. Elde edilen modeller sonucu elde edilen tablolar ve ilişkiler yorumlanmıştır. Veri madenciliği yöntemlerinde karşılaştırma yapılması hususunda çeşitli model başarı kriterleri kullanılmaktadır. Ele alınan kriterler ışığında modellerden elde edilen test verisinde başarı oranları elde edilmiş, karşılaştırmalar yapılmıştır. Çözümlenen veri seti için en uygun yöntemin C5.0 karar ağacı yöntemi olduğu belirlenmiştir. İşgücü verileri ile

C5.0 karar ağacı yöntemi sonucu elde edilen sonuçlara göre; işgücü durumu istihdam kategorisi için, cinsiyet en önemli değişken olarak belirlendiğinden ağaç cinsiyet değişkenine göre bölünmektedir. C5.0 karar ağacı yöntemiyle, işgücü durumunu en çok etkileyen değişkenin cinsiyet ve bunu takiben yaş grubu değişkeninin olduğu belirlenmiştir. Eğitime devam değişkeni ve İBBS bölge düzeyleri değişkenleri veri setinde en düşük etkiye sahip değişkenler olarak belirlenmiştir. Söz konusu kriterlerin değerlendirilmesiyle model başarıları karşılaştırılmasında C5.0 karar ağacı yöntemini sırasıyla CHAID karar ağacı ve Lojistik regresyon yöntemi takip etmektedir. Veri seti için Bayes ağları yönteminin başarı kriterleri açısından en son tercih edilecek model olduğu sonucuna varılmıştır.

Uygulanan veri madenciliği yöntemlerinden veri seti için en uygun modelin belirlenmesinde kullanılan doğruluk oranı, hata oranı, kesinlik, duyarlılık, F-ölçütü gibi kriterlerin yardımıyla belirlenen model ile işgücü durumu bilinmeyen bir bireyin işgücü durumunun yaklaşık %76' lık doğruluk payıyla tahmin edilmesi mümkün olmaktadır.

Türkiye'deki işgücü profilinin belirlenmesi ve ileriye yönelik bireylerin işgücü durumlarına ilişkin tahminlerin yapılmasında çalışmada kullanılan yöntemlerin yanında Yapay Sinir Ağları ve Karar Destek Makineleri gibi diğer sınıflama yöntemleri kullanılarak model başarısı karşılaştırma kriterlerinden modelin çalışma hızı, ölçeklenebilirlik gibi çeşitli kriterler de değerlendirilerek yorumlanabileceği düşünülmektedir. Kullanılan yöntemler yardımıyla, Türkiye' de işgücü durumlarına ait tahminler daha doğru yapılabilmekle birlikte işgücünde etkili olabilecek gruplara yönelik uygulama ve teşvikler yapılarak işgücüne dâhil edilme olanağı sağlanabilecektir. Özellikle işgücüne katılma potansiyeli oldukça yüksek olan genç nüfusun çeşitli uygulamalarla, istihdam dışındaki kadınların ise çeşitli teşvikler uygulanarak istihdama katılmasının sağlanabileceği düşünülmektedir. Ayrıca ilgili kurum ve kuruluşlar nezdinde oluşturulacak projeksiyonlarla birlikte; uygun istihdam politikasından, toplumda uygulanacak eğitimlere, işverenlere sağlanacak teşviklere ve şirketlerin uygulayacağı ücret politikalarına kadar geniş yelpazede birçok destek sağlanabilecektir. Bunun sonucu olarak da ülkenin istihdam hedefleri gerçekleştirilerek toplumsal refaha kavuşturulabilecektir.

KAYNAKÇA

[1] Kara Ö.S., Lojistik Regresyon Analizi Ve Kadın İşgücü Üzerine Bir Uygulama, Yüksek Lisans Tezi, Uludağ Üniversitesi Sosyal Bilimler Enstitüsü İstatistik Bilim Dalı, Bursa, **2015**

[2] Altunkaya H.İ., Ülkelerin Uzun Dönem Kredi Notlarının Derecelendirilmesinde Önemli Değişkenlerin Veri Madenciliği Teknikleri Kullanılarak Belirlenmesi, Yüksek Lisans Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı, Ankara, **2013**

[3] Kocabaş F.M., Veri Madenciliği Süreci ve Gerçek Bir Veri Seti Üzerinde Uygulanması, Yüksek Lisans Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı, Ankara, **2010**

[4] Yılmaz E., İstatiksel Analiz Yöntemi Olarak Veri Madenciliğinde CHAID Algoritması Ve Türkiye'de İşgücü Piyasasının Durumunun ve Bunun Nedenlerinin Belirlenmesine İlişkin Bir Uygulama, Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi Sosyal Bilimler Enstitüsü İşletme Ana Bilim Dalı İşletme Yönetimi Yüksek Lisans Programı, İstanbul, **2012**

[5] Yakut E., Veri Madenciliği Tekniklerinden C5.0 Algoritması ve Destek Vektör Makineleri ile Yapay Sinir Ağlarının Sınıflandırma Başarılarının Karşılaştırılması: İmalat Sektöründe Bir Uygulama, Doktora Tezi, Atatürk Üniversitesi Sosyal Bilimler Enstitüsü İşletme Anabilim Dalı, Erzurum, **2012**

[6] Kuzey C., Veri Madenciliğinde Destek Vektör Makinaları ve Karar Ağaçları Yöntemlerini Kullanarak Bilgi Çalışanlarının Kurum Performansı Üzerine Etkisinin Ölçülmesi ve Bir Uygulama, Doktora Tezi, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü İşletme Anabilim Dalı Sayısal Yöntemler Bilim Dalı, İstanbul, **2012**

[7] Çakır Ö., Veri Madenciliğinde Sınıflandırma Yöntemlerinin Karşılaştırılması "Bankacılık Müşteri Veri Tabanı Üzerinde Bir Uygulama", Doktora Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü İşletme Anabilim Dalı Sayısal Yöntemler Bilim Dalı, İstanbul, **2008**

[8] Kıyak E., Crisp-Dm Yöntembilimi Kullanılarak Deniz Kuvvetleri Verisi Üzerinde Veri Madenciliği Sınıflandırma Tekniklerinin Karşılaştırılması, Yüksek Lisans Tezi, Kocaeli Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı, Kocaeli, **2006**

[9] Özdemir A., Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği-Sağlık Sektöründe Uygulama, Doktora Tezi, Atatürk Üniversitesi Sosyal Bilimler Enstitüsü İşletme Anabilim Dalı, Erzurum, **2004**

[10] Doğan K., Arslantekin S., Büyük Veri: Önemi, Yapısı ve Günümüzdeki Durum, DTCF Dergisi, 56.1, 15-36, **2016**

[11] Dura C., Atik H. ,Bilgi Toplumu, Bilgi Ekonomisi ve Türkiye, İstanbul, Literatür Yayıncılık, 1.Baskı, **2002**

[12] Han J., Kamber M., *Data Mining: Concepts and Techniques*, Second Edition, Morgan Kaufmann Publishers, 2nd Ed., San Francisco, USA, **2006**

[13] Can M.B. vd., Veri kümelerinden bilgi keşfi: veri madenciliği, http://docplayer.biz.tr/14269227-Mehmet-berkay-can-eren-camur-mine-koru-omer-ozkan-zeynep-rzayeva.html (Şubat **2017**)

[14] Özkan Y., *Veri Madenciliği Yöntemleri*, Papatya Yayıncılık Eğitim, **2008**)

[15] Giudici P., *Applied Data Mining, Statistical Methods for Business and Industry*, Italy, **2003**

[16] Karaca İ., Büyük Veri Analizlerinin Kurumsal Faaliyetlerde Kullanım Alanları, Lisans Tezi, Ankara Üniversitesi Dil ve Tarih-Coğrafya Fakültesi Bilgi ve Belge Yönetimi Bölümü Ankara, **2015**

[17] Savaş S. vd., Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri, Tarama makalesi, İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 11-2, **2012**

[18] Olson D.L., Delen D., *Advanced Data Mining Techniques*, Springer-Verlag Berlin Heidelberg, **2008**

[19] Öğüt S., Veri Madenciliği Kavramı ve Gelişim Süreci. Yeditepe Üniversitesi, İstanbul, 2002, http://www.sertacogut.com/blog/wp-content/uploads/2009/03/sertac_ogut_-_veri_madenciligi_kavrami_ve_gelisim_sureci.pdf (Şubat, **2017**)

[20] http://ugurozmen.com/crm/en-iyi-teklif-sss-3 (Şubat, **2017**)

[21] Atılgan E., Karayollarında Meydana Gelen Trafik Kazalarının Karar Ağaçları ve Birliktelik Analizi ile İncelenmesi, Yüksek Lisans Tezi, Hacettepe Üniversitesi, İstatistik Bölümü, Ankara, **2011**

[22] Albayrak A.S., Yılmaz Koltan Ş., Veri Madenciliği: Karar Ağacı Algoritması ve İMKB Verileri Üzerine Bir Uygulama, Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 14-1, 31-52, **2009**

[23] Dolgun M.Ö., Büyük Alışveriş Merkezleri İçin Veri Madenciliği Uygulamaları, Yüksek Lisans Tezi, Hacettepe Üniversitesi İstatistik Anabilim Dalı, Ankara, **2006**

[24] Gullo F., From Patterns in Data to Knowledge Discovery: What Data Mining Can Do, 3rd International Conference Frontiers in Diagnostic Technologies, ICFDT3 2013, Physics Procedia 62, 18-22, **2015**

[25] Şentürk A., *Veri Madenciliği Kavram ve Teknikler*, Ekin Yayınevi, **2006**

[26] Emel G.G., Taşkın Ç., Veri Madenciliğinde Karar Ağaçları ve Bir Satış Analizi Uygulaması, Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi, 6-2, 221-239, **2005**

[27] Ayık Y.Z. vd., Lise Türü ve Lise Mezuniyet Başarısının, Kazanılan Fakülte ile İlişkisinin Veri Madenciliği Tekniği ile Analizi"., Atatürk Üniversitesi , Sy.445,

http://docplayer.biz.tr/3976459-Lise-turu-ve-lise-mezuniyet-basarisinin-kazanilan-fakulte-ile-iliskisinin-veri-madenciligi-teknigi-ile-analizi.html (Mayıs, **2018**)

[28] Witten I.A., Frank E., *Data Mining Practical Machine LearningTools and Techniques*, The Morgan Kaufmann Series in Data Management Systems, 2nd Ed., **2005**

[29] Yohannes Y., Hoddinott J., Classıfıcatıon And Regressıon Trees: An Introductıon Technical Guide, International Food Policy Research Institute 2033 K Street, N.W. Washington, D.C. 2006 U.S.A., Mart 1999, S.2, https://www.researchgate.net/publication/242370834_Classification_and_Regression_Trees_An_Introduction (Nisan, **2018**)

[30] Michie D., *Machine Learning, Neural and Statistical Classification*, Cambridge, U.K., **1994**

[31] Berry M.J.A. , *Linoff G.S., Data Mining Techniques for Marketing, Sales and Customer Relationship Management*, Wiley Publishing, 2nd Ed., Indiana, **2004**

[32] Doğan N., Özdamar K., CHAID Analizi ve Aile Planlaması ile ilgili Bir Uygulama, T Klin Tıp Bilimleri, 23, 392-397, **2003**

[33] Kass GV., An Exploratory Technique For İnvestigating Large Quantities Of Categorical Data, Applied Statistics, 29-2, 119-127 , **1980**

[34] Diaz-Perez F.M., Bethencourt-Cejas M., Chaid Algorithm As An Appropriate Analytical Method For Tourism Market Segmentation, Journal of Destination Marketing & Management, 5-3, 275–282, **2016**

[35] Ritschard G., CHAID and Earlier Supervised Tree Methods , Dept of Econometrics, University of Geneva, Switzerland, 2010-02, 1-28, **2010**

[36] Karakoyun M., Hacıbeyoğlu M., Biyomedikal Veri Kümeleri ile Makine Öğrenmesi Sınıflandırma Algoritmalarının İstatistiksel Olarak Karşılaştırılması, Dokuz Eylül Üniversitesi Mühendislik Fakültesi Mühendislik Bilimleri Dergisi, 16-48, 30-41, **2014**

[37] Siyah B., Spor Metinleri Sınıflama Projesi, MKÜ Bilgisayar Mühendisliği Bölümü, S.1, http://www.bulentsiyah.com/naive-bayes-ile-spor-metinleri-siniflandirma/, (Nisan, **2017**)

[38] Larose D.T., *Data Mining Methods and Models*, A John Wiley & Sons, Inc Publication, United States of America, **2006**

[39] Murat N., Model Seçiminde Bayesci Yaklaşımların Kullanımı, Yüksek Lisans Tezi, Ondokuz Mayıs Üniversitesi Fen Bilimleri Enstitüsü İstatistik Ana Bilim Dalı, Samsun, **2007**

[40] Aktaş C., Erkuş O., Lojistik Regresyon Analizi ile Eskişehir'in Sis Kestiriminin İncelenmesi, İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 8-16, 47-59, **2009**

[41] Önder H., Cebeci Z., Lojistik Regresyonlarda Değişken Seçimi, Çukurova Üniversitesi Ziraat Fakültesi Dergisi, 17 (2), 105-114, **2002**

[42] Tatlıdil, H., *Uygulamalı Çok Değişkenli İstatistiksel Analiz,* Ziraat Matbaacılık, Ankara, **2002**

[43] Kıran Z.B., Lojistik Regresyon ve Cart Analizi Teknikleriyle Sosyal Güvenlik Kurumu İlaç Provizyon Sistemi Verileri Üzerinde Bir Uygulama, Yüksek Lisans Tezi, İstatistik Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara, **2010**

[44] Başarır G., Çok Değişkenli Verilerde Ayrımsama Sorunu ve Lojistik Regresyon Analizi, Doktora Tezi, Hacettepe Üniversitesi, İstatistik Bölümü, Ankara, **1990**

[45] Hosmer D.W., Lemeshow S., *Applied Logistics Regression, Wiley Series in Probability and Statistics*, John Wiley & Sons, Inc. , 2nd Ed., **2000**

[46] Ürük E., İstatistiksel Uygulamalarda Lojistik Regresyon Analizi, Yüksek Lisans Tezi, Marmara Üniversitesi Fen Bilimleri Enstitüsü, Matematik Anabilim Dalı Uygulamalı Matematik Programı, İstanbul, **2007**

[47] Aladağ Ç.H. vd, Improving Weighted Information Criterion By Using Optimization, Journal of Computational and Applied Mathematics, 233-10, 2683-2687, **2010**

[48] Bircan H., Lojistik Regresyon Analizi: Tıp Verileri Üzerine Bir Uygulama, Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 2, 185-208, **2004**

[49] Aladağ Ç.H. vd., Forecasting Nonlinear Time Series With A Hybrid Methodology, Applied Mathematics Letters, 22-9, 1467-1470, **2009**

[50] Coşkun C., Baykal A., Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması, Akademik Bilişim'11 - XIII. Akademik Bilişim Konferansı Bildirileri, Şubat 2011 İnönü Üniversitesi, Malatya, 51-58, **2011**

[51] Dolgun M.Ö., Veri Madenciliği Sınıflama Yöntemlerinin Başarılarının; Bağımlı Değişken Prevelansı, Örneklem Büyüklüğü ve Bağımsız Değişkenler Arası İlişki Yapısına Göre Karşılaştırılması, Doktora Tezi, Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü, Ankara, **2014**

[52] Akçayol M.A., Web Madenciliği (Web Mining) Ders Notları, Gazi Üniversitesi Bilgisayar Mühendisliği Bölümü,

http://w3.gazi.edu.tr/~akcayol/files/WM_L5ClassifierEvaulation.pdf (Nisan, **2018**)

[53] Powers D.M.W., Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation, Journal of Machine Learning Technologies, 2-1, 51-58, 2011

[54] Beliner E., Müşteri İlişkileri Yönetiminde İstatistiğin Yeri ve Bir Uygulama, Yüksek Lisans Tezi, Hacettepe Üniversitesi İstatistik Anabilim Dalı, Ankara, **2015**

[55] TÜİK web sayfası,

http://www.TÜİK.gov.tr/MicroVeri/HİA_2015/turkce/metaveri/tanim/index.html (Ocak, **2018**)

[56] TÜİK web sayfası,

http://www.TÜİK.gov.tr/MicroVeri/HİA_2015/turkce/metaveri/tarihce/index.html (Ocak, **2018**)

[57] Taş B., AB Uyum Sürecinde Türkiye İçin Yeni Bir Bölge Kavramı: İstatistiki Bölge Birimleri Sınıflandırması (İbbs), Afyon Kocatepe Üniversitesi Fen-Edebiyat Fakültesi Coğrafya Bölümü, 03200 Afyonkarahisar, Sosyal Bilimler Dergisi, 8-2, 185-198, **2006**

[58] IBM web sayfası,

https://www.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.modeler.help/partition_settingstab.htm (Ocak, **2018**)

ÖZGEÇMİŞ

Kimlik Bilgileri

Adı Soyadı : Merve BARAN KILIÇALAN
Doğum Yeri : Konak
Medeni Hali : Evli
E-Posta : merveebaran@hotmail.com

Eğitim

Lisans : Hacettepe Üniversitesi İstatistik (2007-2011)
Lisans : Anadolu Üniversitesi, İşletme (2008-2014)
Yüksek Lisans : Hacettepe Üniversitesi İstatistik (2011-2018)

Yabancı Dil ve Düzeyi

İngilizce - İyi

İş Deneyimi

(2012-2016) Sosyal Güvenlik Kurumu
Sosyal Güvenlik Uzman Yardımcısı
(2016-Halen) Sosyal Güvenlik Kurumu
Sosyal Güvenlik Uzmanı

Deneyim Alanları

Veri ambarı, veri analizi ve raporlama

Tezden Üretilmiş Projeler ve Bütçesi

-

Tezden Üretilmiş Yayınlar

-

Tezden Üretilmiş Tebliğ ve/veya Poster Sunumu ile Katıldığı Toplantılar

-

HACETTEPE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
YÜKSEK LİSANS/DOKTORA TEZ ÇALIŞMASI ORJİNALLİK RAPORU

HACETTEPE ÜNİVERSİTESİ
FEN BİLİMLER ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI BAŞKANLIĞI'NA

Tarih: 08/06/2018

Tez Başlığı: HANEHALKI İŞGÜCÜ ARAŞTIRMA VERİLERİ İLE VERİ MADENCİLİĞİ YÖNTEMLERİNİN UYGULANMASI VE MODELLERİN KARŞILAŞTIRILMASI

Yukarıda başlığı gösterilen tez çalışmamın a) Kapak sayfası, b) Giriş, c) Ana bölümler d) Sonuç kısımlarından oluşan toplam 77 sayfalık kısmına ilişkin, 29/05/2018 tarihinde tez danışmanım tarafından *Turnitin* adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı %9' dur.

Uygulanan filtrelemeler:

1- Kaynakça hariç
2- Alıntılar hariç
3- 5 kelimeden daha az örtüşme içeren metin kısımları hariç

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü Tez Çalışması Orjinallik Raporu Alınması ve Kullanılması Uygulama Esasları'nı inceledim ve bu Uygulama Esasları'nda belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini saygılarımla arz ederim.

Tarih ve İmza

Adı Soyadı:	Merve BARAN KILIÇALAN 08.06.2018
Öğrenci No:	N10324021
Anabilim Dalı:	İSTATİSTİK
Programı:	İSTATİSTİK
Statüsü:	☒ Y.Lisans ☐ Doktora ☐ Bütünleşik Dr.

DANIŞMAN ONAYI

UYGUNDUR.

Doç. Dr. G. Hakan Hacıoğlu
(Unvan, Ad Soyad, İmza)