T.C.
MALTEPE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

TÜRKİYE'DE MOBİL VERİ KULLANIMININ VERİ MADENCİLİĞİNDE KULLANILAN ALGORİTMALAR İLE ANALİZİ

YÜKSEK LİSANS TEZİ

Muhammet Ali ALTINIŞIK

Tez Danışmanı
Yrd. Doç. Dr. Erdal GÜVENOĞLU

İSTANBUL – 2017

T.C.
MALTEPE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

TÜRKİYE'DE MOBİL VERİ KULLANIMININ VERİ MADENCİLİĞİNDE KULLANILAN ALGORİTMALAR İLE ANALİZİ

YÜKSEK LİSANS TEZİ

Muhammet Ali ALTINIŞIK

Tez Danışmanı
Yrd. Doç. Dr. Erdal GÜVENOĞLU

İSTANBUL – 2017

T.C. Maltepe Üniversitesi
Fen Bilimleri Enstitüsü Müdürlüğüne,

22.12.2017 tarihinde tezinin savunmasını yapan Muhammet Ali ALTINIŞIK' a ait "Türkiye'de mobil veri kullanımının veri madenciliğinde kullanılan algoritmalar ile analizi" başlıklı çalışma, Jürimiz Tarafından Fen Bilimleri Bilgisayar Mühendisliği Anabilim Dalı, Bilgisayar Mühendisliği Tezli Yüksek Lisans Programında Yüksek Lisans Tezi Olarak **Oy Birliği/~~Oy Çokluğuyla~~** Kabul Edilmiştir.

.....

Yrd.Doç.Dr. Erdal GÜVENOĞLU
(Başkan)
(Danışman)

.....

Prof. Dr. A. Mesut RAZBONYALI
(Üye)

.....

Yrd. Doç. Dr. Can RAZBONYALI
(Üye)

YEMİN METNİ

26/12/2017

Yüksek Lisans tezi olarak sunduğum " Türkiye'de mobil veri kullanımının veri madenciliğinde kullanılan algoritmalar ile analizi " adlı çalışmanın, proje safhasından sonuçlanmasına kadar olan bütün süreçlerinde bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurulmaksızın tarafımca yazıldığını ve yararlandığım bütün eserlerin "Kaynakça"da gösterilenlerden oluştuğunu, "Kaynakça"da yer alan bu eserlerden metin içinde atıf yaparak yararlanmış olduğumu belirtir ve onurumla doğrularım.

151402202
Muhammet Ali Altınışık

TÜRKİYE'DE MOBİL VERİ KULLANIMININ VERİ MADENCİLİĞİNDE KULLANILAN ALGORİTMALAR İLE ANALİZİ

ÖZET

Yüksek Lisans Tezi, Türkiye'de Mobil Veri Kullanımının Veri Madenciliğinde Kullanılan Algoritmalar ile Analizi, T.C. Maltepe Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı.

Günümüzde hızla gelişen teknolojilere paralel olarak artış gösteren veri miktarları, verilerin karmaşık bir şekilde birikmesine neden olmaktadır. Artan verilerin en büyük problemleri; saklanması, işlenmesi ve anlamlandırılmasıdır. Verilerin işlenme ve anlamlandırılma noktasındaki problemlerine çözüm olarak veri madenciliği ortaya çıkmıştır. Veri madenciliği, belirli algoritmalar kullanılarak verilerin daha önceden bilinmeyen anlamlı bilgilerinin tahmin edilmesidir. Veri madenciliğinde verileri ortak özelliklerine göre gruplamak için kümeleme ve sınıflandırma algoritmaları yardımıyla daha farklı yöntemler bulunabilir.

Bu çalışmada belirli bölgeler ve yaş grupları arasında, mobil veri kullanımına ait veri seti elde edilerek bu verilerin veri madenciliğinde sınıflandırma, regresyon ve kümeleme algoritmalarıyla işlenmesi ve bununla beraber Türkiye'de bireylerin mobil veri kullanımının belirli özelliklerine göre (bulunduğu bölge, yaş grubu ve cinsiyet) ortaya çıkarılması amaçlanmıştır.

Kasım 2017 yılında yazılmış olan bu tez 69 sayfadan oluşmaktadır.

Anahtar Kelimeler: Veri Madenciliği, Karar Ağaçları, Çoklu Regresyon Modeli, Kümeleme Yöntemleri, Mobil Veri

ANALYSIS WITH ALGORITHMS USED IN DATA MINING OF MOBILE DATA USAGE IN TURKEY

ABSTRACT

Thesis, Analysis with Algorithms Used in Data Mining of Mobile Data Usage in Turkey, T.C. Maltepe University, Graduate School of Natural and Applied Sciences, Department of Computer Engineering.

Nowadays, increasing volume of data parallel to rapidly developing technologies, data causes complex accumulation of data. The storage, processing and giving meaning of increasing datas are a major problem. Data mining is used for all datas that are stored for giving meaning. Data mining is the prediction of information in a meaning previously unknown using specific algorithm. In data mining, different methods can be found with the help of clustering and classification algorithms to group data according to common properties. In this study, data set of mobile data usage between certain regions and age groups is obtained and processed by classification, regression and clustering algorithms in data mining.

In this study, dataset of mobile data usage between certain regions and age groups is processed by classification, regression and clustering algorithms in data mining. And then, the use of mobile data by individuals in Turkey is determined according to specific demographic information (region, age group and gender).

This thesis which was written in november 2017 consist of 69 pages.

Key Words: Data Mining, Decision Tree, Multiple Regression, Clustering Methods, Mobile Data

URKUND

Urkund Analysis Result

Analysed Document: TEZ_Ali_Altınışık.docx (D32734002)
Submitted: 11/22/2017 6:40:00 AM
Submitted By: erdalguvenoglu@maltepe.edu.tr
Significance: 1 %

Sources included in the report:

Burcu ALAN -YL.pdf (D20976524)

Instances where selected sources appear:

1

Yrd. Doç. Dr. Erdal GÜVENOĞLU
Bilgisayar Mühendisliği Bölümü

ÖNSÖZ

Bu tez çalışmasında Türkiye'de bireylerin mobil veri kullanımının belirli özelliklerine göre (bulunduğu bölge, yaş grubu ve cinsiyet) veri madenciliğinde sınıflandırma, regresyon ve kümeleme algoritmalarıyla işlenerek ortaya çıkarılması amaçlanmıştır.

Bu araştırmadaki önerileri, yardımları, desteği ve ilgisi için değerli hocam Yrd. Doç. Dr. Erdal Güvenoğlu'na, yüksek lisans eğitimim boyunca benden bilgisini ve desteğini esirgemeyen tüm hocalarıma sonsuz teşekkürlerimi sunuyorum. Bilgisi ve deneyimiyle desteğini esirgemeyen değerli kardeşim Arş. Gör. Muhammet Emin Altınışık'a teşekkür ederim. Her zaman yanımda olan çok değerli aileme, gösterdikleri anlayış ve sabır için şükranlarımı sunuyorum.

Muhammet Ali ALTINIŞIK

İÇİNDEKİLER

KISALTMALAR LİSTESİ

TÜİK	:	Türkiye İstatistik Kurumu
IOT	:	Nesnelerin interneti
M2M	:	Makineler arası iletişim
ITU	:	Uluslararası Telekomünikasyon Birliği

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

1. GİRİŞ

Günümüzde bazı kurum ve şirketler, ürettikleri verileri depolama alanlarında saklamaktadırlar. Bu verilerin saklanması, toplanması ve işlenmesi de ek problemler olarak karşımıza çıkmaktadır. Depolanan veriler her geçen gün daha da büyümekte ve verilere bakılarak anlamlı ifade çıkarılması güçleşmektedir. Karmaşık halde saklanan veriler, belirli bir amaç için işlenmesiyle anlamlı hale gelebilmektedir. Toplanan verilerin anlamlandırılması noktasında ise veri madenciliği çözüm olarak ortaya çıkmıştır. Veri madenciliği, belirli algoritmalar kullanılarak verilerin daha önceden bilinmeyen anlamlı bilgilerinin tahmin edilmesi ve ortaya çıkarılmasıdır.

Gelişen teknoloji ile insanlar veri kullanımı noktasında hatırı sayılır bir çoğunlukta mobil hatları kullanmaktadırlar. İnternet ve bazı iletişim programlarının yoğun kullanıldığı çağımızda artık, mobil iletişim araçları vazgeçilmez bir araç haline gelmiştir. Mobil veri kullanımı bu sayede daha da artmaktadır. Artan mobil veri kullanımı, iletişim hizmeti sunan firmaların sürekli altyapısını geliştirmeye zorlamaktadır. Yapılacak iyileştirmeler için veri kullanım yoğunluğunu önceden tespit edip zamanında yapılan geliştirmeler, rekabet gücünün ön planda olduğu günümüzde firmayı her zaman bir adım öne taşımaktadır. Ayrıca altyapısı için hazırlıkları yapılan 5G teknolojisinde ise internetin yanı sıra tüm iletişim hizmetinin de tamamen veriler üzerinden sağlanacağını düşünülürse, mobil verinin ne kadar önemli derecede kullanılacağı ve veri miktarının artacağı da bir gerçektir. Bu hizmeti sunan firmalar arasındaki rekabet her geçen gün artmaktadır. 5G teknolojisinin ağır maliyetinden dolayı bireylerin hangi bölgelerde, hangi yaş aralığında, ne kadar veri kullandığı bilgilerinin ortaya çıkmasıyla, yatırım yapılacak bölgeler ve şehirler sıralamasına bu çalışmanın bir ön ışık olarak da kullanılması hedeflenmektedir.

Bununla birlikte Türkiye'de bireylerin belirli özellik bilgilerine göre (bulunduğu bölge, yaş grubu ve cinsiyet) mobil veri kullanımının veri madenciliğinde sınıflandırma, regresyon ve kümeleme yöntemleri kullanılarak ortaya çıkarılması amaçlanmıştır. Bu araştırmayı gerçekleştirmek için herhangi bir X firmasından

bölge, yaş grubu, cinsiyet ve veri kullanım bilgileri tahmin edici değişkenler olarak kullanılarak, Türkiye'de belirli bölge ve yaş gruplarına ait mobil veri kullanımının ortaya çıkarılması öngörülmeye çalışılmıştır. Bu çalışmada kullanılan veriler içerisinde kişisel hiçbir veri yoktur. Sadece belirli bölgelerde, belirli yaş aralığı ve cinsiyete ait veri kullanım miktarı bilgileri ile çalışılmıştır.

Bu çalışmada verinin kapsamı, veri kümesi ve çalışmada kullanılabilecek uygun programlarla teknikleri belirlemek için aşağıdaki süreç izlenilmiştir;

- Verinin kapsamı 2 şekilde belirlenmiştir.
    - Kapsanan kişiler: Türkiye sınırları dâhilindeki 16-70 yaş grubuna ait bireylerdir.
    - Coğrafi kapsam: Örnek seçimi için Türkiye sınırları içerisindeki yerleşim yerleri kapsama dâhil edilmiştir.
- Çalışma için R programlama dili ve Oracle veri tabanı kullanılmıştır. Yöntem olarak da veri madenciliğinde sınıflandırma, regresyon ve kümeleme teknikleri kullanılmıştır.

2. LİTERATÜR ARAŞTIRMASI

Veri madenciliğinin terim olarak ilk ortaya çıkış tarihi, 1960'lı yıllara dayanmaktadır. O dönemlerde veri madenciliği yerine veri taraması gibi kavramlar ve 1970'lere gelindiğinde ise ilişkisel Veri Tabanı Yönetim Sistemleri kullanılmıştır. Burada yaşanan zorluklardan en önemlisi, verinin miktarı büyüdükçe sistemlerin doğruluğu sorgulanır olmuştur. Daha sonraki dönemlerde ise ortaya çıkan algoritmalarla veri analizleri yapılmaya başlanmıştır. 1990'lı yıllarda artık veri madenciliği tamamen literatürde yerini almıştır. Bu kavram kullanıldığı ilk zamanlarda veri tabanındaki veriler üzerinde çalışılmıştır. Fakat zamanla büyüyen veri miktarının etkisiyle veri tabanında tutulamayacak kadar büyük verilerin işlenmesi olarak gelişmiştir. Veri madenciliği, büyük miktarlardaki verinin tek başına ortaya çıkaramadığı bilgiyi belirli yöntemler yardımıyla tahmin eden veya ortaya çıkaran analiz sürecidir [1].

Veri madenciliği yöntemi, önceden bilinmeyeni tahmin edici veya tanımlayıcı modeller olarak kümeleme, sınıflama ve regresyon modellerini içermektedir. Bu modeller istenen sonucu başarılı bir şekilde elde etmek amacıyla birlikte de kullanılabilir. Bu modellerde temel olarak yapılan şey, eğer yeni bir nesnenin niteliklerini inceleme ve bu nesneyi önceden tanımlama ise sınıflandırma; eğer bu sınıflandırma denetimsiz ve öngörülecek alanların belirlenmesini, birbirine benzeyen verilerin altkümelere ayrılmasını hedefler ise kümelemedir. Bir nesnenin varlığı ile diğer bir nesnenin varlığı arasında tahmin istenir ve bu tahmin sayısal değişken ise regresyon modelidir [2,3].

Bilgi sistemleri, mevcut verileri ve geçmişteki verileri sorgulayabilmektedir. İşletmeler çoğu zaman stratejik kararlar almak ya da müşterilerine daha iyi hizmet verecek yeni politikalar uygulamak zorundadır. Örneğin, market sorumluları daha fazla ürün satın almayı teşvik etmek için marketleri yeniden tasarlamaktadırlar. Telefon şirketleri ise müşterilerini daha fazla çağrı yapmaya teşvik etmek için yeni fiyat yapıları oluşturmaktadırlar. Her iki örnekte de şirketler, veri madenciliği ile

geçmişteki müşteri davranışlarını anlayıp ona göre stratejik kararları alabilmektedirler. Veri madenciliği, büyük verilerdeki bilgiyi ortaya çıkarır ve veriler arasında muhtemel ilişkileri belirtmektedir. Veri madenciliği teknolojisinin temel bileşenleri, istatistik, yapay zekâ ve makine öğrenimi gibi araştırma alanlarında on yıllardır geliştirilmektedir. Günümüzde bu teknolojiler ilerleyerek daha önce sistemler içine gömülmüş olan bilgilerden faydalanabilecek bir iş ortamı oluşturabilmektedirler. Bankalar, aşağıdaki gibi çeşitli uygulamalar için veri madenciliğini kullanabilmektedirler [3]:

- Kart pazarlaması: Kart veren kuruluşlar, müşteri segmentlerini belirleyerek hedeflenen ürün geliştirme ve özelleştirilmiş fiyatlandırma ile kârlılıklarını artırabilmektedirler.
- Kart fiyatlandırması ve karlılık: Kart veren kuruluşlar, ürünlerin maddi imkânlarını kullanarak, kârlarını en üst düzeye çıkarmak ve müşteri kayıplarını en aza indirmek için veri madenciliği teknolojisinden yararlanabilmektedirler.
- Dolandırıcılık algılama: Veri madenciliği ile müşterilerin geçmiş işlemlerinin incelenip varsa sahtekârlık olduğu tespit edilebilmektedir.

Telekomünikasyon şirketleri, her geçen gün artmakta olan rekabetle karşı karşıya kalmaktadırlar; bu durum, mevcut müşterileri koruyabilmek ve yenilerini çekebilmek için özel fiyatlandırma programları sunmaya zorlamaktadır. Telekomünikasyonda veri madenciliği ile kullanılan bazı uygulamalar ise şunlardır [3]:

- Çağrı detay kaydı analizi: Telekomünikasyon şirketleri müşterilere ait ayrıntılı çağrı kayıtlarını inceleyip benzer kullanım modellerine sahip müşteri segmentlerini belirleyebilmektedirler.
- Müşteri sadakati: Bazı müşteriler, rakip şirketler tarafından cazip teşviklerden faydalanabilmek için hizmet aldığı şirketi sürekli değiştirmektedirler. Şirketler ise veri madenciliğini, sadık kalacak olan müşterilerin özelliklerini belirlemek ve böylece şirketlerin en fazla kâr sağlayacak müşterilere yaptıkları harcamaları hedef almalarını sağlamak için kullanabilmektedirler.

Telekomünikasyon sektörü ses, faks, cep telefonu, görüntü, e-posta, web veri iletimi ve diğer veri trafiği dâhil olmak üzere birçok kapsamlı iletişim hizmeti sunmaktan dolayı hızla gelişmiştir. Birçok ülkede bu sektördeki düzenlemelerin değişmesi, iletişim teknolojileri ve yeni bilgisayarların gelişmesinden dolayı hızla genişlemekte ve rekabet gücünü yükseltmektedir. Bu gelişmeler telekomünikasyon modellerinin belirlememize, kaynakları daha iyi kullanmamıza ve hizmet kalitesini iyileştirmemize yardımcı olmak için veri madenciliğinin kullanılmasına büyük bir talep oluşturmaktadır [4].

Dünyada iletişim alanında çok fazla değişiklikler görülmektedir. Bu değişiklerle birlikte günümüzde artık sabit hatlar kullanılmamakta bunun yerine herkesin sahip olduğu gün içerisinde çalışan bir cep telefonu kullanılmaktadır. Cep telefonları sadece bizim dünyaya bağlanmamıza aracı olmakla kalmayıp, eğlence cihazlarının da amacına hizmet etmektedir. Gelecek teknolojilere, uygun fiyatlı paketlere, kaliteli ve hızlı iletişime ilişkin müşterilerin farkındalığının artmasıyla; mobil üreticilerin müşteri talebine tamamen uygun bir paket vermeleri çok önemlidir. Önde gelen cep telefonu üreticileri en iyi ve en yeni teknolojileri oluşturarak yenilikçi pazar devleriyle rekabet edebilmektedir. Özellikle profesyonel kamera çekimleri, fotoğraf işleme ve yüksek grafikli oyun özellikleri bugünkü cep telefonunu el bilgisayarına dönüştürmüştür. Bu cep telefonları arasında veri paylaşımı başlangıçta kızıl ötesi ile sağlanmaktaydı. Kızıl ötesi ile veri aktarılmasındaki veri kayıpları çokça olmaktaydı ancak Bluetooth'un ortaya çıkışıyla bu eksiklikler giderilmiştir ve iki cihaz arasında 50 metre aralıkta veri paylaşılması sağlanmıştır. Cep telefonu dünyasında veri paylaşımındaki hızlılık ile telekomünikasyon dünyasının mobil geniş bant üzerinde yeni bir iletişim ve navigasyona odaklanmasını sağlamıştır. 3G ve 4G teknolojilerinin gelişmesiyle veri aktarım hızı ve miktarında yeni bir devrim yaşanmıştır. 5G teknolojisi ile insanların veri iletişimi, hızı ve miktarında artışlar tamamen yükseleceği ve çalışma ofisleri, yirmi birinci yüzyılın kişisel dijital asistanı olan bu cep telefonlarıyla daha da küçüleceği öngörülmektedir [5].

Ülkemiz de bu değişimlerden etkilenerek 4.5G'ye geçiş yapıp bu konuda altyapısal olarak önemli bir adım atmıştır. Bu teknolojinin getirdiği en önemli yenilikler;

yüksek hız, daha düşük gecikme süresi ve yüksek kapasitede mobil internet erişimidir. 4.5G ile yüksek hızda ve kesintisiz bağlantı sağlanabildiği için nesnelerin interneti (IoT), makineler arası iletişim (M2M), akıllı şehirler, bulut bilişim gibi birtakım konularda gelişmeler sağlanmakta ve bu sayede yaşam kalitesi daha da yükselmektedir. Daha gelişmiş 5G teknolojisinde ise, 4.5G'de 1 Gbit/s'e varan hızların 5G ile 10 kat artarak 10 Gbit/s'e kadar ulaşması öngörülmektedir. Mobil haberleşme standartlarını belirleyen Uluslararası Telekomünikasyon Birliği (ITU)'nin 5G'de hangi frekansların kullanılacağına 2019 yılında karar vermesi ve ticari olarak kullanımına ise 2020 sonrasında geçilmesi beklenmektedir. 5G ile yüksek frekans bantlarında yüksek performans sunabilecek yeni bir telsiz erişim teknolojisinin geliştirileceği tahmin edilmektedir [6].

Bugün dünya nüfusunun yarısından fazlası internet kullanmakta ve kullanılan internetin yarıdan fazlası ise cep telefonları aracılığıyladır. Bu cep telefonlarının kullandığı internet bağlantılarının çoğu yüksek hızda veri aktarımı sağlayan geniş bant teknolojisiyledir. İnternet kullanıcılarına yönelik 2017 yılında küresel ölçekte yapılan bir araştırmanın sonucunu aşağıda özetlersek:

- Küresel internet kullanıcısı sayısı yaklaşık 3.77 milyar.
- Küresel mobil kullanıcı sayısı yaklaşık 4.92 milyar.
- Küresel sosyal medya kullanıcısı sayısı 2.80 milyar.
- Küresel e-ticaret kullanıcısı sayısı 1.61 milyar.

Bu bilgiler ışığında daha önceki yapılan araştırmalara kıyasla, bir önceki yıla göre internet kullanımının 354 milyon artışla %10 büyüdüğü, benzersiz mobil kullanıcıların 222 milyon artarak %5 kadar büyüdüğü ve mobil sosyal medya kullanıcılarının ise 581 milyon artışla % 30 büyüdüğü görülmüştür. Bu oranlar veri kullanım artışını doğrudan etkileyen en önemli etmenlerdir [7].

Ülkemiz için yapılan bir diğer araştırmanın sonucunu incelediğimizde Türkiye'deki mobil kullanıcı sayısı 71 milyon iken, sosyal medyaya mobilden bağlanan kullanıcı sayısı ise 42 milyondur ve mobil cihaz kullanıcılarının %75'i akıllı telefon

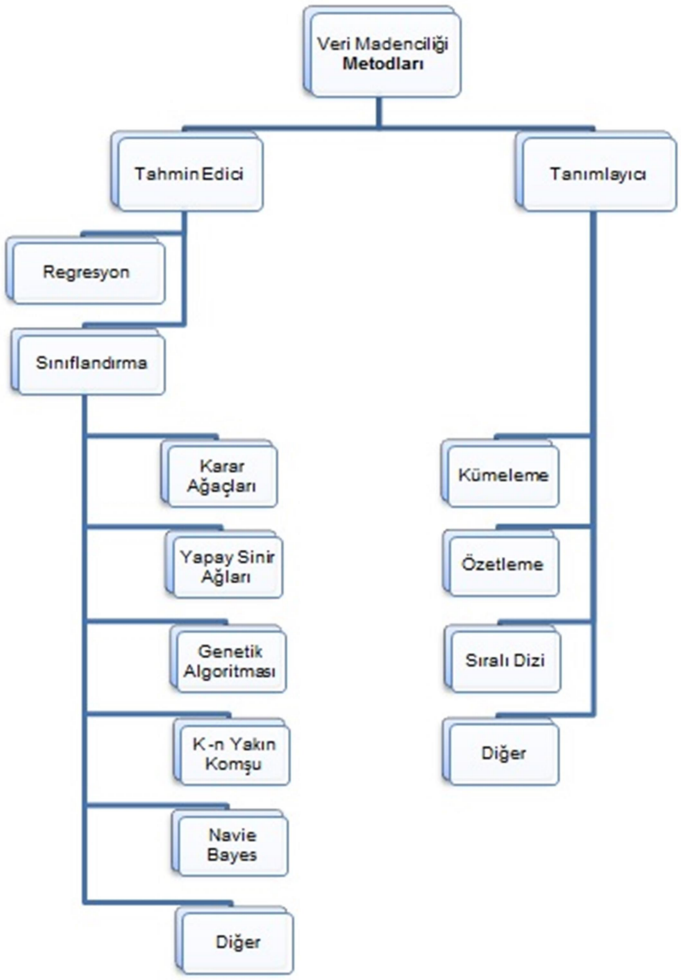
kullanmaktadır. Büyümeye rakamlarını incelediğimizde Türkiye'de 2016 Ocak ayından 2017 Ocak ayına kadar internet kullanıcı sayısının 2 milyon, aktif sosyal medya kullanıcısı sayısının ise 6 milyon arttığı ortaya çıkmıştır. Ayrıca Türkiye'nin web trafiğine göre; bilgisayar üzerinden web kullanımının %29 oranında gerilediği, mobil trafiğin ise %33 oranında artarak ülke genelinde kullanım oranının %61'e çıktığı görülmektedir. Bu durum Türkiye'de web trafiği konusunda mobil kullanımın ne derece önemli olduğunu bir kez daha gözler önüne sermektedir [8].

Literatür araştırması sonucunda günümüzde artık iletişimin ayrılmaz bir parçası olan mobil iletişim cihazlarının kullanımının yanı sıra, mobil internet, eğlence ve sosyal ağlar gibi hızlı ve büyük miktarlarda veri iletişimine ihtiyaç duyulan uygulamaların kullanılması, rekabetçi telekomünikasyon sektöründeki firmaları sürekli altyapısal iyileştirmelere zorlamaktadır. Bu çalışma da bireylerin belirli bölgelere ve yaş gruplarına göre kullandığı mobil verinin ortaya çıkarılmasıyla, sektördeki firmaların sürekli bir iyileştirmeye ihtiyaç duyduğu altyapısal yetersizliklerini önceden belirleyebilmesine ön ayak olacaktır. Ayrıca iletişimin tamamen veri üzerinden olacağı ve şu an ki teknolojiye oranla hızın ve dolayısıyla veri miktarının 10 kat artacağı 5G teknolojisi içinde altyapısal yatırım yapan firmaların yüksek maliyet açısından veri kullanım yoğunluğuna göre önceden belirlenen konumlardan başlayarak yatırım yapabilmesine yardımcı olacaktır.

3. MATERYAL VE YÖNTEM

Veri madenciliğinde kullanılan yöntemler, verinin türüne ve veri analizinden sonra ortaya çıkan sonuçlara göre tanımlayıcı ve tahmin edici olmak üzere 2 ana modele ayrılmaktadır. Tanımlayıcı modellerde, anlamlandırılması güç olan veri setinin karar vermeye öncülük etmede kullanılabilecek gizli kalmış örüntülerinin ortaya çıkarılmasıdır yani karmaşık verilerden, insanlar tarafından yorumlanabilen ilginç desenler oluşturmaktır. Bu modellerde kümeleme, özetleme gibi teknikler kullanılmaktadır. Örneğin, Bir banka, hizmet verdiği müşterilerini maaş ve diğer varlık bilgilerine göre tanımlayıcı modeller uygulayarak onları belirli müşteri gruplarına ayırabilmektedir.

Tahmin edici modellerde ise eldeki veri setinden bir model veya kural ortaya çıkarılarak sonuçları önceden bilinmeyen veri kümelerine bu kuralları uygulayarak sonuç değerlerinin tahmin edilmesidir. Örneğin, bankalar kendi aralarında ortak bilgi paylaşımı ile herhangi bir müşteriye ait verilere sahip olabilirler. Müşterinin bu bilgileri bağımsız değişken olarak ve bankalara ait borçlarının zamanında geri ödenip ödenmediği bilgileri ise bağımlı değişken olarak kullanılıp bir model kurulabilir. Kurulan bu model, bankadan kredi isteyen müşterinin bu krediyi zamanında geri ödeyip ödeyemeyeceğini tahmin edebilmektedir. Tahmin edici modellerde karar ağaçları, yapay sinir ağları, genetik algoritmalar, k-en yakın komşu ve navie bayes gibi teknikleri kullanan sınıflandırma ve regresyon yöntemleri kullanılmaktadır. [10]. Veri madenciliğinde kullanılan metotlar Şekil 3.1'de gösterilmiştir [11].



Şekil 3.1. Veri madenciliği yöntemleri

3.1. Sınıflandırma Yöntemi

Veri madenciliğinde her metot farklı amaca hizmet etmekte, kendi avantaj ve dezavantajlarını sunmaktadır. Sınıflandırma en önemli metotlardan biridir. Temel olarak yaptığı şey nesneleri tanımlamak için bazı niteliklere sahip özelliklere dayalı bir sınıflandırıcı oluşturmaktır. Daha sonra bu yeni nesnenin niteliklerini inceleyerek bu nesneyi önceden tanımlanmış bir sınıfa atamaktadır [4].

Denetimli öğrenme olarak ta adlandırılan sınıflandırmanın amacı, nesneleri tanımlamak için bazı niteliklere sahip verilerin grubunu tanımlamak ve nitelik temelinde bir sınıflandırıcı oluşturmaktır. Ayrıca veri setindeki örnekleri kullanarak

her bir sınıfa ait farklı özelliklerin ortaya çıkarılması ve bu özelliklerin belirli kurallarla açıklanmasıdır. Veri kümesinden belirli kuralların ortaya çıkarılma süreci tamamlandığında, bu kurallar yeni örneklem veri kümelerine uygulanır ve bu verilerin hangi sınıfa ait olduğu kullanılan model ile belirlenir. Bu modelde veri kümesi, kullanılan algoritmaya uygun olarak hazırlanır. Daha sonra veri kümesinin bir kısmı eğitim diğer bir kısmı ise test için ayrılır. Eğitim için ayrılan veri kümesiyle modelin öğrenimi gerçekleşir, test için ayrılan veri kümesiyle de modelin doğruluk derecesi belirlenir. Doğruluk derecesi, veri kümesinin uygulandığı modelde doğru olarak sınıflanan olayların sayısının gerçekleşen tüm olayların sayısına bölünmesiyle ortaya çıkar. Hatalı olarak sınıflanan olay sayısının, gerçekleşen tüm olayların sayısına bölünmesiyle de hata oranı ortaya çıkmaktadır [4]. Sınıflandırma yönteminde en yaygın kullanılan teknikler [14];

- Karar ağaçları teknikleri
- Yapay sinir ağları teknikleri
- Genetik algoritmalar
- K-En yakın komşu algoritmaları
- Navie-Bayes teknikleri
- Kural tabanlı algoritmalar

Navie-Bayes, her bir değişkenin sonuca olan etkilerinin olasılık olarak hesaplanmasına dayanan bir istatistiksel sınıflandırma tekniğidir. Bayes sınıflandırma, büyük veri tabanlarına uygulandığı zaman yüksek hız ve doğruluk oranı sergilemektedir. Bu yöntem bir örnekle açıklanırsa; bayes modeli, hava şartları, nem, sıcaklık ve rüzgâr bilgilerine göre tenis oynayıp oynayamama tahmininde bulunabilmektedir. Sisteme değişken bilgilerinin tüm olasılıkları tenis oynar veya oynayamaz şeklinde öğretilmektedir. Eğitilmiş sistemde bu model, hava yağmurlu, nem var, soğuk ve rüzgârın var olduğu bilgilerine göre tenis oynanamaz tahminini bildirebilmektedir.[10]

Kural tabanlı sınıflandırmada, IF-THEN kuralları kümesi kullanılır. Bilgiyi veya bilginin küçük parçalarını temsil etmenin en iyi yolu kurallardır. IF-THEN kuralında

IF koşul THEN sonucun bir ifadesidir. Örnek olarak, yaş ve öğrenci bilgilerinin bulunduğu değişkenler ile kontörlü mobil hat kullanımının tahmini için önceden eğitilmiş sisteme IF yaş = gençlik ve öğrenci = evet THEN ön ödemeli hat kullanımı vardır bilgisini verebilmektedir. Bir kuralın "IF" kısmı (veya sol tarafı), öncül kural veya önkoşul olmaktadır. "THEN" kısmı (veya sağ taraf) sonuçta ortaya çıkan kuraldır. Kuralın sonucunda ise bir sınıf tahmini bulunmaktadır [10].

K-en yakın komşu sınıflandırıcılar benzetme yoluyla öğrenmeye, yani belirli bir test grubunu ona benzeyen eğitim gruplarıyla karşılaştırmayı temel almaktadır. Eğitim grupları n niteliklerle tanımlanıp ve her bir grup, n-boyutlu bir uzayda bir noktayı temsil etmektedir. Bu şekilde, tüm eğitim grubu n-boyutlu bir desen alanında saklanmaktadır. Bilinmeyen bir örneklem verildiğinde, k-en yakın komşu sınıflandırıcısı, bilinmeyen örnekleme en yakın olan k desen alanını aramaktadır. Bu yakınlık Öklid uzaklığı ile tanımlanmaktadır. Bilinmeyen örneklem, k en yakın komşu içinden en çok benzediği sınıfa atanmaktadır. En yakın komşu sınıflandırıcılar, sayısal tahmin için yani verilen bir bilinmeyen örneklem için gerçek değerli bir tahminin döndürülmesi içinde kullanılabilir. Bu durumda, sınıflandırıcı ortalama değeri döndürmektedir. Bu yöntem büyük eğitim setleri ile daha yavaş çalışmaktadır [10].

Genetik algoritmalar, doğal evrim fikirlerinin sürecine benzer şekilde çalışan algoritmalardır. Genetik fonksiyonların, bilgisayar problemlerine uygulanması ile problemlerin çözümünü hedefleyen bir yaklaşımdır. En güçlü kişinin hayatta kalma kavramına dayanarak, mevcut nüfustaki en uygun kuralların yanı sıra bu kuralların soyundan gelen yeni bir nüfus oluşturmaktadır. Problemlere tek bir çözüm yerine farklı çözümler içeren çözüm kümesi üretmektedirler. Genetik algoritmalar genellikle ikili kodlama, permütasyon kodlama ve değer kodlama türlerinde çalışmaktadır. İkili kodlama türünde kromozomlar “1011” gibi değerlerle ifade edilmektedir. Değer kodlama türünde, kromozomlar “ABCFG” veya “ileri-geri-yukarı” gibi ifadelerle temsil edilmektedir. Permütasyon kodlama türünde ise “(8 5 6 7 2)” gibi değerlerle ifade edilmektedir. Genetik algoritmalar, yeniden üretim, çaprazlama ve mutasyon gibi genetik operatörler kullanmaktadırlar. Yeniden üretim,

var olan neslin yerine yeni bir nesil oluşturmaktır. Çaprazlama, atanın sahip olduğu ikili sisteme ait kromozom değerlerinin yerleri değiştirilerek en uygun değere sahip çocuk kromozomlarını ikili sistemde üretmektir. Mutasyonda, kuralın dizesindeki rasgele seçilen bitler ters çevrilerek kromozom çeşitliliği artan yeni nesil elde edilmektedir. Genetik algoritmalar için kullanılan en yaygın örnek gezgin satıcı probleminin çözümüdür. Genetik algoritmalar günümüzde yaygın olarak quantum hesaplama araçlarında ve CPU hızlandırma gibi konularda kullanılmaktadır [10].

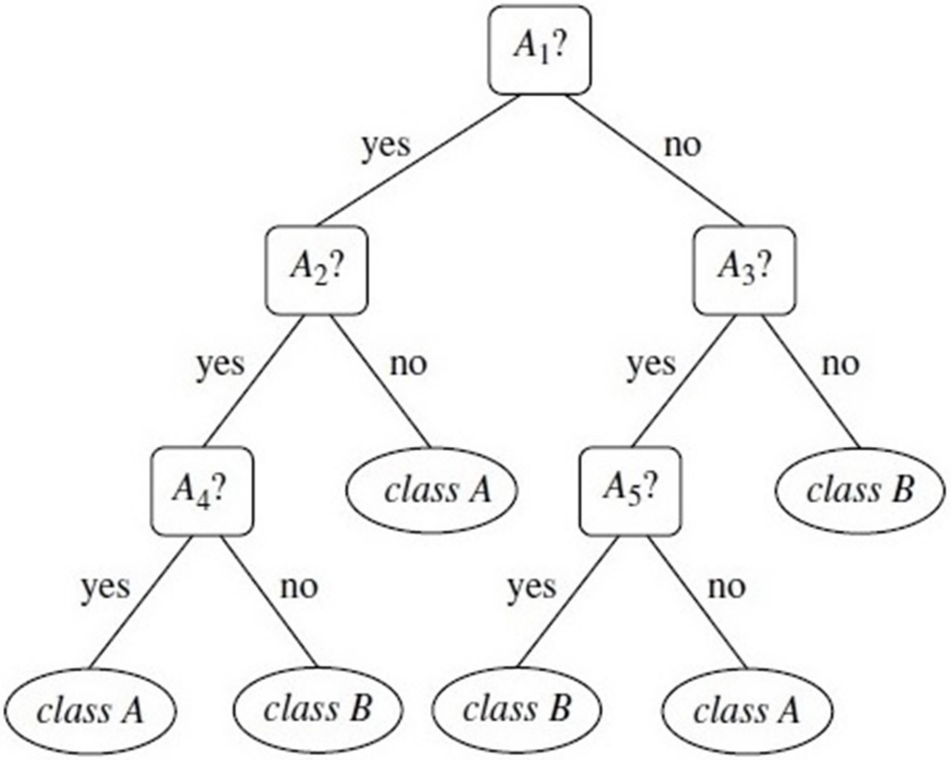
Yapay sinir ağları insan beyninin yapısını temel alarak oluşturulmuş tekniktir. Yapay sinir ağlarının en önemli özelliği öğrenme veya eğitim sürecinde öğretilmeyen girişler için uygun sonuçlar üretebilmesidir. Bu sebeple karmaşık problemlerin çözümünde yaygın olarak kullanılmaktadır. Yapay sinir ağları yaygın olarak görüntü ve ses tanımlama sistemleri, tahmin ve kestirim sistemleri, arıza analizi ve tıp gibi alanlarda kullanılmaktadır [14].

3.1.1. Karar ağaçları teknikleri

1970'lerin sonları ve 1980'lerin başında, makine öğrenme araştırmacısı olan J. Ross Quinlan tarafından ID3 (Yinelemeli Ayrıştırıcı) olarak bilinen bir karar ağacı algoritması geliştirmiştir. Daha sonra Quinlan, C4.5 algoritmasını (ID3'ün yerini alan) ortaya çıkarmıştır. 1984'te ise bir grup istatistikçi ikili karar ağaçlarının üretimini tanımlayan Sınıflama ve Regresyon Ağaçlarını (CART) yayınlamışlardır. ID3 ve CART algoritmaları aynı zamanlarda birbirinden bağımsız olarak bulunmuştur ancak karar ağaçlarını oluşturmak için benzer bir yaklaşım izlemektedirler. ID3, C4.5 ve CART algoritmaları, karar ağaçlarının yukarıdan aşağıya, yinelemeli böl - fethet biçiminde yapılandırıldığı geriye dönüşü olmayan bir yaklaşımı benimsemektedirler. Karar ağaçlarında çoğu algoritma yukarıdan aşağıya doğru olan yaklaşımı uygulamaktadır. Karar ağaçlarında yaygın olarak kullanılan diğer algoritmalar ise rastgele orman (random forest), hızlandırılmış ağaçlar (boosted trees), döndürme ağacı (rotation forest), C5.0 ve MARS algoritmalarıdır [4].

Veri madenciliğinde karar ağaçları, kolayca yorumlanabilir olması, güvenilirliklerinin iyi olması ve ağaç yapısıyla da kolay, anlaşılır kuralların oluşturulabilmesi nedenlerinden dolayı sınıflama modelleri içerisinde en popüler sınıflama tekniğidir. Tahmin edici bir teknik olan karar ağacı kök, düğüm ve yaprakları olan bir ağaç görünümündedir [12].

Karar ağacı oluşumunda gerçekleşecek modeli karar düğümü belirler. Bu modelin karar düğümü üzerinde uygulanması sonucunda ağaç dallara ayrılır. En üst seviyedeki ayırımlara bağlı olarak ağacın ardışık olarak düğüm ve dallara ayrılması işlemi gerçekleşir. Dallara ayrılan ağaçta eğer dalın ucundaki veriler bir sınıf oluşturamıyorsa burada bir düğüm oluşur ve düğümden sonra tekrar dallara ayrılma işlemi ardışık olarak gerçekleşir. Eğer ağacın dalındaki verilerle bir sınıf oluşabiliyorsa burada artık bir yaprak oluşur ve bu yaprak veri kümesinden bir grup benzer niteliklere sahip veriyi temsil eder. Bu işlemler yukarıdan aşağıya doğru ilerler. Kök düğümünden başlayıp yaprağa ulaşıncaya kadar ardışık olarak devam eder [14]. Basit bir karar ağacı Şekil 3.2'de gösterilmiştir [4].



Şekil 3.2. Karar ağacı teknikleri

Basit bir şekilde karar ağacı, veri kümesinde bulunan verilerin belirli niteliklerine bakarak daha küçük veri kümesi gruplarına bölme işlemidir. Başarılı olarak bölünen grupların üyeleri nitelik bakımından birbirleriyle daha benzer olmaktadır. Veri miktarının büyük olduğu veri kümelerinin sınıflandırma problemlerinde en uygun çözüm yöntemi karar ağaçları yöntemleridir [13].

Karar ağaçları çok boyutlu verileri işleyebilmektedir. Elde edilen bilginin ağaç biçiminde temsil edilmesi insanlar tarafından kolayca benimsenmesini sağlamaktadır. Karar ağacında öğrenme ve sınıflandırma adımları basit ve hızlı olmaktadır. Genellikle karar ağaçlarının doğruluğu da iyidir. Bununla birlikte, başarı oranı eldeki verilere bağlı olabilmektedir. Bazı karar ağacı algoritmaları yalnızca ikili ağaçlar üretirken, bazıları ise ikili olmayan ağaçlar üretebilmektedirler. Karar ağacı algoritmaları, tıp, imalat ve üretim, finansal analiz, astronomi ve moleküler biyoloji gibi birçok uygulama alanında kullanılmaktadır. Karar ağaçları, birçok ticari kurallar sistemlerin temelini oluşturmaktadır [4].

Karar ağacı algoritmalarının bilinen en önemli avantajları;

- Ağaç yapısından dolayı insanlar tarafından anlaşılması ve yorumlanması kolay olmaktadır.
- Ön işleme hızlıdır ve büyük verilerde çok az bir ön işleme ile veri kullanılabilir hale gelmektedir.
- Bazı algoritmalar sadece sayısal problemlerde kullanışlı iken bazıları ise sadece kategorik verilerde kullanışlıdır. Karar ağaçları ise hem sayısal hem kategorik verilerle çalışabilmekte ve kullanışlı sonuçlar üretebilmektedir.
- Her adım ayrı ayrı yorumlanabilir ve görüntülenebilmektedir.
- Büyük verileri kısa sürede işleyebildiğinden dolayı basit ve hızlı bir yapıya sahip olmaktadır [10].

3.2. Regresyon Yöntemi

Sınıflandırma yöntemi kategorik değerlerin tahmin edilmesinde, regresyon yöntemi ise süreklilik gösteren sayısal değerlerin tahmin edilmesinde kullanılmaktadır. Regresyon yöntemi, bir veya daha fazla sürekli değişken niteliğe sahip verilerin diğer özniteliklerini temel alan önceden tahmin yöntemidir. Regresyon, veri kümesinde bulunan ölçüt değişkeni ile diğer tahmin değişkenleri arasındaki ilişkiyi sayısal bir formüle dönüştürmede kullanılan istatistiksel analizdir. Bir başka ifadeyle değişkenler arasında ilişkinin niteliğini bulmayı amaçlayan bir analiz yöntemidir [10].

Regresyon, değişkenlerin sayısal olduğu bir veya daha fazla bağımsız değişkenden bir bağımlı değişkenin değerini tahmin etmek için kullanılır. Veri madenciliğinde kullanılan yaygın regresyon modelleri doğrusal ve çoklu regresyondur. Doğrusal Regresyon, bağımlı ve bağımsız değişken arasında ikili bir ilişki var olduğu durumlarda kullanılmaktadır. Sadece bir bağımlı değişken ile bir bağımsız değişkenin bulunduğu veri setlerine bu regresyon modeli uygulanabilmektedir. Kullanılan tahmin değişkenlerinin değerlerinden hareketle ölçüt değerinin tahmin edilebilmesi amaçlanmaktadır Doğrusal regresyon yönteminde bir adet tahmin değişkeni kullanılırken çoklu regresyon yönteminde ise birden fazla tahmin değişkeni kullanılmaktadır [15].

3.2.1. Çoklu regresyon teknikleri

Birden fazla bağımsız değişkenin bir bağımlı değişkeni etkileyebildiği durumlarda çoklu regresyon tekniği kullanılmaktadır. Çoklu regresyon, bağımlı bir değişken üzerinde etkisi olan birden fazla bağımsız değişken ile bu bağımlı değişken arasındaki ilişkiyi ortaya çıkaran analiz yöntemidir [15]. Bu analiz yönteminin sonucunda bağımlı değişken ile bu değişkeni etkileyen bağımsız değişkenler arasındaki ilişkiyi matematiksel formüllerle açıklayabilmek amaçlanmaktadır.

Çoklu regresyonda, modele dâhil edilecek bağımsız değişkenlerin neler olacağı ve bunların dâhil ediliş biçimi regresyon analizinin sonucunu etkilemektedir. Araştırmacının bu modele değişkenlerin tümünü birden dâhil edebildiği gibi değişkenleri istediği biçimde ya da sırada dâhil edebilir. Burada amaç araştırmacının belirlediği değişkenlerin, bağımlı değişken üzerindeki etkilerinin incelenmesidir [15]. Bu modelde ortaya çıkarılan katsayıların güvenilirliği standart hata değerlerine bakılarak değerlendirilmektedir. Regresyon modellerinde belirlilik katsayısı 1'e en yakın değerler için kötü sonuç elde edilirken 0'a en yakın değerler için en iyi ve anlamlı sonuç elde edilmektedir [16].

Çoklu regresyonda doğru model oluşturulmaya çalışılırken değişken eleme veya değişken ekleme yöntemleri kullanılmaktadır. Değişken ekleme yönteminde tüm değişkenleri modele katmadan sadece belirlenen sabit değişkenle model oluşturulur. Daha sonra oluşturulan bu model ile bağımsız değişkenler arasından modele en yakın korelasyon değerine sahip bağımsız değişken seçilerek modele eklenilir. Eklenen bu değişken modele anlamlı katkı sağlıyorsa modelde tutulur. Bu şekilde eklendikten sonra anlamlı katkı getiren diğer bağımsız değişkenler içinde işlemler tekrarlanarak model oluşturulur. Değişken eleme yönteminde ise, oluşturulan başlangıç modeline diğer bağımsız değişkenlerin tümü dâhil edilir. Sonrasında bağımsız değişkenler arasından oluşturulan bu modele göre, sonuç ile en düşük korelasyon sergileyen bağımsız değişken modelden çıkarılır. Modelden çıkarılan bu değişken, modelin bağımsızlığını anlamlı derecede düşürürse modelde tutulur. Bu işlemler, modelde sadece anlamlı derecede bağımsız değişkenler kalana kadar tekrar edilir [15].

Çoklu doğrusal regresyon modeli denklem 3.1' de verilmiştir.

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n \tag{3.1}$$

b_0 : Doğrunun y eksenini kestiği nokta

$b_1,\ b_2, \ldots, b_n$: Regresyon katsayıları

$X_1,\ X_2, \ldots, X_n$: Bağımsız değişken değerleri

3.3. Kümeleme Yöntemi

Kümeleme yöntemi, sınıflandırma yönteminden farklı olarak denetimsizdir ve veriler arasında benzer niteliklere sahip olanları altkümelere ayrılmasını hedefler. Kümeleme yönteminde hangi nesnenin hangi altkümeye ait olduğu belli değildir. Bu yöntem nesnelerin birbirine benzeyen özelliklerine göre gruplara ayrılmasıdır. Aynı gruptaki nesneler farklı gruptaki nesnelere göre nitelik bakımından birbirlerine daha çok benzerler. Kümeleme modellerinde amaç, nesnelerin birbirlerine çok benzediği halde özellikleri bakımından birbirlerinden çok farklı altkümelere bölünmesi ve bu şekilde veri tabanlarında saklanmasıdır [9].

Kümeleme, veri nesnelerinin kümesini birden çok gruba ayırmaktadır. Küme analizi veya basitçe kümeleme, bir takım veri nesnelerinin alt gruplara bölünmesi işlemidir. Her bir alt küme, bir kümedir ve kümedeki nesneler birbirine benzemektedir, ancak diğer kümelerdeki nesnelerle benzer değildir. Farklı kümeleme yöntemleri, aynı veri kümesinde farklı kümeler oluşturabilir. Bölümleme, insanlar tarafından değil kümeleme algoritması tarafından gerçekleştirilmektedir. Bu nedenle, kümeleme, veri içindeki daha önce bilinmeyen grupların keşfedilmesine yol açabilmesi açısından yararlıdır [4].

Küme analizi, iş zekâsı, görüntü kalıbı tanıma, web araması, biyoloji ve güvenlik gibi birçok alanda yaygın şekilde kullanılmaktadır. İş zekâsında kümeleme, çok sayıda müşteriyi bir grup içindeki müşterilerin güçlü benzer özelliklere sahip olduğu gruplar halinde organize etmek için kullanılabilir. Bu, gelişmiş müşteri ilişkileri yönetimi için iş stratejilerinin geliştirilmesini kolaylaştırır. Ayrıca, çok sayıda projeye sahip bir danışman şirketini örnek vermek gerekirse; proje yönetimini iyileştirmek için kümeleme ve bölünme yöntemlerini uygulayarak projeleri benzerliğe dayalı kategorilere dönüştürebilir. Böylece proje denetimi ve teşhisi proje teslimini ve sonuçlarını iyileştirmek için etkin bir şekilde yönetilebilir [4].

Bir veri madenciliği işlevi olarak küme analizi, verilerin dağılımı hakkında fikir edinmek, her kümenin özelliklerini gözlemlemek ve daha ileri analizler için bağımsız

bir araç olarak kullanılabilmektedir. Ayrıca tespit edilen kümeler nitelik altkümesi seçimi ve sınıflandırma gibi diğer algoritmalar için bir ön işlem basamağı görevi görebilmektedir [4].

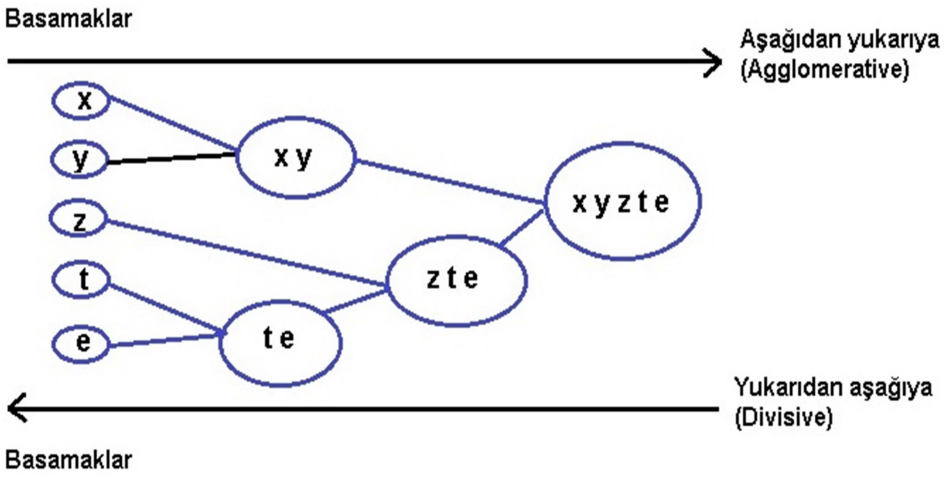
Kümeleme yönteminde, sınıflandırma yöntemleri gibi verilerin ait olduğu herhangi bir sınıf bulunmamaktadır. Sınıflandırma yöntemlerinde eğitilen modelin sınıfları bilinmektedir ve bilinmeyen bir veri geldiğinde hangi sınıfa ait olacağı tahmin edilebilmektedir. Fakat kümeleme yönteminde, veriler kendi içinde benzer gruplara ayrılmaktadırlar. Bazen kümele işlemleri, sınıflandırma yöntemi için bir önişlem olarak ta kullanılabilmektedir. Kümeleme yöntemlerinde kullanılan birçok algoritma mevcuttur. Bu algoritmalar verinin tipine ve amacına göre kullanılabilmektedirler. Yaygın olarak kullanılan bu algoritmalar aşağıdaki şekilde gruplanabilir [10].

1. Bölme metodu (Partitioning methods)
2. Yoğunluk tabanlı metodlar (Density-based methods)
3. Hiyerarşik metodu (Hierarchical methods)
4. Model tabanlı metodları (Model-based methods)
5. Izgara tabanlı metodları (Grid-based methods)

Bölme metodu, n adet verinin k adet altkümeye bölünmesidir. Verilerin bölünmesiyle oluşan altkümelerde, nesneler kendi içinde benzer fakat altkümeler arasında farklıdır. En çok kullanılan bölme metodu algoritmaları ise k-means ve k-medoids yöntemleridir. K-means yöntemi, altkümelerin ortalamasının hesaplanabildiği durumdaki veriler için kullanılabilmektedir. K-medoids yöntemi ise n adet nesnede, verilerin çeşitli özelliklerini temsilen k adet temsilci nesne bulabilmektir. Temsilci nesneler oluşturulan kümelerin merkezidir. K-medoids ile k-means yöntemlerinin en önemli farkı merkez noktalarının belirlenme şeklidir [10].

Hiyerarşik kümeleme metodu, n adet nesne başlangıçta n adet küme olarak seçilmektedir ve bu nesneler en yakın kümelerle birleştirilme esasına dayanmaktadır. Oluşturulan altkümeler ise bir küme ağacı şeklinde gruplara ayrılmaktadır. Hiyerarşik kümelemede küme ağaçları, aşağıdan yukarıya ve yukarıdan aşağıya

olmak üzere 2 sınıfa ayrılmaktadır. Aşağıdan yukarıya doğru olan kümelemede, öncelikle her nesne kendi kümesini oluşturmakta ve bu küçük kümeler birleşerek daha büyük kümeleri oluşturmaktadır. Bu birleşme her nesnenin dâhil olduğu tek bir küme oluşuncaya kadar devam etmektedir. Yukarıdan aşağıya doğru olan kümelemede ise her nesnenin dâhil olduğu tek bir küme oluşturulmakta ve bu tek küme bölünerek daha küçük kümeleri oluşturmaktadır. Bu bölünme her bir nesnenin kendi kümesini oluşturana kadar devam etmektedir. Yukarından aşağıya ve aşağıdan yukarı doğru oluşan hiyerarşik kümeleme metodu Şekil 3.3'de gösterilmiştir [10].



Şekil 3.3. Hiyerarşik kümele yöntemleri

3.3.1. K-Means kümeleme teknikleri

En iyi bilinen ve en yaygın kullanılan bölümleme yöntemlerinden biri k-means yöntemidir [9]. K-means algoritması, veri tabanında bulunan nesnelerin k adet farklı altkümeye bölümlenmesini sağlar. Kümeleme sonucu her bir altkümenin kendi içindeki elemanlar arası benzerlikler çok iken, altkümeler arası eleman benzerlikleri çok düşüktür [4].

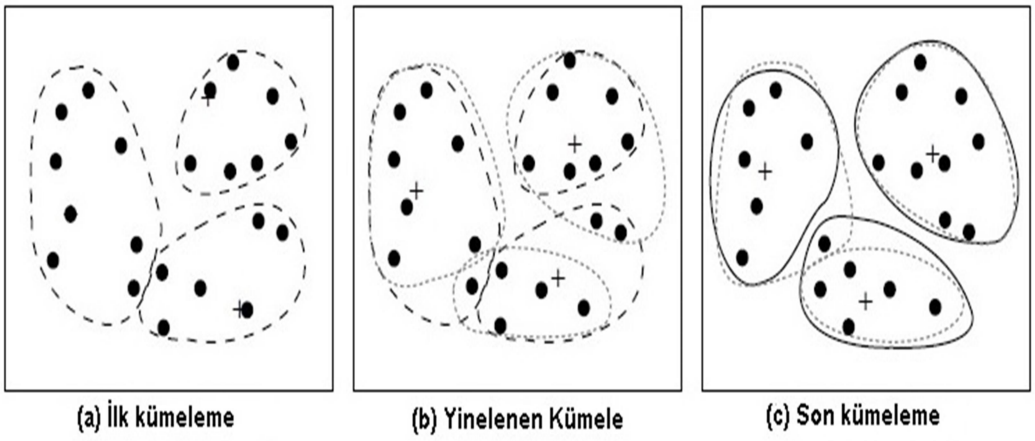
K-means algoritması, bir kümenin merkezini nesnelerin ortalama değeri olarak tanımlamaktadır. İlk olarak, n sayıda nesnelerin içindeki merkezi rastgele seçer, bunların her biri başlangıçta bir kümenin ortalamasını veya merkezini temsil eder.

Kalan nesnelerin her biri için, nesne ile küme ortalaması arasındaki Öklid mesafesine dayalı olarak, en yakın olan kümeye atanmaktadır. Daha sonra k-means algoritması, kümelemedeki varyasyonları tekrar tekrar devam ettirmektedir. Her küme için bir önceki yinelemede kümeye atanan nesneleri de kullanarak kümenin yeni ortalamasını hesaplanmaktadır. Mevcut yinelemede oluşturulan kümeler önceki yinelemede oluşturulan kümelerle aynı olana kadar devam etmektedir [4]. Belirlenmiş bir k değeri için k-means kümeleme algoritması 4 aşamada gerçekleşmektedir:

1. Veri kümesi her grup bir altküme olacak şekilde k sayıda altkümeye ayrılır.
2. Kümedeki nesnelerin niteliklerinin ortalaması merkez nokta olduğundan dolayı her demetin ortalaması hesaplanır.
3. Nesneler kendi merkez noktalarına en yakın olan kümeye atanır.
4. Nesnelerin bulunduğu kümelerde değişiklik olmayana kadar işlemler aşama 2'den devam edilir [4].

K-means algoritmasının bilinen en yaygın problemi ise merkez nokta için başlangıçta kötü bir seçim yapılırsa kümelemedeki değişiklik çok sık olur ve istenmeyen farklı sonuçlar oluşabilir. Ayrıca dışarda kalan değeri büyük nesnelerde, dâhil olacağı kümenin merkez noktasında büyük derecede sapma meydana getirebilir [17].

K-means kümeleme metodunun sonuçları, küme merkezlerinin ilk rastgele seçimine bağlı olabilmektedir. Uygulamada iyi sonuçlar almak için, k-means kümelemenin farklı ilk kümelenme merkezleri ile birden çok defa çalıştırılması yaygındır [4] K-means yöntemini kullanarak bir takım nesnelerin kümelenmesi basit şekilde Şekil 3.4'te gösterilmiştir [4].



Şekil 3.4. K-means kümeleme teknikleri

K-means kümeleme yönteminin bilinen en büyük avantajları ise küme sayısının başlangıçta bilinmesi sayesinde doğruluk oranı en yüksek olan yöntemdir. Ayrıca en büyük özelliklerinden biri büyük miktardaki veriler üzerinde yüksek doğruluk payıyla çalışabilmesidir.

4. GERÇEKLEŞTİRİLEN ÇALIŞMA

Bu çalışmada herhangi bir X firmasından belirli bölgelerde ve yaş grupları arasında mobil veri kullanımına ait veri seti elde edilerek bu verilerin veri madenciliğinde kullanılan yaygın yöntemlerle işlenmesi ve Türkiye'de bireylerin mobil veri kullanımının belirli özellik bilgilerine göre (bulunduğu bölge, yaş grubu ve cinsiyet) ortaya çıkarılmasıdır. Açık kaynak kodlu olması, diğer diller ve programlar ile bağlantı desteği, geniş bir yelpazede grafiksel teknikler içermesinden dolayı R programlama dili tercih edilmiştir.

Çalışmada veri madenciliği uygulamasının Bölüm 3'te açıklanan model ve yöntemler kullanılarak verinin hazırlanması, çalışmanın gerçekleştirilmesi ve model oluşturma aşamalarındaki işlemler anlatılacaktır.

4.1. Veri Setlerinin Oluşturulması

Türkiye sınırları içerisinde ve yaş aralığı 16-70 olan bireylere ait kullanılan mobil verilerin veri kümesi çalışmada kullanılmıştır. Bu veri kümesine ek olarak Türkiye'deki şehirlerin enlem ve boylam bilgilerini içeren veri seti Google haritalar şirketinden elde edilmiştir. Veri setimiz şu 4 ana nitelikten oluşmuştur: Şehir, cinsiyet, yaş aralığı ve kullanılan mobil veri bilgisidir. Bu ana bilgiler Tablo 4.1'de gösterilmiştir.

Tablo 4.1. Veri kümesi ana nitelikler

ALAN ADI	VERİ TÜRÜ	VERİ BİLGİSİ
CITY	Karakter	Şehir bilgisi
GENDER	Karakter	Cinsiyet bilgisi
AGE_GROUP	Karakter	Yaş aralığı bilgisi
USAGE_DATA_KB	Sayı	Kullanılan veri bilgisi -KB-

Veri kümesindeki ana niteliklerin yanında ek olarak bölge ismi, kullanılan mobil veri aralığı, enlem-boylam bilgisi, kullanılan verinin MB ve GB nitelikleri de yer

almaktadır. Bu bilgiler Tablo 4.2'de gösterilmiştir. Kullandığımız modellere uygun olan nitelikler veri setinden seçilerek modele uygulanmıştır.

Tablo 4.2. Veri kümesi diğer nitelikler

ALAN ADI	VERİ TÜRÜ	VERİ BİLGİSİ
REGION_NAME	Karakter	Bölge adı
USAGE_DATA_MB	Sayı	Kullanılan veri bilgisi -MB-
USAGE_DATA_GB	Sayı	Kullanılan veri bilgisi -GB-
USAGE_DATA_GB_GROUP	Karakter	Kullanılan veri aralığı bilgisi
LAT	Sayı	Enlem Bilgisi
LNG	Sayı	Boylam Bilgisi

Veri madenciliği öngörü modeli ile ilgili daha önce yapılan çalışmalar incelendiğinde, birden çok tahmin edici değişkene sahip ve tahmin edilmesi istenen değişkenin veri türünün sayısal değer olduğu durumlarda, öğrenme modeli olarak regresyon yöntemi kullanıldığı görülmüştür. Tahmin edilmesi istenen değişkenin veri türünün belirli grup verisi olduğunda ise sınıflandırma ve kümeleme tekniklerinin kullanıldığı görülmüştür. Elimizdeki veri setinde bireyin mobil veri kullanım bilgisi sayısal değer içerdiğinden dolayı regresyon yöntemi uygulanmıştır. Sınıflandırma ve kümele tekniğini uygulayabilmek için mobil veri kullanım bilgisi belirli kategorik aralıklarla gruplandırılarak yeni bir değişken olarak modellere uygulanmıştır.

Veri seti, Oracle veri tabanında scriptler ve SQL sorgularıyla düzenlenmiştir. Veri setinde KB şeklinde olan veri kullanım bilgilerinin MB ve GB çevrimleri yapılmıştır ve her bir GB kullanım aralığıyla ilgili gruplamalar yapılıp bu grup değerlerine karşılık gelen sayısal değerler verilmiştir. Daha sonra elde edilen bu veri setinde şehir değişkenine karşılık gelen sayısal enlem ve boylam bilgileri her şehir için ayrı ayrı eşitlenip veri setine eklenilmiştir. Bu konum bilgileri şehir merkezinin dünya üzerindeki geçerli enlem boylam bilgileridir. Ayrıca her şehir bilgisinin karşılığı olan şehir isimleri ve bulunduğu bölge isimleri de veri setine eklenmiştir.

Veriler, Oracle veri tabanında SQL sorguları kullanılarak eksik nitelik barındıran kayıtlar da temizlenerek sınıflandırma, kümeleme ve regresyon analizi için kullanılabilir biçime getirildikten sonra, herhangi bir eksik veri içermeyen 128.581

kayıttan oluşan bir veri kümesi elde edilmiştir. Elde edilen veri kümesindeki değişkenler, veri türleri ve değişkenlerin açıklamaları Tablo 4.3'te gösterilmiştir.

Tablo 4.3. Veri kümesi değişken bilgileri

ALAN ADI	VERİ TÜRÜ	VERİ BİLGİSİ
ID	Sayı	Veri kayıt numarası
CITY_ID	Sayı	Şehir plaka bilgisi
AGE_GROUP	Karakter	Yaş aralığı bilgisi
AGE_ID	Sayı	Yaş aralığı sayısal bilgisi
GENDER	Karakter	Cinsiyet bilgisi
GENDER_ID	Sayı	Cinsiyet sayısal bilgisi
DATA_USAGE_GB_ID	Sayı	Veri kullanımı sayısal bilgisi
DATA_USAGE_GB_GROUP	Karakter	Veri kullanım grup bilgisi
DATA_USAGE_GB_GROUP_ID	Sayı	Veri kullanım grup sayısal bilgisi
DATA_USAGE_GB	Sayı	Veri kullanımı GB
DATA_USAGE_KB	Sayı	Veri kullanımı KB
DATA_USAGE_MB	Sayı	Veri kullanımı MB
CITY	Karakter	Şehir bilgisi
REGION_ID	Sayı	Bölge sayısal bilgisi
REGION_NAME	Karakter	Bölge bilgisi
LAT	Sayı	Enlem bilgisi
LNG	Sayı	Boylam bilgisi

Veri setinde karakter olarak tutulan değişkenlerin her birine karşılık yeni sayısal değerler tanımlanarak bu değişkenler sayısallaştırılmıştır. Veri kullanım aralığı değişkeninin sayısallaştırılması işlevinde her bir veri kullanım değeri için 1'den başlanarak artan numerik değerlerle eşleştirilmiştir. Mobil veri kullanım aralığına karşılık artan numerik değerler eşleştirilerek yeni değişken olarak eklenmiştir. Yeni sayısal değerler Tablo 4.4'de gösterilmiştir.

Tablo 4.4. Veri kullanım bilgisi için sayısal değer

DATA_USAGE_GB_GROUP	DATA_USAGE_GB_GROUP_ID
Veri kullanım miktar grubu	**Veri kullanım grubu sayısal değeri**
0-1	0
1-2	1
2-4	2
4-6	3
6-8	4
8-10	5
10-12	6
12-14	7
14-16	8
16-18	9
18-20	10
20+	11

Veri setinde bireylerin cinsiyet bilgisinin tutulduğu değişkenin sayısallaştırılması işlevinde cinsiyet değerleri 1'den başlanarak artan numerik değerlerle eşleştirilmiştir. Yeni sayısal değerler Tablo 4.5'te gösterilmiştir.

Tablo 4.5. Cinsiyet için sayısal değer

GENDER	GENDER_ID
Cinsiyet	**Cinsiyet sayısal değeri**
E	2
K	1

Veri setinde bölge bilgisinin tutulduğu değişkenin sayısallaştırılması işlevinde bölge isim bilgileri 1'den başlanarak artan numerik değerlerle eşleştirilmiştir. Yeni sayısal değerler Tablo 4.6' da gösterilmiştir.

Tablo 4.6. Bölge bilgisi için sayısal değer

REGION_ID	REGION_NAME
Bölge adı sayısal bilgisi	**Bölge adı**
1	Akdeniz Bölgesi
2	Doğu Anadolu Bölgesi
3	Ege Bölgesi
4	Güneydoğu Anadolu Bölgesi
5	İç Anadolu Bölgesi
6	Marmara Bölgesi
7	Karadeniz Bölgesi

Veri setinde bireylerin yaş aralığı bilgisinin tutulduğu değişkenin sayısallaştırılması işlevinde yaş aralık bilgileri 1'den başlanarak artan numerik değerlerle eşleştirilmiştir. Yeni sayısal değerler Tablo 4.7'de gösterilmiştir.

Tablo 4.7. Yaş bilgisi için sayısal değer

AGE_ID	**AGE_GROUP**
Yaş aralığı sayısal bilgisi	**Yaş aralığı**
1	0-20
2	20-30
3	30-40
4	40-50
5	50-60
6	60+

Elde edilen verilerle, veri madenciliği çalışması için kullanılacak programlara uygun dosya formatlarının oluşturulması aşamasında ise veri madenciliğinde sınıflandırma, kümeleme ve regresyon modellerini oluşturmak için R Studio programı kullanılmıştır. R Studio, R programlama dilinin kullanımını kolaylaştıran kullanıcı dostu bir uygulama geliştirme aracıdır. Bu program veri girdilerini örüntü dosyalarından almaktadır. Bu dosya formatını oluşturmak için Oracle veri tabanını kullanan ve uygulama geliştirme aracı olan Toad for oracle yazılımı kullanılmıştır. Veriler, Oracle veri tabanında Tablo 4.8'deki gibi tek bir tablo üzerinde birleştirilmiştir. Daha sonra bu veriler Toad yardımıyla, sekmeyle ayrılmış metin formatında kaydedildikten sonra R studio ile içeri alınmıştır.

Tablo 4.8. Veri kümesinde bulunan değerler

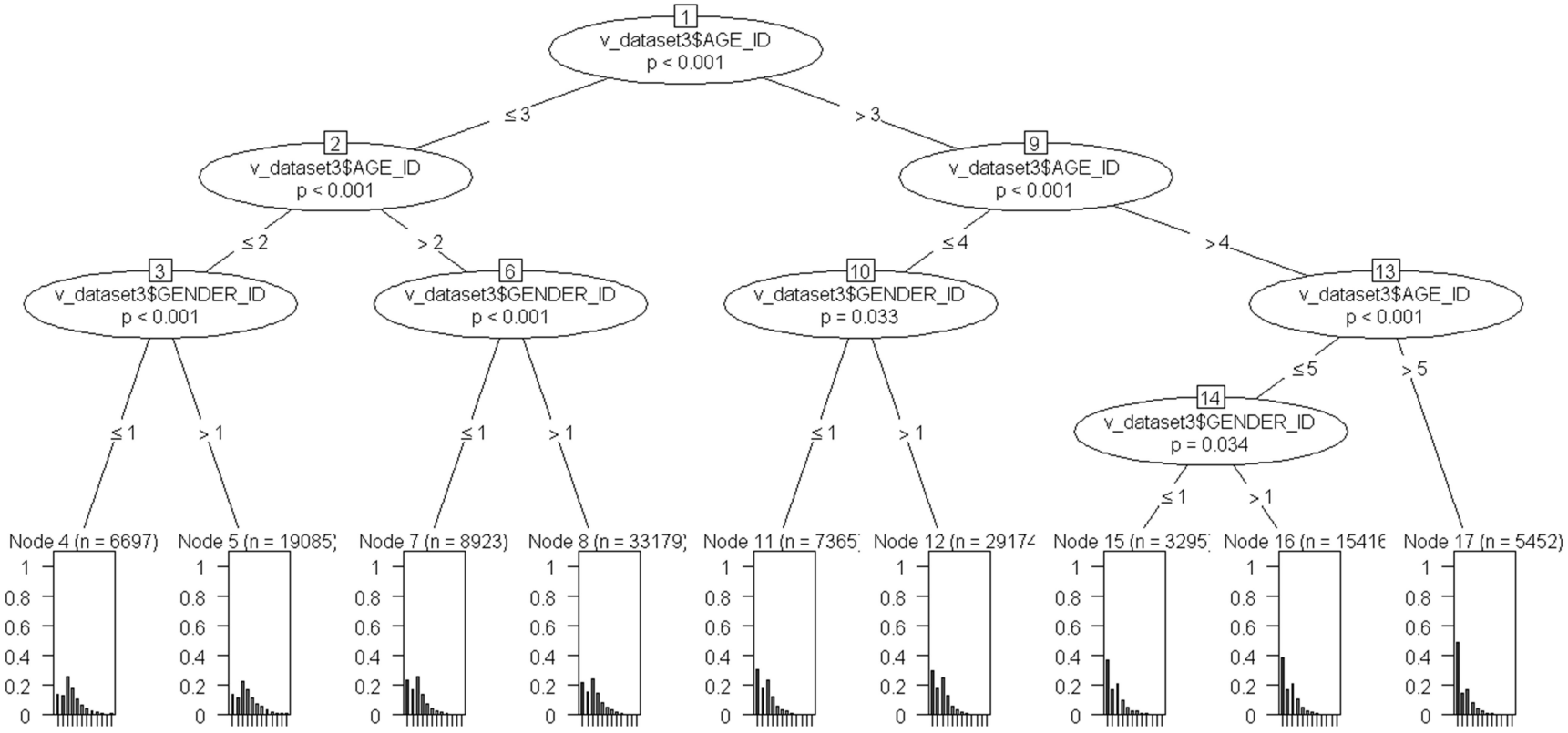
USAGE_GB	USAGE_GB_GROUP	USAGE_GB_ID	GENDER	GENDERID	CITY	CITY_ID	REGION_NAME	REGION_ID	AGE_ID	AGE	LNG	LAT
11,7	10-12	12	K	1	Bursa	16	Marmara Bölgesi	6	2	20-30	29,06687	40,18257
2,3	2-4	3	K	1	Konya	42	İç Anadolu Bölgesi	5	5	50-60	32,4833333	37,8666667
1,1	1-2	2	E	2	Edirne	22	Marmara Bölgesi	6	4	40-50	26,5666667	41,6666667
4,4	4-6	5	E	2	Kocaeli	41	Marmara Bölgesi	6	3	30-40	29,8815203	40,8532704
3,1	2-4	4	E	2	Samsun	55	Karadeniz Bölgesi	7	2	20-30	36,33128	41,292782
1,7	1-2	2	E	2	Afyonkarahisar	3	Ege Bölgesi	3	6	60+	30,54034	38,76376
6,8	6-8	7	E	2	Çankırı	18	İç Anadolu Bölgesi	5	3	30-40	33,6166667	40,6
4,8	4-6	5	K	1	Bilecik	11	Marmara Bölgesi	6	3	30-40	29,983061	40,150131
7,3	6-8	8	E	2	Kahramanmaraş	46	Akdeniz Bölgesi	1	3	30-40	36,9333333	37,5833333
8,1	8-10	9	E	2	Kastamonu	37	Karadeniz Bölgesi	7	2	20-30	33,78273	41,38871
8,4	8-10	9	E	2	Kütahya	43	Ege Bölgesi	3	4	40-50	29,9833333	39,4166667
4,4	4-6	5	E	2	İzmir	35	Ege Bölgesi	3	4	40-50	27,12872	38,41885
3,7	2-4	4	E	2	İstanbul	34	Marmara Bölgesi	6	6	60+	28,97696	41,00527
2,2	2-4	3	E	2	Afyonkarahisar	3	Ege Bölgesi	3	3	30-40	30,54034	38,76376

4.2. Model Oluşturma

Bu çalışmada oluşturulan sınıflandırma, kümeleme ve regresyon modelleri R Studio 1.0.153 programı yardımıyla hazırlanmıştır. Ayrıca verideki değişkenlik, benzerlik eğilimleri, verilerin hangi alanlarda kümelendiği, ayrıştığı veya nasıl bir eğilim izlediği hakkında fikir sahibi olmak için R Studio'daki grafik ve haritalar gibi veri görselleştirme araçları kullanılmıştır. Çalışmada Yeni Zelanda'daki Aucland Üniversitesinde Ross Ihaka ve Robert Gentleman tarafından yazılan R dili kullanılmıştır. Daha sonra dünyanın çeşitli yerlerindeki araştırmacılar R'yi geliştirmek için bir araya gelmiş ve R Geliştirme Takımı adlı bir ekip oluşturmuşlardır. R programlama dili, Unix, Windows ve Mac OS platformlarında çalıştırılabilen içerisinde birçok makine öğrenmesi algoritması barındıran bir veri madenciliği ve istatistiksel hesaplama programıdır. R'nin asıl alanı istatistiki hesaplama ve grafik gösterimi olmasına rağmen, bünyesinde çeşitli sınıflandırma, kümele ve regresyon analizi algoritmalarını da barındırmaktadır.

4.2.1. Sınıflandırma yönteminde karar ağaçları modelinin oluşturulması

Karar ağaçları, sınıflandırma yönteminde kullanılan diğer algoritmalara göre kolay anlaşılmasından dolayı yaygın olarak kullanılan modeldir. Bu modelde eldeki veriler oluşturulan ağaca uygulanır. Karar ağaçları modeli oluşturulurken elimizdeki veri setine uygun olan ve R de yaygın olarak kullanılan koşullu çıkarım ağaç (ctree) fonksiyonu kullanılmıştır. R'de kullandığımız ctree fonksiyonunda formül ve veri seti parametrelerini kullanarak karar ağacı oluşturulmuştur. Formül kısmında karar değişkeni, gruplanmış veri kullanımı değerine karşılık gelen sayısallaştırılmış değişken (data_usage_gb_group_id) olarak seçilmiştir. Modelin uygulanması sonucu oluşan karar ağacı Şekil 4.1'de gösterilmiştir.



Şekil 4.1. Karar ağaçları modeli

Karar ağacı modeli incelendikten sonra her bir nodun ayrı ayrı hata payları Şekil 4.2'de gösterilmiştir.

```
[1] root
|  [2] AGE_ID <= 3
|  |  [3] AGE_ID <= 2
|  |  |  [4] GENDER_ID <= 1: 2 (n = 6697, err = 74.2%)
|  |  |  [5] GENDER_ID > 1: 2 (n = 19085, err = 77.5%)
|  |  [6] AGE_ID > 2
|  |  |  [7] GENDER_ID <= 1: 2 (n = 8923, err = 73.9%)
|  |  |  [8] GENDER_ID > 1: 2 (n = 33179, err = 75.3%)
|  [9] AGE_ID > 3
|  |  [10] AGE_ID <= 4
|  |  |  [11] GENDER_ID <= 1: 0 (n = 7365, err = 69.3%)
|  |  |  [12] GENDER_ID > 1: 0 (n = 29174, err = 70.4%)
|  |  [13] AGE_ID > 4
|  |  |  [14] AGE_ID <= 5
|  |  |  |  [15] GENDER_ID <= 1: 0 (n = 3295, err = 62.8%)
|  |  |  |  [16] GENDER_ID > 1: 0 (n = 15416, err = 61.7%)
|  |  |  [17] AGE_ID > 5: 0 (n = 5452, err = 51.0%)
```

Şekil 4.2. Karar ağaçları modeli sonuçları

Bu modelin sonuçlarının gösterildiği Şekil 4.2'nin incelenmesinde hata oranlarının %50 - %77 aralığında olduğu gözlemlenmiştir. Buradaki hata oranlarından kastedilen değerler ise nodun sonucu oluşan sınıfın içerdiği mobil veri kullanım miktarının dışında kalan diğer mobil veri kullanım aralıklarıdır. Tüm nodları aşağıda açıklayacak olursak:

NOD4 : Hata oranı %74,2'dir. Kurala uyan 6697 verinin %25,8'inin sonucu 2, yani mobil veri kullanımı 2-4 GB aralığında; %74,2'sinin sonucu ise 2 dışındaki tüm değerler, yani 2-4 GB dışındaki tüm değer aralıklarıdır.

NOD5 : Hata oranı %77,5'dir. Kurala uyan 19085 verinin %22,5'inin sonucu 2, yani mobil veri kullanımı 2-4 GB aralığında; %77,5'inin sonucu ise 2 dışındaki tüm değerler, yani 2-4 GB dışındaki tüm değer aralıklarıdır.

NOD7 : Hata oranı %73,9'dur. Kurala uyan 8923 verinin %26,1'inin sonucu 2, yani mobil veri kullanımı 2-4 GB aralığında; %73,9'unun sonucu ise 2 dışındaki tüm değerler, yani 2-4 GB dışındaki tüm değer aralıklarıdır.

NOD8 : Hata oranı %75,3'dür. Kurala uyan 33179 verinin %24,7'sinin sonucu 2, yani mobil veri kullanımı 2-4 GB aralığında; %75,3'ünün sonucu ise 2 dışındaki tüm değerler, yani 2-4 GB dışındaki tüm değer aralıklarıdır.

NOD11 : Hata oranı %69,3'dür. Kurala uyan 7365 verinin %30,7'sinin sonucu 0, yani mobil veri kullanımı 0-1 GB aralığında; %69,3'ünün sonucu ise 0 dışındaki tüm değerler, yani 0-1 GB dışındaki tüm değer aralıklarıdır.

NOD12 : Hata oranı %70,4'dür. Kurala uyan 29174 verinin %29,6'sının sonucu 0, yani mobil veri kullanımı 0-1 GB aralığında; %70,4'ünün sonucu ise 0 dışındaki tüm değerler, yani 0-1 GB dışındaki tüm değer aralıklarıdır.

NOD15 : Hata oranı %62,8'dir. Kurala uyan 3295 verinin %37,2'inin sonucu 0, yani mobil veri kullanımı 0-1 GB aralığında; %62,8'inin sonucu ise 0 dışındaki tüm değerler, yani 0-1 GB dışındaki tüm değer aralıklarıdır.

NOD16 : Hata oranı %61,7'dir. Kurala uyan 15416 verinin %38,3'ünün sonucu 0, yani mobil veri kullanımı 0-1 GB aralığında; %61,7'sinin sonucu ise 0 dışındaki tüm değerler, yani 0-1 GB dışındaki tüm değer aralıklarıdır.

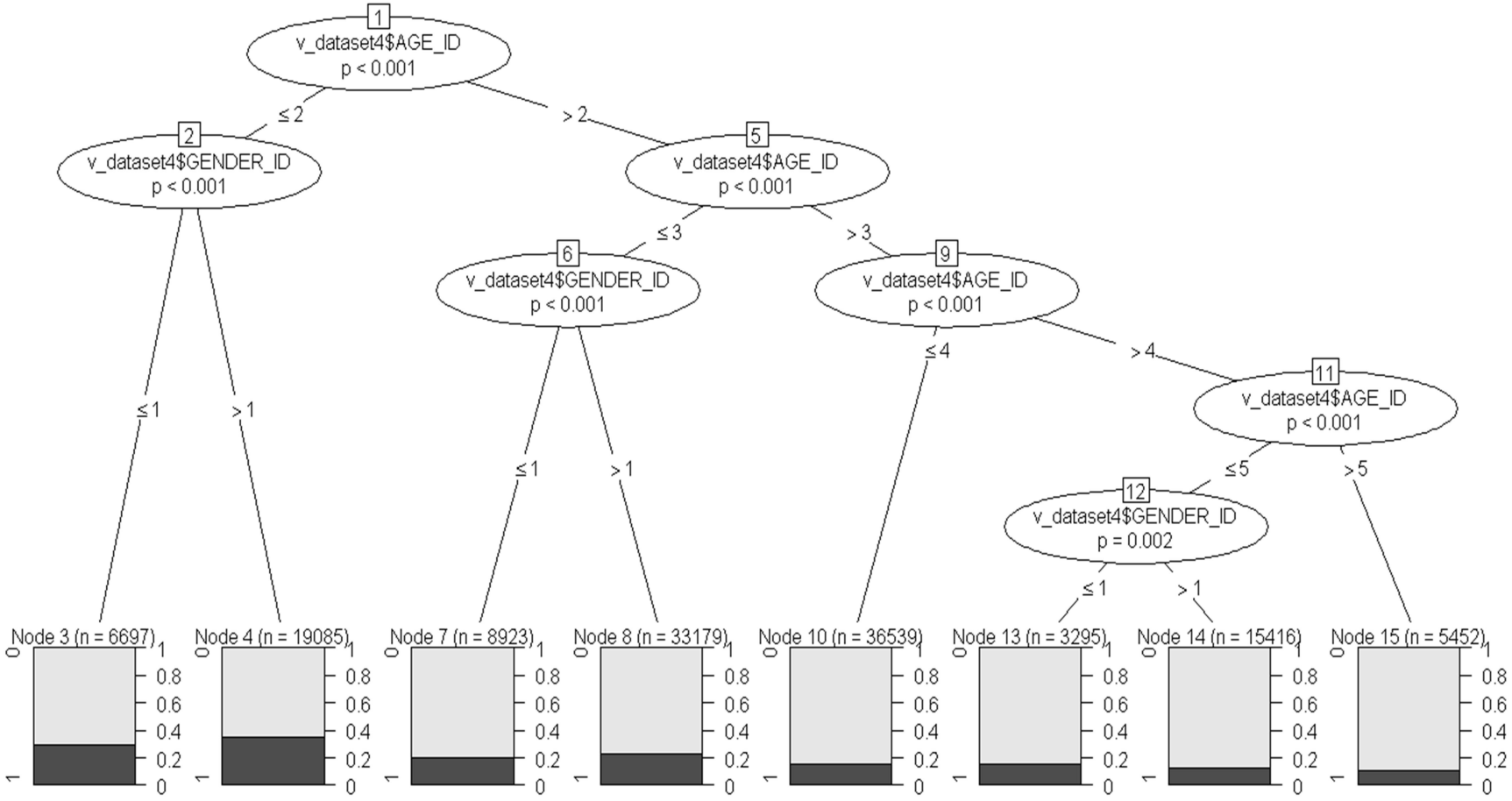
NOD17 : Hata oranı %51'dir. Kurala uyan 5452 verinin %49'unun sonucu 0, yani mobil veri kullanımı 0-1 GB aralığında; %51'inin sonucu ise 0 dışındaki tüm değerler, yani 0-1 GB dışındaki tüm değer aralıklarıdır.

Karar ağacı modelinin sonuçlarının gösterildiği Şekil 4.2'de hata oranları çok yüksek çıkmaktadır. Bu hatanın oluşmasına sebep olan karar değişkeninin değişmesi gerekmektedir. Bunun için veri setinde karar değişkeninin sahip olduğu sonuç değerlerinin daha geniş kapsamlı olması düşünülmüştür. Veri setinde karar değişkeni olarak kullanılan verinin yerine, aşağıdaki özelliklere uygun şekilde daha genel bir gruplama verisi oluşturulmuş ve yeni oluşturulan bu verinin karar değişkeni olarak kullanılmıştır.

Yeni eklenen karar değişkeni özellikleri:

- 0-6 GB arası veri kullanımı: 0: Normal veri kullanımı.
- 6 GB ve daha yüksek veri kullanımı: 1: Yüksek veri kullanımı.

Veri setinde veri kullanım miktarları incelendiğinde ortalama veri kullanım miktarı 6 GB gibi bir değer ortaya çıkmaktadır. Bu değer temel alınarak 6 GB ve alt değerlerine ait veri kullanım miktarı normal, 6 GB üzeri veri kullanımı ise yüksek veri kullanım miktarı olarak gruplanmıştır. Bu özelliklere sahip değişkeni usage_data_decision alan adıyla sadece bu modelde kullanılmak üzere veri setine dahil edilmiştir. Yeni değişkenin bulunduğu veri seti tekrardan R programına yüklenerek karar ağacı modeli oluşturulmuştur. Tekrardan oluşturulan karar ağacı Şekil 4.3'de gösterilmiştir.



Şekil 4.3. Karar ağaçları 2. modeli

Oluşturulan karar ağacı incelendiğinde bir önceki karar ağacına kıyasla çok daha az hata oranı ile karşılaşılmıştır. Node bazında incelendiğinde en sağdaki nodun hata payının en az olduğu gözlemlenmiştir. Node15'in hata payı en az, Node4'ün ise hata payı en çok olan değer olarak göze çarpmaktadır.

Karar ağacı oluşturulduktan sonra ortaya çıkan kurallar:

Eğer Birey yaşı <=30 ise
Ve Cinsiyet = Kız ise NOD3
Cinsiyet = Erkek ise NOD4
Eğer Birey yaşı >30 ise
Ve birey yaşı <= 40 ise
Ve Cinsiyet Kız ise NOD7
Cinsiyet=Erkek ise NOD8
Ve Bireyin yaşı >40 ise
Ve Bireyin yaşı <= 50 ise NOD10
Ve Bireyin yaşı >50 ise
Ve bireyin yaşı 60 <= ise
Ve cinsiyet=Kız ise NOD13
Cinsiyet=Erkek ise NOD14
Ve bireyin yaşı>60 ise NOD15

Nod'ların incelenmesinde ise:

NOD3 : Hata oranı %28,9 ile 0'dır (normal veri kullanımı). Yani NOD3 kuralına uyan bireylerin %71,1'i normal veri kullanımına %28,9'u da yüksek veri kullanımına sahiptir.

NOD4 : Hata oranı %34,7 ile 0'dır (normal veri kullanımı). Yani NOD4 kuralına uyan bireylerin %65,3'ü normal veri kullanımına %34,7'si de yüksek veri kullanımına sahiptir.

NOD7 : Hata oranı %19,6 ile 0'dır (normal veri kullanımı). Yani NOD7 kuralına uyan bireylerin %80,4'ü normal veri kullanımına %19,6'sı da yüksek veri kullanımına sahiptir.

NOD8 : Hata oranı %22,4 ile 0'dır (normal veri kullanımı). Yani NOD8 kuralına uyan bireylerin %77,6'sı normal veri kullanımına %22,4'ü de yüksek veri kullanımına sahiptir.

NOD10 : Hata oranı %15 ile 0'dır (normal veri kullanımı). Yani NOD10 kuralına uyan bireylerin %85'i normal veri kullanımına %15'i de yüksek veri kullanımına sahiptir.

NOD13 : Hata oranı %14,9 ile 0'dır (normal veri kullanımı). Yani NOD13 kuralına uyan bireylerin %85,1'i normal veri kullanımına %14,9'u da yüksek veri kullanımına sahiptir.

NOD14 : Hata oranı %12,7 ile 0'dır (normal veri kullanımı). Yani NOD14 kuralına uyan bireylerin %87,3'ü normal veri kullanımına %12,7'si de yüksek veri kullanımına sahiptir.

NOD15 : Hata oranı %10,3 ile 0'dır (normal veri kullanımı). Yani NOD15 kuralına uyan bireylerin %89,7'si normal veri kullanımına %10,3'ü de yüksek veri kullanımına sahiptir.

Karar ağacının incelenmesi sonucunda her bir nodun ayrı ayrı kuralları ve hata oranları Şekil 4.4'de gösterilmiştir.

```
[1] root
|  [2] AGE_ID <= 2
|  |  [3] GENDER_ID <= 1: 0 (n = 6697, err = 28.9%)
|  |  [4] GENDER_ID > 1: 0 (n = 19085, err = 34.7%)
|  [5] AGE_ID > 2
|  |  [6] AGE_ID <= 3
|  |  |  [7] GENDER_ID <= 1: 0 (n = 8923, err = 19.6%)
|  |  |  [8] GENDER_ID > 1: 0 (n = 33179, err = 22.4%)
|  |  [9] AGE_ID > 3
|  |  |  [10] AGE_ID <= 4: 0 (n = 36539, err = 15.0%)
|  |  |  [11] AGE_ID > 4
|  |  |  |  [12] AGE_ID <= 5
|  |  |  |  |  [13] GENDER_ID <= 1: 0 (n = 3295, err = 14.9%)
|  |  |  |  |  [14] GENDER_ID > 1: 0 (n = 15416, err = 12.7%)
|  |  |  |  [15] AGE_ID > 5: 0 (n = 5452, err = 10.3%)
```

Şekil 4.4. Karar ağaçları 2. modeli sonuçları

Yukarıda gösterilen karar ağacı algoritmasıyla yapılan çalışma neticesinde elde edilen sonuçlara göre, hata oranı %15'den küçük nodlar temel alınarak, Türkiye'de 60 yaş üzeri bireylerin %89,7'si, 50 ve 60 yaş arası erkek bireylerin %87,3'ü, 50 ve 60 yaş arası bayan bireylerin %85,1'i, 40 ve 50 yaş arası bireylerin %85'i 0-6 GB aralığında mobil veri kullandığı ortaya çıkarılmıştır.

4.2.2. Regresyon yönteminde çoklu regresyon modelinin oluşturulması

Çoklu Regresyon modeli oluşturulurken veri setimizde bağımlı değişken olarak mobil veri kullanımı değişkeni kullanıldı. Bağımsız değişkenler olarak da bölge, yaş, cinsiyet ve şehir bilgileri kullanıldı. R'de doğrusal model kestirimi için lm(Formül, Veri) fonksiyonundan yararlanıldı. Fonksiyonun, model oluşturulurken kullanılan

parametrelerinden Formül, uygun olduğu düşünülen modelin sembolik gösterimidir. Veri ise modeldeki değişkenleri içeren listedir.

Regresyon modellerinde anlamlılık düzeyi, 0'a yakın değer anlamlı olarak, 1'e yakın değer ise kötü sonuç olarak kabul görmektedir. Bu yakınlık araştırmacı tarafından belirlenen eşik değerle ölçülür. Örneğin, 0.05 değeri eşik kabul edilerek bu değerin üstündeki değerler kötü sonuç, altındaki değerler anlamlı olarak nitelendirilir. Bu çalışmada eşik değer 0.05 olarak kabul edildi. Mobil veri kullanımının bağımlı değişken, bölge, yaş, cinsiyet ve şehir bilgilerinin ise bağımsız değişken olduğu veri setimizi girdi olarak kullanıp çoklu regresyon modelimizi oluşturuyoruz. Oluşturulan bu modelin sonuçları Tablo 4.9'da gösterilmiştir.

Tablo 4.9. Çoklu regresyon modeli sonuçları

	Estimate	**Prediction(>|t|)**	**Sign. codes**
(Intercept)	6.7018341	< 2e-16	***
REGION_ID	0.0104980	0.060454	.
AGE_ID	-0.7573595	< 2e-16	***
GENDER_ID	0.1507664	5.08e-09	***
CITY_ID	-0.0019096	0.000165	***
Signif. codes: 0 '***' | 0.001 '**' | 0.01 '*' | 0.05 '.' | 0.1 ' ' | 1			
Residual standard error: 3.765 on 128581 degrees of freedom			

Regresyon modelinde her bir bağımsız değişkenin Tablo 4.9'da gösterilen anlamlılık düzeyleri incelendiğinde,

- Yaş bilgisini veren (AGE_ID) değişkenin anlamlılık değeri 2e-16'dan daha küçüktür, yani 0 değerine çok yakındır ve 2e-16<0.05 olduğundan anlamlıdır.
- Cinsiyet bilgisini veren (GENDER_ID) değişkenin anlamlılık değeri 5.08e-09'dur, yani 0 değerine çok yakındır ve 5.08e-09 < 0.05 olduğundan anlamlıdır.

- Şehir bilgisini veren (CITY_ID) değişkenin anlamlılık değeri 0,000165'dir ve 0,000165 <0.05 olduğundan bu değişkende anlamlıdır.
- Bölge bilgisini veren (REGION_ID) değişkenin anlamlılık değeri 0,060454'tür ve 0,060454 > 0.05 olduğunda bu bağımsız değişken kötü sonuç vermektedir.

Çoklu Regresyon modelinde en uygun modeli bulmak için bir kaç yöntem vardır. Bunlardan en yaygın olarak kullanılanı geriye doğru eleme yöntemiyle en uygun modelin bulunmasıdır. Geri doğru eleme yönteminde ilk adım, anlamlılık düzeyi için kötü sonuç olarak belirlenen değişkenler içerisinde en kötü anlamlılık düzeyine sahip değişken, modelden atılır. Bu değişken olmadan geri kalan değişkenlerle model tekrar çalıştırılır. Elde edilen yeni model tekrar incelenir. Kötü anlamlı değişken yine varsa bu durum tekrar edilir. Şayet tüm değişkenler anlamlı olunca model tamamlanmıştır.

Geriye doğru eleme yöntemini uygularken elimizde anlamlılık düzeyi kötü olan tek değişken bölge bilgisidir. En uygun modeli bulabilmek için bölge bilgisi değişkeni veri setinden çıkarılmıştır. Mobil veri kullanımının bağımlı değişken, yaş, cinsiyet ve şehir bilgilerinin ise bağımsız değişken olduğu veri setimizi girdi olarak kullanılıp çoklu regresyon modelimi tekrar oluşturulmuştur. Elde edilen yeni sonuçlar Tablo 4.10'da gösterilmiştir.

Tablo 4.10. Çoklu regresyon 2. modeli sonuçları

	Estimate	**Prediction(>|t|)**	**Sign. codes**
(Intercept)	6.7466737	< 2e-16	***
AGE_ID	-0.7567465	< 2e-16	***
GENDER_ID	0.1481422	8.86e-09	***
CITY_ID	-0.0017262	0.00052	***
Signif. codes: 0 '***' | 0.001 '**' | 0.01 '*' | 0.05 '.' | 0.1 ' ' | 1			
Residual standard error: 3.765 on 128581 degrees of freedom			

Elde edilen yeni regresyon modeli incelendiğinde yaş bilgisini veren (AGE_ID) değişkenin anlamlılık değeri 2e-16'dan daha küçük yani 0 değerine yakındır. Bu değer 2e-16<0.05 olduğundan anlamlıdır. Cinsiyet bilgisini veren (GENDER_ID) değişkeninin anlamlılık değeri 5.08e-09'dur yani 0 değerine daha yakındır. Bu değer 5.08e-09<0.05 olduğundan anlamlıdır. Şehir bilgisini veren (CITY_ID) değişkeninin anlamlılık değeri 0,00052'dir yani 0 değerine daha çok yakındır. Bu değer 0,00052<0.05 olduğundan bu değişkende anlamlıdır. Burada bir önceki regresyon modelimize göre cinsiyet ve şehir değişkenimizin anlamlılık değerinde küçük değişiklik olmuştur. Bu da bölge bilgisinin değişken listesinden çıkarılmasının sonucudur. Bu şekilde çoklu regresyon modeli tamamlanmıştır.

Sonuç olarak çoklu regresyon modelinin elimizdeki veri setine uygulanmasındaki amacımız, denklem 5.1'deki matematiksel formüle ulaşarak bağımsız değişkenler arasında bu formül kullanılıp bağımlı değişkenin tahmin edilmesini sağlamaktır. Oluşturulan anlamlı modelde Tablo 4.10'da görüldüğü gibi bağımsız değişkenlerin katsayı bilgileri oluşturulmuştur. Buna göre:

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n \qquad (5.1)$$

Y : Mobil veri kullanımı

X_1 : Yaş bilgisi (AGE_ID)

X_2 : Cinsiyet bilgisi (GENDER_ID)

X_3 : Şehir bilgisi (CITY_ID)

b_0 : 6.7466737

b_1 (X_1 in katsayısı): -0.7567465

b_2 (X_2 nin katsayısı): 0.1481422

b_3 (X_3 ün katsayısı): -0.0017262

Elde edilen formül;

$$Y = (6.7466737) + (-0.7567465)X_1 + (0.1481422)X_2 + (-0.0017262)X_3$$

Örneğin, 32 yaşında, erkek ve İstanbul'da yaşayan bireyin mobil veri kullanım bilgisi bu formül yardımıyla hesaplanırsa;

Yaş verisinin değeri: 32 = 30-40= 3
Cinsiyet verisinin değeri: Erkek = 2
Şehir verisinin değeri: İstanbul = 34

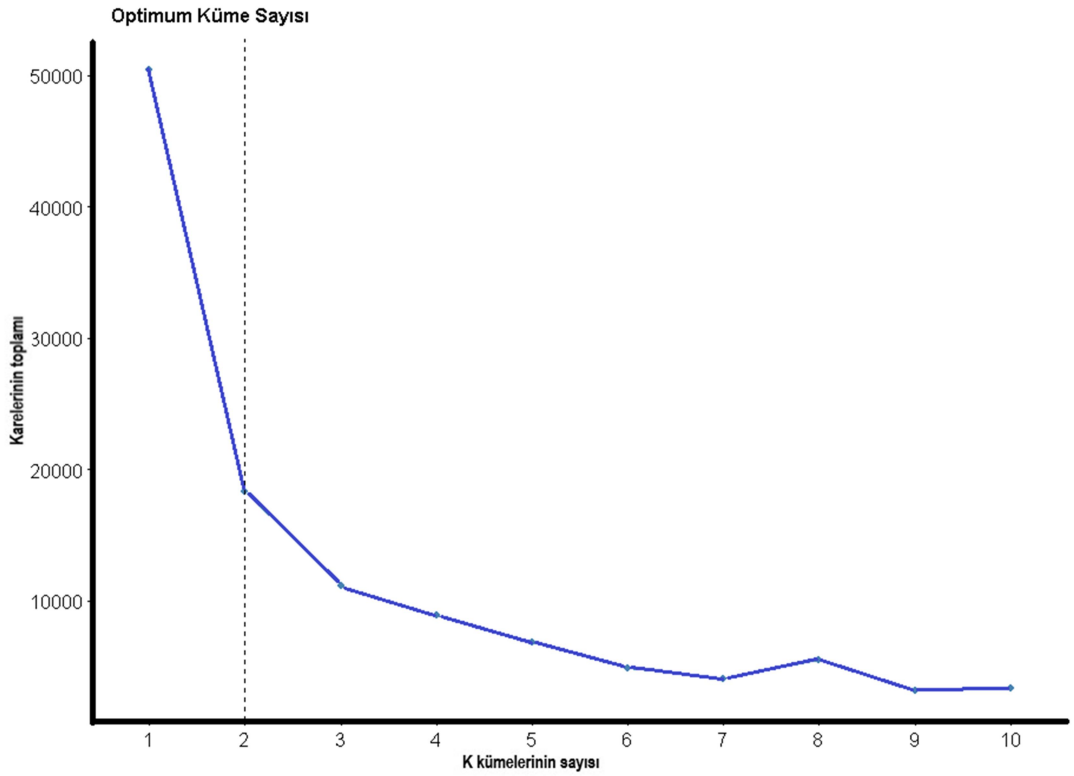
Bu değerler elde edilen formülde yerlerine yazılıp aşağıdaki sonuç elde edilmiştir.

Y = (6.7466737) + (-0.7567465) x 3 + (0.1481422) x 2 + (-0.0017262) x 34
Y = 5,4707749 = 5.5 GB mobil veri kullanımına sahiptir.

Bu formülden elde edilen bir diğer çıkarım ise tüm değişkenler sabit kalmak koşuluyla cinsiyet değişkenini belirten X_2 verisinin katsayısı (0.1481422) pozitif bir değer ve X_2 verisinin alabileceği değerler 2: Erkek, 1: Bayan olduğundan erkeklerin bayanlardan daha çok mobil veri kullandığı ortaya çıkarılmıştır.

4.2.3. Kümeleme yönteminde k-means modelinin oluşturulması

K-means kümeleme modeli oluşturulurken öncelikle k değerinin belirlenmesi gerekmektedir. R'de optimal k sayısını belirleyen fonksiyonlar mevcuttur ve varsayılan maksimum k sayısı 10'dur. Bu fonksiyon yardımıyla veri setimize uygun k değeri hesaplanıp sonucu Şekil 4.5'te gösterilmiştir.



Şekil 4.5. Uygun k değeri grafiği

Yukarıdaki Şekil 4.5'te en büyük kırılmanın k=2'de olduğunu gözlemlemekteyiz. Dolayısıyla k=2'den sonra çok büyük değişimler olmadığından optimal k değerimizi 2 olarak belirliyoruz. K=3 ve k=4'te de kırılmalar oluşmuştur fakat k=1 ile k=2 arasındaki değişim kadar olmamıştır. 4'ten sonraki k değerleri içinde değişim çok azdır.

R'de k-means kümeleme algoritması kmeans fonksiyonuyla gerçekleştirilir. Kmeans fonksiyonunun dataset, k ve nstart olmak üzere aktif olarak kullanılan 3 parametresi vardır;

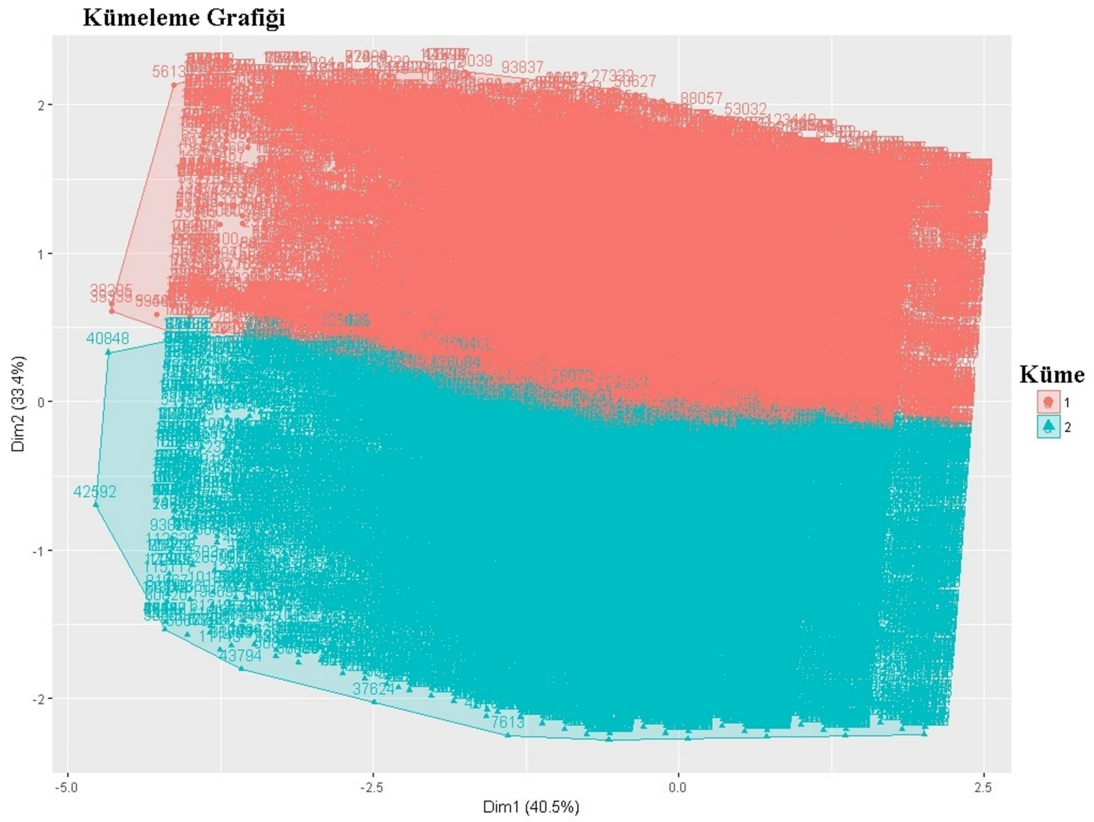
Dataset: Veri seti bilgisini matrix formatında alır.

K: Alt küme sayısıdır. Çalışmada k=2 değeri verildi.

Nstart: Başlangıçta bölünmesi istediğimiz küme sayısı. Daha iyi sonuç için çalışmada 100 değeri verildi.

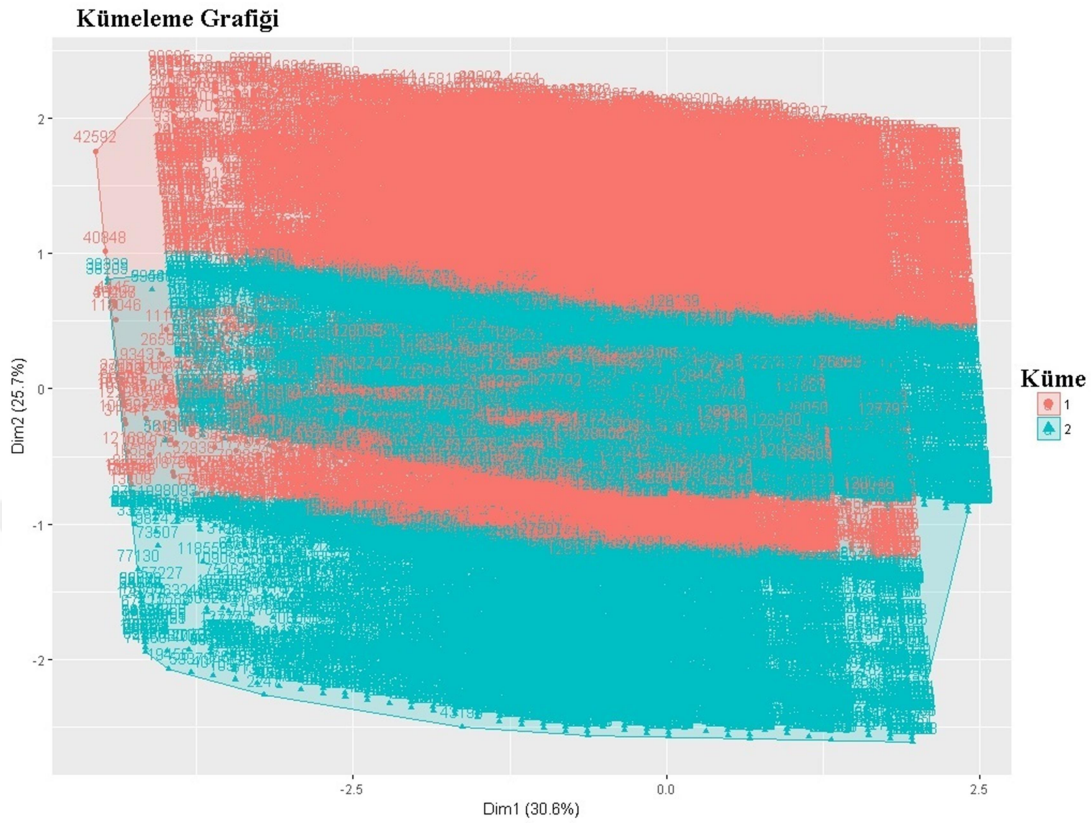
K-means algoritmasının sonuçları baştaki rastgele merkez seçimlerinden çok etkilendiğinden dolayı nstart değerinin 1 den büyük olması gerekmektedir. Nstart>1 seçildiği zaman birden fazla başlangıç konfigürasyonunu dener ve en iyi konfigürasyonla ilgili raporlar üretir. Örneğin, nstart = 25'in eklenmesi, 25 başlangıç rastgele merkezi oluşturur ve algoritma için en iyi olanı seçer.

K-means algoritması veri setimizdeki farklı değişkenlerin bir arada kullanılmasıyla birkaç kez denenmiştir. Veri setindeki şehir, yaş ve veri kullanım değişkenleri kullanılarak oluşturulan k-means kümeleme modeli Şekil 4.6'da gösterilmiştir.



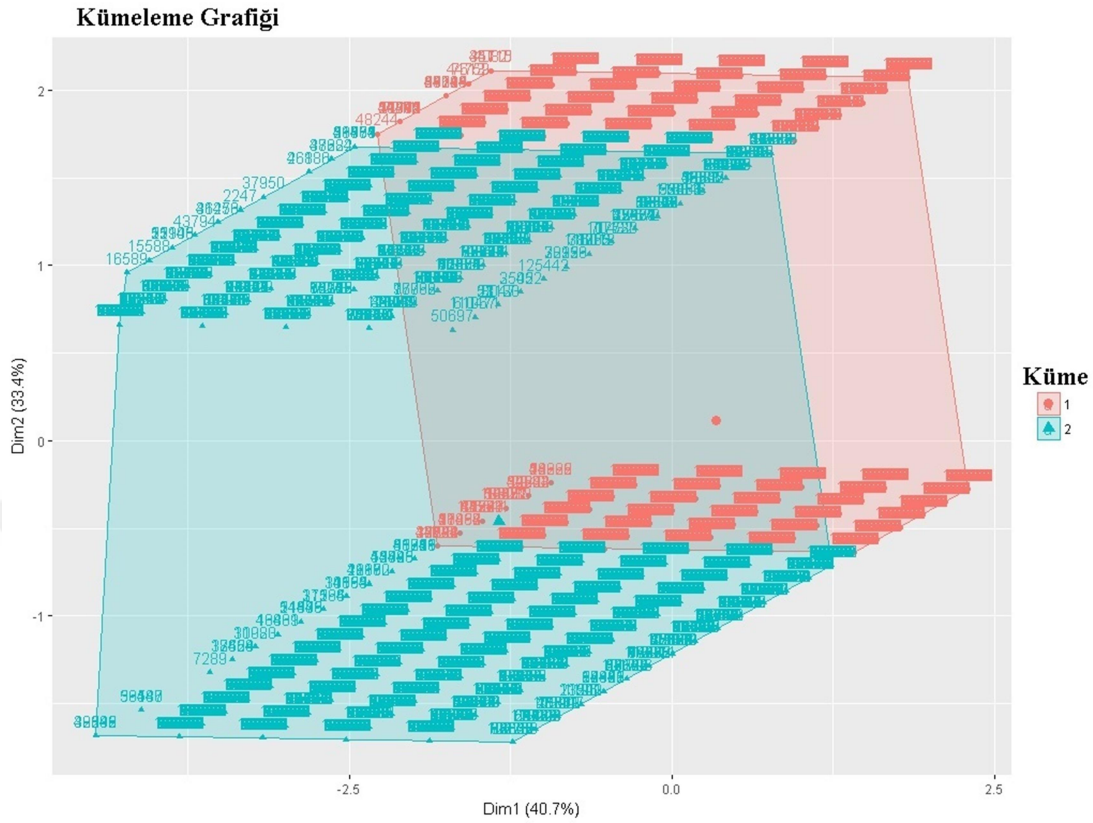
Şekil 4.6. Şehir, yaş ve veri değişkenleriyle k-means modeli

Veri setinde bulunan şehir, cinsiyet, yaş ve veri kullanım değişkenleri kullanılarak oluşturulan k-means kümeleme modelinin sonucu Şekil 4.7'de gösterilmiştir.



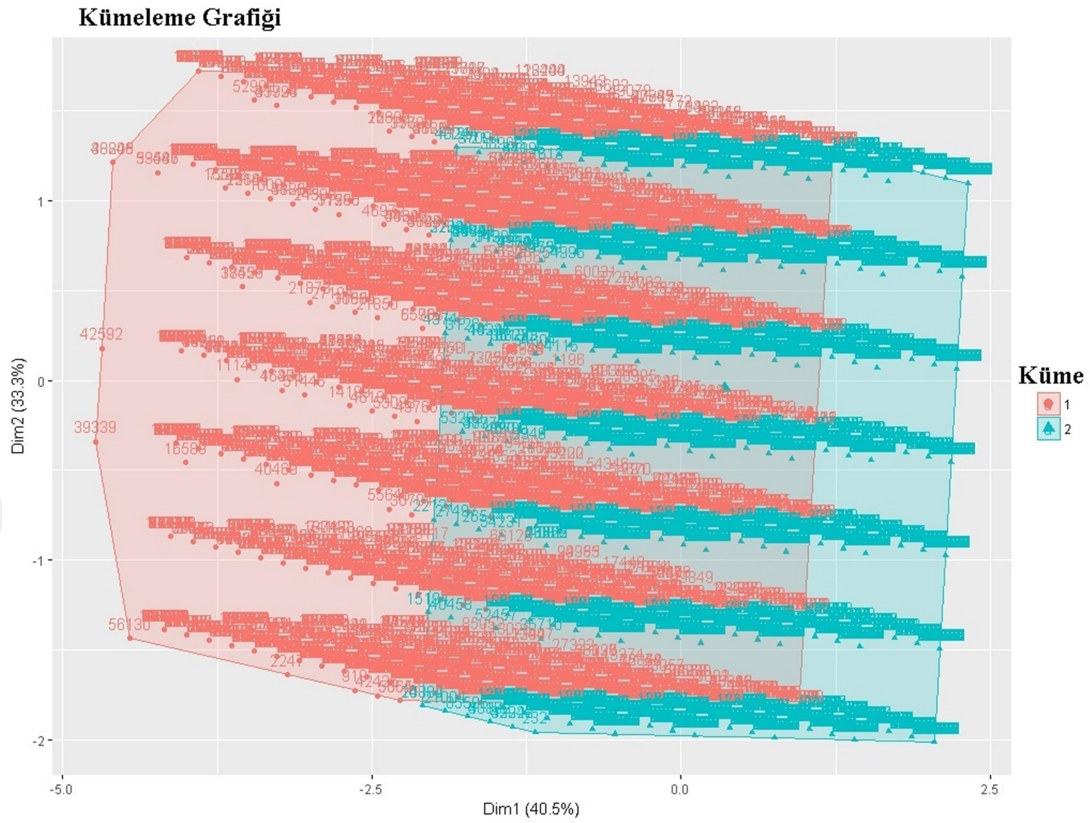
Şekil 4.7. Şehir, cinsiyet, yaş ve veri değişkenleriyle k-means modeli

Veri setinde bulunan cinsiyet bilgisi, yaş bilgisi ve veri kullanım bilgilerini içeren değişkenler kullanılarak oluşturulan k-means kümeleme modelinin sonucu Şekil 4.8'de gösterilmiştir.



Şekil 4.8. Cinsiyet, yaş ve veri değişkenleriyle k-means modeli

Veri setinde bulunan bölge bilgisi, yaş aralığı ve veri kullanım bilgisini içeren değişkenler kullanılarak oluşturulan k-means kümeleme modelinin sonucu Şekil 4.9'da gösterilmiştir.



Şekil 4.9. Bölge, yaş ve veri değişkenleriyle k-means modeli

Yukarıdaki k-means kümeleme modelinin uygulanması sonucu oluşan görsel grafikler incelendiğinde Şekil 4.6'nın nispeten Şekil 4.7, Şekil 4.8 ve Şekil 4.9'a göre daha ayrık kümelenebilir veriler olduğu gözlemlenmiştir. Kümele yöntemiyle elde edilen en uygun sonuç şehir, yaş ve veri değişkenleriyle oluşturulan k-means modelidir.

Sonuç olarak K-means kümele yöntemini elimizdeki veri setimize uyguladığımızda cinsiyet bilgisi, yaş bilgisi ve veri kullanım bilgisini içeren değişkenlerin kendi aralarında ayrıca kümeleme yoluyla da ayrılabilmekte veya veri tabanında da bu şekilde bölümlenerek saklanabilmektedir.

4.2.4. Veri görselleştirilme metodu

Veri görselleştirme, verilerin grafik ve haritalar gibi araçlarla sunularak en iyi şekilde anlaşılabileceği bir yöntemdir. Bu şekilde verideki değişkenlikler, benzerlikler, ayrışmalar ve verinin nasıl bir eğilim izlediği daha kolay görülebilir. Ayrıca veri kümesinde bulunan ortalamadan sapan farklı özelliklere sahip verileri de tanımlamak mümkündür.

Çalışmamızda elimizdeki veri setinin daha okunabilir ve anlaşabilir olabilmesi için veri görselleştirme metodu kullanılmıştır. Bu metodu R uygulamalarıyla beraber Google haritaları üzerinde yapılması mümkün olan görsel çalışmalar ele alınmıştır. Modelin gerçekleşmesi için R'de ggmap ve mapproj kütüphaneleri kullanılarak Şekil 4.10 elde edilmiştir.

Şekil 4.10. Şehirlere göre veri kullanımı gösterimi

Şekil 4.10'da elde edilen çalışma incelendiğinde elimizdeki veri setinde bulunan her bir şehir değişkeni için kullanılan veri miktarı ortaya çıkarılmıştır. Bu haritada hangi şehirlerde ortalama ne kadar veri kullanıldığı tahmin edilebilmektedir. Çalışmayı daha da iyileştirmek adına her bir şehrin nüfusunun, ülke nüfusuna etki eden ağırlık katsayısını bulmak için şehirlerin nüfus sayısı ile Türkiye'nin toplam nüfus sayısı Türkiye İstatistik Kurumundan (TÜİK) elde edilmiştir. Daha sonra her bir şehrin nüfus sayısı ülke nüfusuna bölünerek şehirlerin ağırlık katsayısı elde edilmiştir. Türkiye'de şehirlerin ülke nüfusuna etki eden ağırlık katsayıları Tablo 4.11'de gösterilmiştir.

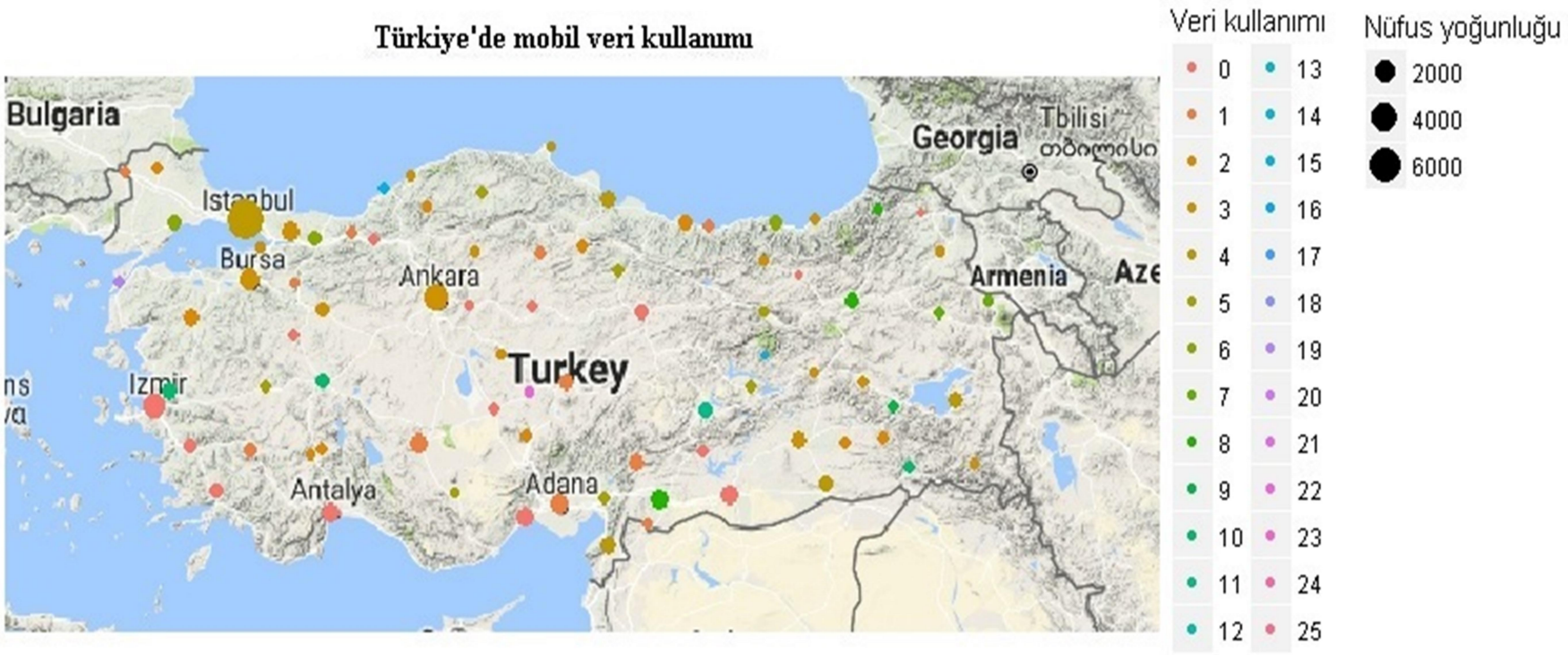
Tablo 4.11. Şehirlere göre nüfus ağırlık katsayısı

CITY_ID	NÜFUS AĞIRLIK KATSAYISI
34	0,185480673144231
6	0,0669864892721558
35	0,0529167678539504
16	0,0363515716262951
7	0,0291744504604913
1	0,0275847091201839
42	0,0270789512395503
27	0,0247352902443456
...	...

Bu ağırlık katsayısına göre eldeki veri setinden her bir şehrin katsayı oranı kadar rastgele seçimle yeni bir veri kümesi elde edilmiştir. Bu veri kümesinde her şehre ait veriler ağırlık katsayısına göre hesaplanmıştır. Hesaplama sonucunda ise her şehir nüfus yoğunluğu oranında veri içererek toplamda 40 bin civarında veri içeren yeni bir veri kümesi elde edilmiştir. Veri kümesi, veri görselleştirme modelimizde kullanılarak Şekil 4.11 elde edilmiştir.

Şekil 4.11. Şehir nüfusu katsayısına göre veri kullanımı gösterimi

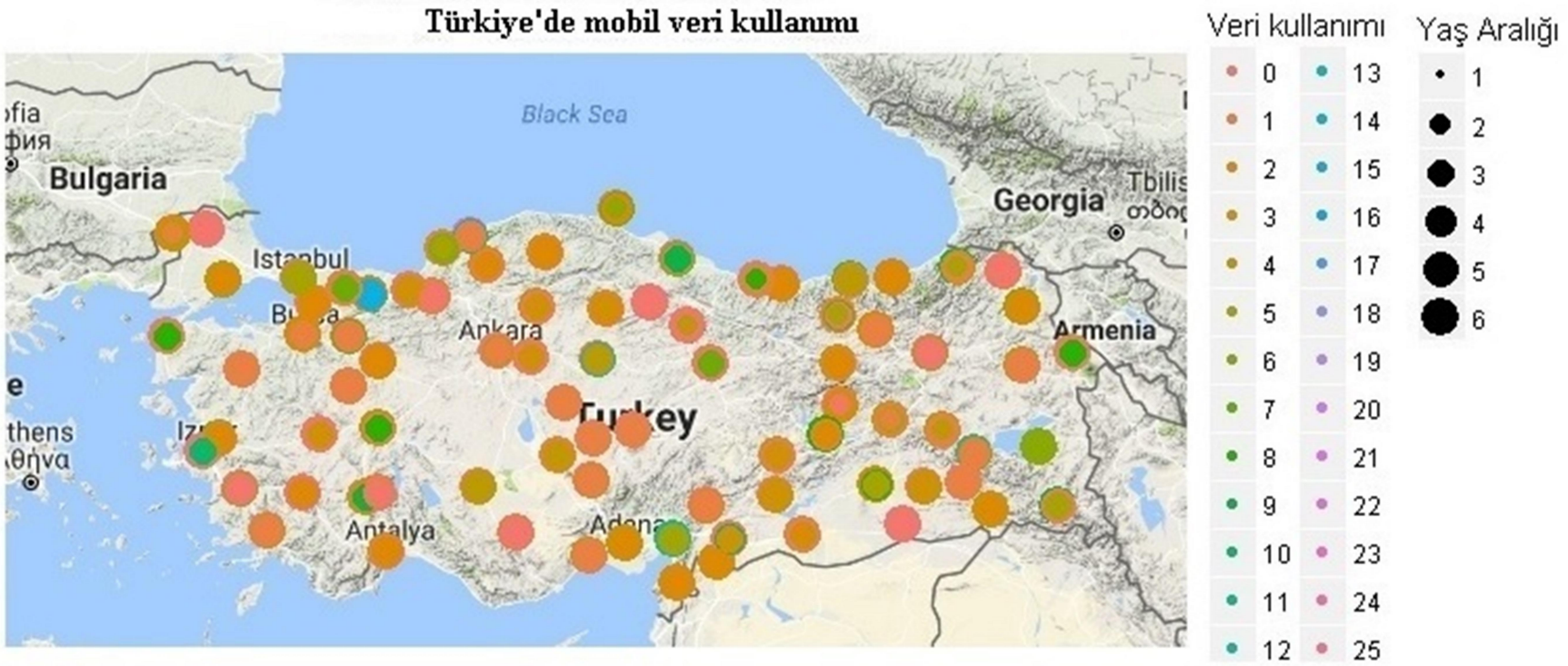
Şekil 4.11'deki harita incelendiğinde, veri kullanım miktarı arttıkça kırmızı dairesel işaretlerin çapı da artmaktadır. Veri kullanan birey sayısı arttıkça da kırmızı renklerin bulanık görünümden daha net görünüme kavuştuğu izlenilmiştir. Burada dairelerin küçük ve bulanık olduğu bölgelerde veri kullanımı miktarı ve veri kullanan birey sayısının az olduğu, dairelerin küçük fakat net olduğu bölgelerde ise veri kullanımı miktarı az fakat veri kullanan birey sayısının fazla olduğu ortaya çıkmaktadır. Dairelerin büyük ve net olduğu bölgelerde veri kullanımı miktarı ve veri kullanan birey sayısının büyük olduğu, dairelerin büyük fakat bulanık olduğu bölgelerde ise veri kullanımı miktarı büyük fakat veri kullanan birey sayısının az olduğu ortaya çıkmaktadır. Aynı veri seti kullanılarak R'de farklı grafiksel fonksiyonlar ile veri görselleştirme metodu uygulanarak nüfus yoğunluğuna bağlı veri kullanım miktarını gösteren Şekil 4.12'deki harita elde edilmiştir.



Şekil 4.12. Şehir nüfusu katsayısına göre detaylı veri kullanımı gösterimi

Şekil 4.12'de ise her bir veri kullanım miktarı farklı renklerde dairesel sembollerle gösterilmiştir. Veri kullanan birey sayısı ise dairesel sembolün çapıyla doğru orantılı olacak şekilde gösterilmiştir. Daha yoğun nüfusa sahip şehirlerde veri kullanım yoğunluğu dairenin büyüklüğüyle gözlemlenmektedir. Örneğin, İstanbul'un üzerindeki dairenin büyüklüğüne bakılarak en yoğun mobil veri kullanımına sahip bölge olarak ortaya çıkmakta ve dairesel sembolün rengine bakılarak yaklaşık 4 GB veri kullanım miktarı olduğu ortaya çıkmaktadır.

Aşağıdaki Şekil 4.13'de ise Türkiye'deki şehirlerde bireyin yaş aralığına bağlı mobil veri kullanımı gösterilmiştir. Burada veri kullanım miktarı farklı renklerdeki dairesel sembollerle gösterilirken, bireyin yaş aralığı dairesel sembolün boyutları ile gösterilmektedir. Şekil 4.13'de görüldüğü üzere çoğu şehirlerde veri kullanım miktarına ait birden çok renk hâkimdir. Dairenin en dış çeperindeki renk, yaş aralığı en büyük bireylerin kullandığı mobil veri iken, dairenin merkezine doğru gidildikçe azalan yaş aralığının kullandığı veri miktarı gösterilmiştir.



Şekil 4.13. Şehirlerde bireylerin yaş aralıklarına göre veri kullanımı gösterimi

5. SONUÇ

Bu çalışmada veri madenciliği, Türkiye'de mobil veri kullanımının bireylerin belirli özelliklerine göre (bulunduğu bölge, yaş grubu, cinsiyet) gruplandırılarak, insan gözlemiyle belirlenmesi kolay olmayan örüntü ve kuralları ortaya çıkarmak için kullanılmıştır.

Yapılan sınıflandırma, regresyon ve kümeleme analizleri sonunda analiz öncesi eldeki veri setine bakılarak çeşitli sonuçlara varılmıştır. Sınıflandırma yöntemi kullanılarak Türkiye'de bireylerin yaş grupları ile kullanılan veri miktarı arasındaki ilişki bulunmuştur. Hangi yaş grubunun yaklaşık ne kadar veri miktarı harcayacağı ortaya konulmuştur. Regresyon yönteminde ise bireylerin yaş bilgisi, şehir bilgisi ve cinsiyet bilgisi verileriyle tahmini olarak kullandığı mobil veriyi ortaya çıkaran matematiksel formül bulunmuştur. Bir başka kullandığımız yöntem olan kümeleme yönteminde yaş, cinsiyet, şehir ve mobil veri kullanım bilgilerinin saklandığı veri tabanlarını yönetmek çok güç hale gelmiştir. Bu verileri bir şekilde bölerek saklayabilmek, erişim için çok büyük kolaylık sağlayacaktır. Birbiriyle bağımsız gibi görünen değişkenler, kümeleme yöntemiyle ortaya çıkarılan kurallar ile bölünebilir ve ayrı ayrı saklanabilir hale getirilmiştir. Veri görselleştirme yöntemi kullanılarak verilerin en iyi şekilde anlaşılabileceği ve yorumlanabileceği görsel haritalar ortaya çıkarılmıştır. Bu haritalar ışığında bölgesel tabanlı veri kullanım miktarları ve veri miktar yoğunluğu ortaya çıkarılmıştır.

Gelişen teknolojiler ile artık insanlar veri kullanımı noktasında hatırı sayılır bir çoğunlukta mobil hatları kullanmaktadır. İnternet ve bazı iletişim programlarının yoğun kullanıldığı çağımızda artık mobil iletişim araçları vazgeçilmez olmuştur. Mobil veri kullanımı bu sayede daha da artmaktadır. Artan internet kullanımları mobil teknoloji sunan şirketler için veri iletişiminde gecikme ve kaliteli iletişim sorunlarını da beraberinde getirmiştir. Bu veri iletişimini sağlayan veri miktarlarının sürekli artışı, firmaları daha kaliteli ve daha hızlı veri alışverişi için altyapısal iyileştirmelere itmektedir. Ayrıca altyapısı için hazırlıklar yapılan 5G teknolojisinde

ise internetin yanı sıra veri iletişiminin de tamamen veriler ile sağlandığını düşünürsek mobil verinin ne kadar önemli derecede kullanılacağı ve veri miktarını artıracağı ortaya çıkmaktadır. Bu hizmeti sunan firmalar arasındaki rekabet her geçen gün artmaktadır. Altyapısal çalışmaların devam ettiği 5G teknolojisinin ağır maliyetinden dolayı bireylerin hangi bölgelerde hangi yaş aralığında ne kadar veri kullandığı bilgilerinin ortaya çıkmasıyla yatırım yapılacak bölgeler ve şehirler sıralamasına bu çalışma bir ön ışık olarak da kullanılabilir. Ayrıca bu konuda daha farklı çalışmalara da kaynak olması hedeflenmektedir.

KAYNAKLAR

[1] Savaş S., Topaloğlu N., Yılmaz M., "Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri", İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, Volume:11, Number:21, 2012.

[2] Rygielski C., Wang J. Y., Yen D. C., "Data mining Techniques For Customer Relationship Management", Technology in Society, Volume:24, Number:4, 2002.

[3] Taşkın Ç., Emel G., "Veri Madenciliğinde Kümele Yaklaşımları ve Kohonen Ağları ile Perakendecilik Sektöründe Bir Uygulama", Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, Volume:15, Number:3, 2010.

[4] Han J., Kamber M., "Data mining: Concepts and techniques", Volume:2, 2001.

[5] Patil S., Patil V., Bhat P., "A Review on 5G Technology", International Journal of Engineering and Innovative Technology, Volume:1, Number:1, 2012.

[6] http://btk.gov.tr/Files/45g-BROSUR.pdf, (19.08.2017).

[7] https://wearesocial.com/special-reports/digital-in-2017-global-overview, (29.07.2017).

[8] https://wearesocial.com/uk/special-reports/2017-digital-yearbook, (25.07.2017).

[9] Sarıman G., " Veri Madenciliğinde Kümeleme Teknikleri Üzerine Bir Çalışma: K-Means ve K-Medoids Kümeleme Algoritmalarının Karşılaştırılması", Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, Volume:15, Number:3, 2011.

[10] Özekes S., "Veri madenciliği modelleri ve uygulama alanları", İstanbul Ticaret Üniversitesi Dergisi, Volume:3, 2003.

[11] Ayre L. B., "Data Mining for Information Professionals", 2006.

[12] Berry M., Linoff G., "Mastering Data Mining: The Art and Science of Customer Relationship Management", John Wiley ve Sons, 1999.

[13] Albayrak A., Koltan Yılmaz Ş., "Veri Madenciliğinde Karar Ağacı Algoritmaları ve İMKB Verileri Üzerine Bir Uygulama", Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, Volume:14, Number:1, 2009.

[14] Ayık Y. Z., Özdemir A., Yavuz U., "Lise Türü ve Lise Mezuniyet Başarısının, Kazanılan Fakülte ile İlişkisinin Veri Madenciliği Tekniği ile Analizi", Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, Volume:10, Number:2, 2007.

[15] https://www.onlineistatistik.com/single-post/2017/01/01/Coklu-Dogrusal-Regresyon-Analizi-Nedir-Amaclari-ve-Varsayimlari-Nelerdir, (15.08. 2017).

[16] https://acikders.ankara.edu.tr/pluginfile.php/233/mod_resource/content/4/11-Coklu%20Regresyon.pdf, (01.08.2017).

[17] Feldman R., Sanger J., "The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data", Cambridge Universty Press, Volume:1, 2006.

ÖZGEÇMİŞ

Muhammet Ali Altınışık, 1986 Erzurum doğumludur. 2005 yılında Harran Üniversitesi Bilgisayar Mühendisliği bölümünü kazanmıştır. 2009 yılında bu bölümü tamamladıktan sonra askerlik görevini yapmaya başlamıştır. 2010 yılında askerlik görevini tamamladıktan sonra yazılım geliştirme uzmanı olarak özel bir şirkette işe başlamıştır. 2015 yılında Maltepe Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği bölümünde yüksek lisans eğitimine başlamıştır. Yazılım sektöründe çeşitli firmalarda farklı pozisyonlarda görev yapmıştır ve halen özel bir firmada kıdemli yazılım danışmanı olarak çalışmaya devam etmektedir.