T.C.
DÜZCE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TELEKOMÜNİKASYON SEKTÖRÜNDE MÜŞTERİ KAYIP ANALİZİ

MUHAMMET SİNAN BAŞARSLAN

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMAN
YRD. DOÇ. DR. Fatih KAYAALP

DÜZCE, 2017

T.C.
DÜZCE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TELEKOMÜNİKASYON SEKTÖRÜNDE MÜŞTERİ KAYIP ANALİZİ

Muhammet Sinan BAŞARSLAN tarafından hazırlanan tez çalışması aşağıdaki jüri tarafından Düzce Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Yrd. Doç. Dr. Fatih KAYAALP

Düzce Üniversitesi

Jüri Üyeleri

Yrd. Doç. Dr. Fatih KAYAALP
Düzce Üniversitesi _____

Yrd. Doç. Dr. Esra ŞATIR
Düzce Üniversitesi _____

Doç. Dr. Kemal POLAT
Abant İzzet Baysal Üniversitesi _____

Tez Savunma Tarihi: 20/06/2017

BEYAN

Bu tez çalışmasının kendi çalışmam olduğunu, tezin planlanmasından yazımına kadar bütün aşamalarda etik dışı davranışımın olmadığını, bu tezdeki bütün bilgileri akademik ve etik kurallar içinde elde ettiğimi, bu tez çalışmasıyla elde edilmeyen bütün bilgi ve yorumlara kaynak gösterdiğimi ve bu kaynakları da kaynaklar listesine aldığımı, yine bu tezin çalışılması ve yazımı sırasında patent ve telif haklarını ihlal edici bir davranışımın olmadığını beyan ederim.

20 Haziran 2017

Muhammet Sinan Başarslan

TEŞEKKÜR

Yüksek lisans eğitimim ve tez süresince gösterdiği her türlü destek ve yardımdan dolayı çok değerli hocam Yrd. Doç. Dr. Fatih KAYAALP'e en içten dileklerimle teşekkür ederim.

Tez çalışmam boyunca değerli katkılarından dolayı Bilgi İşlem Dairesi Başkanı, Doç. Dr. Resul KARA'ya da şükranlarımı sunarım.

Tez çalışmam boyunca her zaman yanımda olup beni destekleyen annem ve kız kardeşime teşekkürlerimi sunarım. Rahmetli dedem, babaannem ve babama saygıyla.

20 Haziran 2017 **Muhammet Sinan Başarslan**

İÇİNDEKİLER

ŞEKİL LİSTESİ

Sayfa No

TABLO LİSTESİ

Sayfa No

KISALTMALAR

ACC	Doğruluk
CART	Classification and regression tree
CRISP	Veri madenciliği üzerine çapraz endüstri standart süreç modeli
CRM	Müşteri ilişkileri yönetimi
dn	Doğru negatif
DOR	Tanısal üstünlük oranı
dp	Doğru pozitif
ERR	Hata oranı
E/K	Erkek/kadın
E,H	E (müşteri ayrıldı), H (müşteri ayrılmadı)
EM	Expectation-maximization(maksimum beklenti)
F-measure	F-ölçü
FNR	Yanlış negatif oranı
FPR	Yanlış pozitif oranı
GSM	Global system for mobile
GKA	Grafiksel kullanıcı arayüzü
GNU	Genel kamu lisansı
GUI	Graphical user ınterface
IBM	International business machines
IDE	Integrated development environment
JDBC	Java database connectivity
KNIME	Konstanz ınformation miner
k-nn	k-en yakın komşu
LEM2	Learning from examples module, version 2
LR-	Negatif olabilirlik oranı
LR+	Pozitif olabilirlik oranı
LP	Lojistik regresyon
MLP-ANN	Multilayer perceptron artificial neural networks
MLP	Multilayer perceptron
N/A	Kayıp veri
NB	Navie bayes
neg	Gerçekte negatiflerin toplamı
NPV	Negatif öngörü değeri
poz	Gerçekte pozitiflerin toplamı
PPV	Pozitif öngörü değeri
ROC	Receiver operating characteristic
SAS	Statistical analysis system
SOM	Self organizing map
SPC	Specificity - Belirleyicilik
SPSS	Statistical package for the social sciences
SVM	Destek vektör makinesi
Sybase	Statistical analysis system
tNeg	Tahmin edilen negatiflerin toplamı
TÜİK	Türkiye istatistik kurumu
TNR	Belirleyicilik
VIM	Visualization and ımputation of missing values

VP	Voted perceptron
yn	Yanlış negatif
yp	Yanlış pozitif
WEKA	Waikato environment for knowledge analysis
.arff	Attribute relationship file format
.txt	Metin dosya uzantısı

SİMGELER

C	Sınıf uzayı
m	Örneklerin sayısı
M	Model
n	Değişkenlerin sayısı
Y	Çıktı uzayı

ÖZET

TELEKOMÜNİKASYON SEKTÖRÜNDE MÜŞTERİ KAYIP ANALİZİ

Muhammet Sinan BAŞARSLAN
Düzce Üniversitesi
Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı
Yüksek Lisans Tezi
Danışman: Yrd. Doç. Dr. Fatih KAYAALP
Haziran 2017, 92 sayfa

İnsanların ihtiyaçlarına göre tüketim tercihleri farklılıklar gösterir. Müşteriye yatırım yapan kurumlar da bu tercihleri öngöremezler. Özellikle müşteri odaklı kurumlar yeni müşteri kazanma ve eldeki müşteriyi memnun ederek müşteri kaybını önlemeye çalışırlar. Müşteri odaklı sektörlerden birisi olan Telekomünikasyon şirketleri de müşteri kazanmak ve mevcut müşterilerini kaybetmemek isterler. İşte bu noktada çeşitli yollar ile müşterilerinin kaybını tahmin etmeye yönelik çalışmalar yaparlar. Bu tez çalışmasında, veri madenciliği ve makine öğrenmesi yöntemlerinden olan sınıflandırma algoritmaları ile müşteri kayıp analizi yapılmıştır. Bu analiz yapılırken makine öğrenmesi süreci adımlarından olan veri madenciliği üzerine çapraz endüstri standart süreç modeli (CRISP) kullanılmıştır. Sınıflandırma algoritmaları ile elde edilen modellerin performansları çapraz geçerleme ve hold-out performans yöntemleri ile değerlendirilmiştir. Çapraz geçerleme katı olarak 4 kat, 5 kat ve 10 kat çapraz geçerleme kullanılmıştır. 4 kat, 5 kat ve 10 kat çapraz geçerleme ile performans değerlendirmesinde karar ağaçları algoritmaları ile kurulan modeller, diğer modellere göre daha iyi bir performans göstermiştir. En iyi performansı gösteren C4.5 karar ağacı algoritmasının performansı yaklaşık olarak 0.98'dir. C4.5 karar ağacından sonra sırasıyla ID3 ve gini karar ağaçları, k-en yakın komşu ve bayes algoritmaları ile oluşturan modeller gelmektedir. *k*-en yakın komşu algoritması karar ağaçlarından sonra gelse de performansı C4.5 karar ağacına yakındır. Hold-out yöntemi ile veri seti %60-%40, %75-%25, %80-%20 ayrım oranlarına sahip sırasıyla eğitim ve test veri setine ayrılmıştır. Bu veri setleri üzerinde yapılan performans değerlendirmelerinde ise k-kat çapraz geçerlemedeki gibi benzer sonuç veren C4.5 karar ağacı en iyi performansı göstermiştir. Sonrasında k-kat çapraz geçerleme performans yönteminde yakın değerlere sahip olduğu ID3 ve Gini karar ağaçlarını geçen k-en yakın komşu algoritması olmuştur. En son sırada ise bayes algoritması yer almaktadır. *k*-en yakın komşu algoritmasının ID3 ve Gini karar ağaçlarını geçmesi hold-out ile rastgele ayrımda daha iyi performans göstermesinden dolayıdır. Veri madenciliği programı olarak kullanılan R sayesinde veri görselleştirme üzerine de bir çalışma yapılmıştır. Bu çalışmalara ek olarak sınıflandırma algoritmalarından en iyi sonucu veren C4.5 Karar ağacı algoritması ile oluşturulan model R paketlerinden Shiny ile web uygulaması yapılarak dinamik hale getirilmiştir.

Anahtar sözcükler: Makine öğrenmesi, R, Veri görselleştirme, Veri madenciliği, Web tabanlı veri madenciliği uygulaması.

ABSTRACT

CUSTOMER CHURN ANALYSIS IN TELECOMUNICATION INDUSTRY

Muhammet Sinan BAŞARSLAN
Düzce University
Graduate School of Natural and Applied Sciences, Department of Computer Engineering
Master's Thesis
Supervisor: Assist. Prof. Dr. Fatih KAYAALP
June 2017, 92 pages

Consumption preferences of people vary depending on their needs. And, institutions investing in clients cannot predict these preferences. Especially, customer-oriented institutions try to gain new customers and prevent customer churn by satisfying existing customers. Telecommunications industry is one of the customer-oriented industries. Telecommunication companies also want to gain customers, without losing existing customers. At this point, they engage in prediction of customer churn using various methods. In this thesis study, customer churn analysis was performed with classification algorithms, which are among the data mining and machine learning methods. In carrying out this analysis, the Cross Industry Standard Process for Data Mining (CRISP) model, which is one of the machine learning process steps, was used. The thesis was explained through the steps of the CRISP model from identification of problem to model selection. The performances of the models obtained by the classification algorithms were evaluated by the cross-validation and hold-out performance methods. The 4-fold, 5-fold and 10-fold cross-validations were used. Models built with decision tree algorithms in performance evaluation with 4-fold, 5-fold and 10-fold cross-validation showed better performance than the other models. The performance of the best performing C4.5 decision tree was approximately 0.98. The C4.5 decision tree was followed by the models created with ID3, Gini decision trees, k-nearest neighbors and Bayes algorithms, respectively. Although the k-nearest neighbor algorithm comes after the decision trees, its performance was closer to that of C4.5 decision tree. In the performance evaluations performed on the training-test dataset with the 60-40%, 75-25% and 80-20% separation ratios with the hold-out method, respectively, the best-performing was the C4.5 decision tree, similar to that of k-fold cross-validation performance. This was followed by ID3 and Gini decision tree and k-nearest neighbor algorithm, with close values as in k-fold cross-validation performance method. The Bayes algorithm had the worst performance. Since the k-nearest neighbor algorithm ID3 and Gini perform better at random distinction with hold-out of decision trees. A study on data visualization has also been carried out through R which is used as a data mining program. In addition to these studies, C4.5, which gives the best result from the classification algorithms, has been rendered dynamic by making web application with Shiny from the R packets generated by the decision tree algorithm.

Keywords: Machine learning, R, Data mining, Data visualization, Web based data mining application.

1. GİRİŞ

Günümüz teknolojileri sayesinde her an bir veri yığınına maruz kalıyoruz ancak tabiri caizse gerçek bilgi için açlık çekiyoruz. Günümüz insanları alış veriş, banka işlemleri, hastane randevu işlemleri, fatura ödeme işlemleri, ilköğretim öğrencisinden üniversite öğrencisine kadar not durum bilgisine bu teknolojiler ile ulaşmaktadırlar. Son zamanlarda yaygın olarak kullanılan e-devlet uygulamaları devlet kurumlarından alınması gereken sabıka kaydı, askerlik durum belgesi gibi her türlü bilgiye de internet üzerinden ulaşılabilmektedir. Çoğu işlemler bilgisayar kullanımı ile yapılmaktadır. Buna ek olarak sürekli elimizin altında olan minimal boyutlardaki akıllı telefon teknolojisiyle beraber bilgiye ulaşmak çok kolay hale gelmiştir. Akıllı telefonlar aracılığı ile oturduğumuz yerden sosyal medyada gezinebilme, anlık yazı, fotoğraf paylaşabilmenin yanı sıra banka işlemleri, fatura ödeme gibi önemli işlemleri de kolayca yapabilir hale geldik. Şekil 1.1'de yer alan TÜİK'in hanelerde bilişim teknolojileri bulunma verilerine göre; 2016 yılı Nisan ayında hanelerin %96,9'unda özellikle akıllı cep telefonu sahipliği görülmektedir. Aynı dönemde evdeki bireylerin %22,9'unda masaüstü bilgisayar, %36,3'ünde dizüstü bilgisayar mevcut iken tablet bilgisayar bulunma oranı %29,6 sahip olduğu görülmektedir [1].

Şekil 1.1. Hanelerde bilişim teknolojileri bulunma oranı.

Bu veriler ışığında teknoloji minimal boyutlarda insanlar arası etkileşimde en üst seviyeye çıkmıştır. Özelikle telefon, tablet gibi minimal boyutta olan ve artık insanların cebine giren cihazlar sayesinde telekomünikasyon firmaları ön plana çıkmış ve telekomünikasyon şirketleri arasında rekabet artmıştır.

Veri miktarı günümüzde sosyal medyanın yaygın olmasıyla beraber anlık veri paylaşımlarıyla sürekli artmaktadır. Sosyal medya dışında finans, telekomünikasyon, kamusal alanlarda yaptığımız işlemler sonucu oluşan bilgilerinin oluşturduğu veri kümelerini kullanmak ticari ve devlet kurumları açısından önem arz eder. Örneğin müşterilerinin her alış verişinin sonunda aldığı tüm ürünleri kaydeden bir süpermarket, müşterileri ile ilgili bilgiler sayesinde bir müddet sonra elinde büyük bir veri seti oluşmaktadır. Bu süper market elindeki veriyi yorumlayarak müşterilerin sıklıkla A ürününü B ürünü ile beraber aldığını görürse;

- ya A ile B ürününü yan yana koyacak ve böylece müşteri A'yı almak için geldiğinde B'yi de kolayca bularak alış verişini tamamlayacaktır.
- ya da A ile B'yi birbirinden uzağa koyacak ve böylece müşteri iki ürünü de buluncaya kadar marketin diğer ürünlerini de görecektir.

Sonuç olarak süpermarket, elindeki çok miktardaki veri arasından kendi satışları için fark yaratacak bir strateji yürütmüş olacaktır. Böyle bir yorumu bir insanın veriye bakarak elde etmesi oldukça zordur. Elde oluşan veri yığınından bilgi elde etmeyi veri madenciliği sağlamaktadır [2].

Örnekleri çoğaltacak olursak;

- Telekomünikasyon şirketinden müşterilerin ayrılma nedenlerine yönelik analizler yapılarak ayrılan ve ayrılma eğiliminde olan müşterileri geri kazanacak kampanyalar düzenlenebilir.
- Bankalar müşterilerinin kredi başvurusuna yanıt verirken müşterilerinin kişisel özellikleri ve davranışlarını değerlendirerek ödenmeyen kredi oranlarında azalma olmasını sağlayabilirler.
- Havayolu şirketleri devamlı müşterilerinin davranışlarını inceler ve müşterilere yönelik fiyatlandırma ile kar oranlarını artırabilirler.
- Hastalıkların teşhisinde, yakın akrabalarından hasta olanlar göz önüne alınarak hastalık riski taşıma ihtimali olan kişiler belirlenebilir.
- Sağlık alanında, insanların kanser riskini öğrenmek için sağlık verileri üzerinde çalışmalar yapılarak kanser hastası olma riski olan hastalar ilgili poliklinik ve servislere yönlendirilebilir.

Günümüzde müşteri odaklı firmalar arasında büyük bir rekabet vardır. Bu rekabet müşteri kazanmanın yanı sıra eldeki müşterileri de tutmaya yöneliktir. Son zamanlarda

özellikle müşteri odaklı firmaların müşteri kayıp analizine daha çok önem vermesinin doğal sonucu olarak bu alanda yapılan çalışmalar artmıştır. Aşağıda özellikle son 5 yılda yapılan çalışmalar derlenerek çalışılan tahmin metotları ve kullanılan veriler hakkında bilgi verilmiştir.

Verbeke ve arkadaşları 2012 yılında müşterilerden maksimum kar elde etmek amacıyla 20 değişken içeren 11 farklı veri seti üzerinde kendi modelleriyle bir çalışma gerçekleştirmişlerdir. Ayrıca çok tercih edilen veri madenciliği ve makine öğrenmesi algoritmaları ile modeller kurmuşlar ve elde edilen sonuçları karşılaştırmışlardır [3].

Kamalraj ve Malathi'nin 2013 yılında bir telekomünikasyon firmasından aldıkları 2835'i operatör değiştirmemiş ve 498'i operatörünü değiştirmiş olan toplam 3333 müşteriden oluşan veri setinde, 10 değişken ile yaptıkları müşteri ayrılma analizi çalışmasında J48 karar ağacı ve C5.0 sınıflandırma tekniğinin performanslarını karşılaştırmışlardır. C5.0 algoritmasının J48 algoritmasına göre daha iyi bir tahmin sonucu verdiği ve daha az bellek kullanarak işlemleri yerine getirdiğini ortaya koymuşlardır [4].

Brandusoiu ve Toderean'ın 2013 yılında bir telekomünikasyon firmasından aldıkları 3333 müşterinin 21 değişkenini içeren görüşme kayıtları veri seti üzerinde yapılan müşteri ayrılma analizi çalışmasında 4 farklı çekirdek fonksiyonu kullanarak Destek Vektör Makineleri (Support Vector Machines) temelli modeller oluşturmuşlar ve bunların performanslarını karşılaştırmışlardır. Bu modellerden polinom çekirdek fonksiyonlu olanın %88,56 ile en iyi sonucu verdiğini belirtmişlerdir [5].

Olle ve Cai tarafından 2014 yılında yapılan çalışmada bir Asya telekomünikasyon firmasına ait 2000 abonenin 23 değişken bulunduran 6 aylık bir veri seti üzerinde hibrit model ile WEKA ortamında bir müşteri ayrılma analiz uygulaması gerçekleştirilmiştir. Model içerisinde sınıflandırma için Lojistik Regresyon (LP) ve tahmin için de Voted Perceptron (VP) yöntemleri kullanılmıştır. Elde edilen sonuçlara göre hibrit modelin tahmin başarısının, tekil yöntemlerle elde edilen sonuçlara göre daha iyi olduğunu belirtmişlerdir [6].

Yabaş tarafından 2014 yılında yapılan çalışmada en son veri madenciliği yöntemleri analiz edilerek, servislerden ayrılmış veya başka bir telekomünikasyon servisini kullanmayı düşünen müşterileri tahmin etmek için bir metot geliştirilmiştir. Bu işlemler için Orange Telekomünikasyon firması tarafından sağlanan 100000 kayıt ve 230

değişken barındıran bir veri seti üzerinde çeşitli sınıflandırma metotları uygulanmış ve bunların başarım sonuçları incelenmiştir [7].

Forhad ve arkadaşları tarafından 2014 yılında yapılan çalışmada 880 telefon numarasına ait 26 aylık görüşme verisinin tutulduğu 6938 kayıt üzerinde kural tabanlı sınıflandırma yöntemi ile bir müşteri ayrılma analizi gerçekleştirmişlerdir ve sonuçlarını sunmuşlardır [8].

Amin ve arkadaşları tarafından 2014 yılında tek ve çok sınıflı modellere ayrıntılı, genetik, kaplama ve LEM2 algoritmaları uygulanarak alınan sonuçların karşılaştırılmasına yönelik bir çalışma gerçekleştirilmiştir. Ayrıca genetik algoritması ile yapılan uygulamanın en başarılı sonucu verdiğini belirtmişlerdir [9].

Kuyzu ve Tufan tarafından 2014 yılında telekomünikasyon sektöründe hizmet veren bir firmanın verilerinden faydalanılarak, telefon, internet ve televizyon gibi ürün gruplarına yönelik tarifeler arasındaki müşteri geçişleri ile bu geçişlere neden olan başlıca faktörler tespit edilmiştir. Bu çalışmanın sonucunda firma için önerilerde bulunulmuştur. Modelleme için bu alanda daha önce yapılan çalışmalarda kullanılmamış olan kesikli seçim modelinden faydalanılmış ve modelleme yapılırken değişken olarak tarife ücreti, müşterilerin gelir düzeyi, hanehalkı sayısı, konut özelliği gibi demografik özellikleri ile konuşma süresi, veri indirme miktarı gibi kullanım bilgilerinden yararlanılmıştır. Bu çalışmanın sonucunda, ürün grupları arası geçişlerde müşterinin gelir düzeyi ve ikamet ettiği konut özelliğinin, telefon tarifeleri arası geçişlerde aylık konuşma süresi ve hane halkı sayısının, internet tarifeleri arası geçişlerde aylık veri indirme miktarı ve gelir düzeyinin, televizyon tarifeleri arası geçişlerdeyse gelir düzeyinin daha belirleyici olduğu görülmüştür. Seçim olasılıkları değerlendirildiğinde ise müşterilerin en rasyonel tercihleri, en az geçiş olacağı düşünülen internet tarifelerinde yapmış olduğu söylenmiştir [10].

Kaur ve Mahajan tarafından 2015 yılında yapılan çalışmada telekomünikasyon sektöründe 21 değişken içeren örnek bir veri setini R programında J48 karar ağacı metodu ile müşteri ayrılmasına yönelik analiz çalışması anlatılmaktadır. Çalışmada ek olarak R programının sunduğu birçok grafik paketiyle müşteri ayrılma tahminine yönelik grafiksel çalışma da gerçekleştirmiştir [11].

Hudaib ve arkadaşları 2015 yılında telekomünikasyon sektöründeki firmalara yönelik olarak yaptıkları müşteri ayrılma analiz çalışması için 3 hibrit model geliştirmişlerdir.

Çalışma kapsamında Ürdün telekomünikasyon firması tarafından sağlanmış 5000 müşterinin 11 değişken bilgisini içeren 3 aylık bir döneme ait veri seti üzerinde hibrit modellerin performansları incelenmiştir. Birinci modelde veri filtreleme için k-means algoritması ve tahmin için de Multilayer Perceptron Artificial Neural Networks (MLP-ANN) yöntemi, ikinci modelde MLP-ANN ile hiyerarşik kümeleme yöntemi ve son model olan üçüncü modelde MLP-ANN ile Self Organizing Map (SOM) yöntemi kullanılmıştır. Hibrit modeller ile C4.5 ve MLP-ANN yöntemlerini tekil olarak kullanan modellerle doğruluk oranı ve müşteri ayrılma tahmin oranları açısından sonuçlarını karşılaştırılmıştır ve hibrit modellerin hepsinin tekli modellere göre daha iyi sonuçlar verdiği görülmüştür. Hibrit modeller arasından da k-means ve MLP-ANN ile yapılan modelin en iyi sonucu verdiği belirtilmiştir [12].

Yıldız tarafından 2015 yılında yapılan çalışmada veri madenciliği ve makine öğrenmesi yöntemlerinden olan sınıflandırma algoritmaları ile modeller kurarak müşteri ayrılma analizi gerçekleştirmiştir. Modellerin çalıştığı zamanı düşürmek için ve başarı oranını artırmak için değişken azaltarak modellerin performanslarını ölçmüştür. Model kurmak için, 5000 abonenin 20 değişken barındıran bir veri seti ile 51306 abonenin 172 değişkeni barındıran 2 farklı veri seti üzerinde çalışma yapmış ve performanslarını kesinlik ve geri çağırma oranı ile ölçmüştür [13].

Backiel ve arkadaşları tarafından 2015 yılında yapılan çalışmada müşterinin bireysel görüşmelerinin incelenmesinin yanısıra müşterinin sosyal ortamındaki kişilerle benzer davranışları gösterme eğilimi olarak tarif edilen benzerlik (homophliy) tabanlı bir müşteri ayrılma analiz uygulaması anlatılmaktadır. Bir mobil operatör şirketinden alınmış 6 aylık süre içinde 1 milyon müşteriye ait 111 değişken bulunduran görüşme bilgileri ve sosyal ağ ortamını tanımlama amaçlı olarak kullanılabilecek hangi müşterinin kimle hangi sıklıkta ne kadar görüştüğü gibi verilerin bulunduğu bir veri seti üzerinde uygulamalar gerçekleştirilmiştir. Bu iki unsurun birarada bulunduğu test sonuçlarının tek başına uygulandığı durumlara göre daha başarılı sonuçlar verdiğini belirtmişlerdir [14].

Dahiya ve Bhatia tarafından 2015 yılında yapılan çalışmada 10 değişken içeren 50 kayıt, 50 değişken içeren 200 kayıt ve 100 değişken içeren 608 kayıttan oluşan 3 veri seti üzerinde müşteri ayrılma analizi yapmak için karar ağaçları ve lojistik regresyon temelli 2 farklı model oluşturup bunları WEKA ortamında gerçekleştirmişlerdir. Elde edilen sonuçlara göre karşılaştırma yapıldığında karar ağaçları ile yapılan modelin daha

iyi bir tahmin sonucu verdiğini belirtmişlerdir [15].

Dalvi ve arkadaşları tarafından 2016 yılında yapılan çalışmada telekomünikasyon sektörü müşterileri olan veri seti üzerinde R programı aracılığıyla veri madenciliği ve makine öğrenmesi tekniklerinden lojistik regresyon ve karar ağaçları temelli müşteri ayrılma analiz modelleri kurulmuştur. Müşterilerin görüşme kayıtlarından elde ettikleri 19 özelliği kullanarak elde ettikleri sonuçların karşılaştırmalarını yapmışlardır. Bu çalışmada karar ağaçları yönteminin daha iyi bir tahmin doğruluğu elde ettiğini belirtmişlerdir [16].

Gordini ve Veglio tarafından 2016 yılında yapılan çalışmada pazarlama stratejilerinin belirlenmesinde müşteri ayrılma analizi üzerinde durulmuştur. Telekomünikasyon alanına yönelik bir çalışma olmasa da yapılan çalışmanın bu sahaya uyarlanabilirlik potansiyeli vardır. İnternet üzerinden çeşitli ürünlerin satışını yapan bir İtalyan şirketten alınmış olan 80000 müşterinin bir yıllık verileri ile 24 değişken üzerinde yapılan çalışmada, AUC parametre seçim tekniği temelli bir Destek Vektör Makine modeli (SVMauc) tasarlanmıştır. Bu çalışmada modelin performansının lojistik regresyon, sinir ağları ve klasik destek vektör makineleri ile karşılaştırıldığı ve başarılı sonuçlar elde edildiği belirtilmiştir [17].

Yihui ve Chiyu tarafından 2016 yılında yapılan çalışmada OOPM ismini verdikleri bir değişken seçim metodu ve FE_RF&T ismini verdikleri özellik çıkarım metodu geliştirdiklerini belirtmişlerdir. Önerdikleri bileşenlerden oluşan modeli China Mobile şirketinden aldıkları 16920 kayıt ve 22 değişken içeren örnek veriler üzerinde uygulayıp elde ettikleri sonuçlara göre değişken seçimi için kullandıkları OOPM'nin Random Forest metoduna göre daha avantajlı olduğu ve FR_RF&T metodunun da PCA metoduna göre daha başarılı olduğunu görmüşlerdir [18].

Branduşoiu ve arkadaşları tarafından 2016 yılında yapılan çalışmada California Üniversitesinden aldıkları 3333 müşteri kaydı ve 21 değişken bilgisi olan veri seti üzerinde sinir ağları, Destek Vektör Makineleri ve Bayes Ağları metotları temelli modeller geliştirerek kontenjans tablosu (confusion matris) değerleri, kazanım oranı (gain measure) ve ROC eğrisi sonuçlarını incelemişlerdir. Bayes ağlarının %99.10, Multi Layer Perceptron (MLP)'nin %99.55 ve SVM'nin de %99.70 doğru tahmin yapabildiğini görmüşlerdir [19].

Oskarsdottir ve arkadaşları tarafından 2016 yılında yapılan çalışmada müşterilerin

çevrelerindeki sık iletişim kurduğu kişiler ile görüşme trafiği kayıtları üzerinden yapılan sınıflandırma tabanlı bir modelde müşteri ayrılma analizi uygulaması anlatılmaktadır. Çalışma içerisinde ortalama 1 milyon kayıt içeren 7 farklı veri seti üzerinde ilişkisel sınıflandırıcılar ile ilgili bilgiler ve karşılaştırma sonuçları verilmektedir [20].

Yu ve arkadaşları tarafından 2016 yılında yapılan çalışmada PBCCP ismini verdikleri parçalı sınıflandırma optimizasyonu temelli geriye yayılımlı sinir ağları esaslı bir algoritma anlatılmaktadır. China Mobile şirketinden aldıkları 7 değişken ve yüzlerce kayıt içeren veriler üzerinde BP, PSO-BP ve PBCCP kullanılarak yapılan testlerin sonuçları karşılaştırılmış ve PBCCP ismi verilen algoritmayla ağırlık ve eşik değerleri optimize edilen BP sinir ağları ile müşteri ayrılma analizinin daha iyi sonuçlar verdiğini belirtmişlerdir [21].

AlOmari ve Hassan tarafından 2016 yılında yapılan çalışmada, segmentasyon ve değişken seçimi işlemleri sonrasında daha önce bir müşteri ayrılma analizinde hiç kullanılmamış olan Kurallar Ailesi'nin 6. Algoritması kullanılarak performans ölçümü anlatılmaktadır [22].

Gürsoy tarafından 2016 yılında Türkiye'de bir telekomünikasyon firmasının müşterileri üzerinde kayıp analizi çalışması gerçekleştirmiştir. Ayrılma ihtimali olan müşteri gruplarını belirlemek için 1000 kayıt ve 24 değişkenden oluşan yaklaşık 4 aylık bir veri seti üzerinde lojistik regresyon ve karar ağaçları ile model oluşturmuş ve bu modellerin performanslarına yönelik karşılaştırma yapmıştır [23].

Yukarıda müşteri ayrılma analizine yönelik çalışmalar incelenmiştir. Genel kanı olan tahmin için sınıflandırma algoritmaları kullanıldığı bu çalışma ile doğrulanmıştır. Bu çalışmalar neticesinde yapılacak çalışmada veri madenciliği tekniklerinden sınıflandırma algoritmaları kullanılmasına karar verilmiştir ve bu çalışmada karar ağaçları, *k*-en yakın komşu ve bayes algoritmaları üzerine çalışma yapılmıştır.

Bu çalışmanın amacı belirli bir dönem içinde bir telekomünikasyon şirketinin müşteri kayıtları üzerinde sınıflandırmaya dayalı veri madenciliği teknikleri kullanılarak müşteri kayıplarını en iyi tahmin eden sınıflandırma modelini belirlemektir.

İkinci bölümde müşteri kayıp analizi, üçüncü bölümde veri madenciliği, dördüncü bölümde makine öğrenmesi ve makine öğrenimi süreçleri, beşinci bölümde veri madenciliği ve makine öğrenmesi ile yapılan çalışmanın aşamalar, altıncı bölümde yapılan çalışma ile kurulan modellere ve bu modellerin performanslarının yer aldığı

sonuçlarla ilgili bilgiler bulunmaktadır.

Tez çalışması kapsamında yapılan diğer iki uygulama da yedinci bölümdeki veri görselleştirme uygulaması ile sekizinci bölümdeki web tabanlı uygulamadır. Tez kapsamında yapılan uygulamaların değerlendirilmesi tartışma ve sonuç bölümünde yer almaktadır.

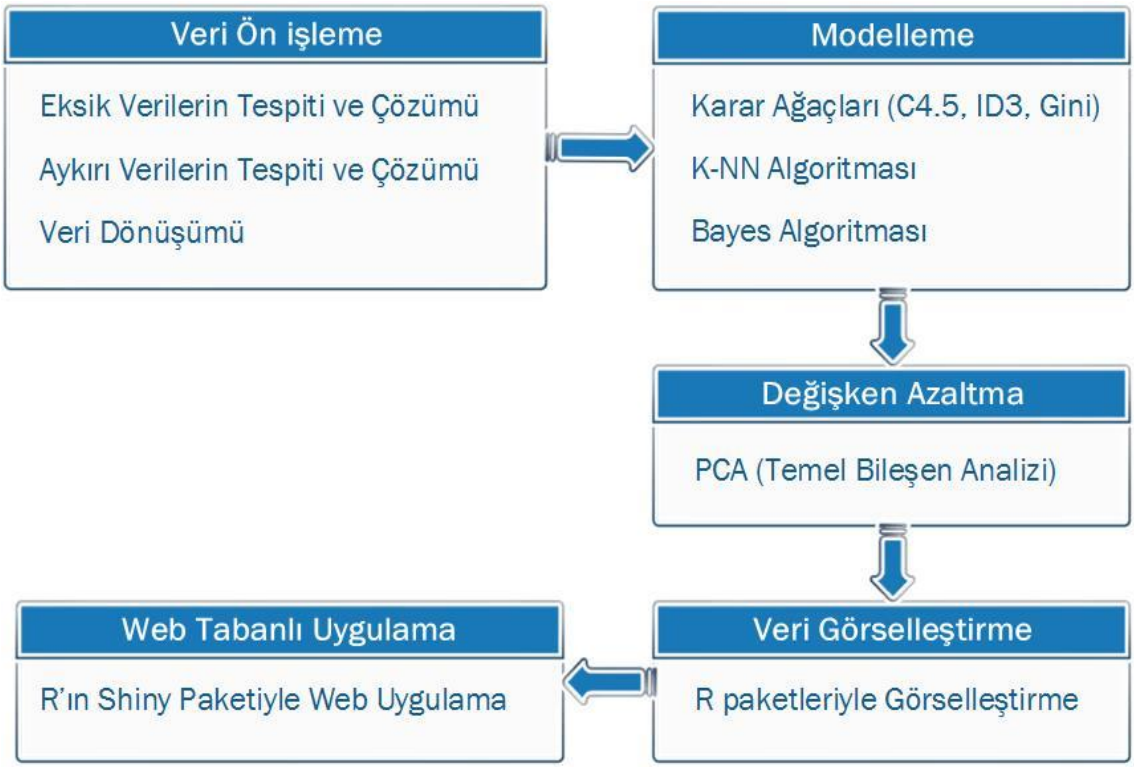
Makine öğrenmesi ve veri madenciliği algoritmaları ile kurulacak olan modellerde veri madenciliği için çarpraz endüstri standart süreci [24] adımları kullanılmıştır. Bu çalışma kapsamında model kurulmasına hazırlık olarak takip edilen CRISP modeli ile model oluşturulurken kullanılan algoritmalar Bölüm 4.1.4'te anlatılmıştır. Makine öğrenmesi sürecinde modellerin hazırlanırken uyulması gereken problem tanımlama, veriyi anlama ve veriyi hazırlama gibi adımları detaylı bir şekilde Bölüm 4.1'de anlatılmıştır. Makine öğrenmesi süreci bölümünde veri seti üzerinde işlem yapıldıktan sonra oluşturulacak modellerin hangi algoritmalar ile yapılacağına ve bu modellerin performanslarının nasıl değerlendirileceğine dair bilgiler Modelleme bölümünde anlatılmaktadır.

Tez çalışmasında sınıflandırma algoritmaları ile müşteri kayıp analizine yönelik en iyi modeli bulma çalışmasına, materyal ve yöntem bölümünde CRISP model adımları takip edilerek açıklanmıştır.

Sınıflandırmaya dayalı veri madenciliği ve makine öğrenmesi algoritmaları ile oluşturulan modeller birbirleriyle karşılaştırılarak en iyi sonucu veren modelin seçildiği Bulgular bölümünde anlatılmaktadır. Ayrıca bu çalışmalara ek olarak bu bölümde kullanılan veri madenciliği programının içerisinde yer alan paketler sayesinde elde edilen grafikler ile veri madenciliği yoluyla veri görselleştirme uygulaması ve web tabanlı müşteri kaybını tahminine yönelik gerçekleştirilen çalışmalar anlatılmıştır.

Son bölüm olan tartışma ve sonuç bölümünde tezin genel bir değerlendirmesi yapılmıştır.

Tez kapsamında yapılan çalışmaların genel gösterimi aşağıda yer alan Şekil 1.2'de gösterilmektedir.



Şekil 1.2. Tez işleyişinin genel gösterimi.

2. MÜŞTERİ KAYIP ANALİZİ

Müşteri odaklı sektörlerde mevcut müşterileri elde tutmak için yapılan analize müşteri kayıp analizi denilmektedir. Müşteri kayıp analizi (customer churn analysis), genellikle Telekomünikasyon, bankacılık veya sigortacılık sektörlerinde kullanılan ve mevcut müşterilerin kaybını önceden tahmin etmeye dayanan analiz yöntemidir. Bu tahminler sayesinde müşteri kaybının önüne geçilmesi için müşteri ilişkileri yönetimi (CRM) kapsamında çözümler üretilebilmektedir. Literatürde, müşteri yıpranması (customer attrition), müşteri sallanması (customer churn), müşteri cayması (customer turnover), müşteri kaçması (customer defection) gibi terimlerle de adlandırılmaktadır. Telekomünikasyon, bankacılık, sigortacılık gibi müşteri sürekliliği bulunan sektörlerde müşteri kaybı kritik bir öneme sahiptir. Çünkü çoğu zaman mevcut müşterilerin tutulması, yeni müşteri kazanmaya göre daha düşük maliyetli operasyonlar gerektirir [25].

Günümüzde firmaları karşı karşıya getiren rekabet ortamında hizmet ile ürünün kalitesi birbirine çok yakın olmaktadır. Bunun en büyük sebebi rekabetin prestij savaşına dönmesidir. Rekabet edilen ortak ürün sayısı az olduğundan rekabet edilen ürüne yönelik çeşitli kampanyalar gibi ekstra işler yapılmaktadır. Bu rekabet sırasında yeni müşteri elde etmek için yapılan maddi yatırımlar çok ciddi boyutlara ulaşmıştır. Fakat yeni müşteri arayışına başlamadan öncelikle eldeki müşteriye yönelik çalışma yapılması maliyet açısından kazanç sağlamaktadır. Hatta mevcut müşterilerine değer veren firmalar yeni müşterisi olmayı düşünen adayları da cezbeder. Bu nedenle mevcut müşteri çok değerlidir. Özellikle de finans sektörü gibi risk barındıran ve müşterilerin nadiren değiştiği alanlarda mevcut müşteriyi elden kaçırmamak önemlidir. Telekomünikasyon firmaları, internet servis sağlayıcıları, bankalar gibi kazançları abonelerinden gelen sektörlerde müşteri kaybı hayati öneme sahiptir. Müşteri kaybı, müşteri memnuniyetsizliği, uygun fiyatlı veya kaliteli hizmet sunan rakiplerin varlığı ve ekonomik sebeplerden dolayı olabilir.

Pazarlama uzmanı Kotler;

"Yeni müşteri kazanmanın maliyeti, mevcut müşterileri memnun etme ve elde tutma maliyetinin 5-10 katını bulabilir." ve "Ortalama bir şirket, her yıl mevcut müşterilerinin %10

ile %30 oranındaki bölümünü kaybeder [26].”

Müşteri elde tutma hakkında bu şekilde görüş belirtmiştir. Bu bilgilere bakılarak mevcut müşterileri kaybetmemenin pazarlama ve satış maliyetlerinin azalmasına önemli etkileri olmaktadır. Bunun nedeni, genellikle mevcut müşterileri korumak, yeni müşteriler kazanmaktan daha düşük maliyet gerektirir. Yeni müşteri kazandığı kadar mevcut müşterisini kaybeden firmaların bu noktada müşteri oranında değişim olmayacağını ve mevcut müşteriyi kaybetmemenin yeni müşteri kazanmak için daha az maliyetli olacağını gözden kaçırırlar [27].

Günümüzde müşterilerine hizmet sağlayan çoğu şirkette müşteri kaybı üzerine çalışma yapan departmanlar bulunmaktadır. Telekomünikasyon şirketlerinin mevcut müşterilerinden bazıları operatör değiştirmek istemektedirler. Bu müşterilere yönelik telekomünikasyon firmaları bazı özel teklifler ile ayrılmak isteyen müşterilerini vazgeçirmeye çalışmaktadırlar. Fakat bu tür kampanyalar çok rağbet görmemektedir. Bunun sebebi rekabet ortamında mevcut müşterilerin düşünülmemesi olabilir. Yeni müşteri kazanmak için firmalar paralarını ve zamanlarını harcarken mevcut müşterilerine yönelik düşük seviyede imkan sundukları için onlar da operatör değiştirmeyi tercih ederler. Bunun doğal sonucu ise yüksek müşteri kaybı olmaktadır.

3. VERİ MADENCİLİĞİ

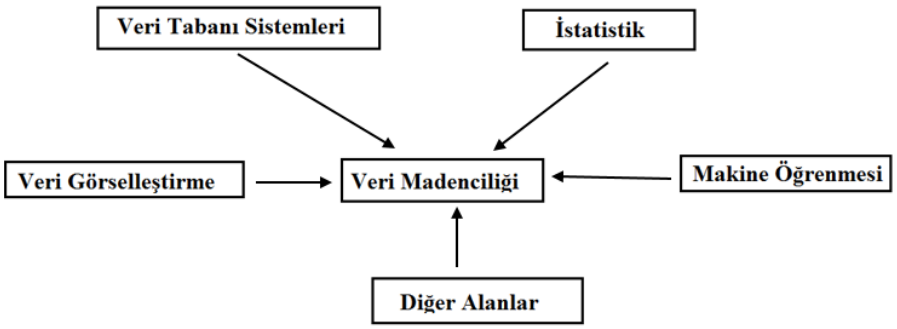
Veri madenciliği konusunda çeşitli tanımlar yapılmaktadır. En genel tanım olarak karmaşık verilerden kullanılabilir bilgi elde edilmesi şeklinde yapılabilir [28]. Veri madenciliği, literatürde geçen isimlendirilmesiyle bilgi keşfi; çok büyük ve karmaşık veri topluluklarından faydalı ve kullanılabilir bilgilerin çıkarılması işlemidir. Bu işlem yapılırken birden çok alan kullanılmaktadır. Bu alanlar veri tabanı yönetim sistemleri, istatistik, yapay zekâ, makine öğrenmesi alanlarıdır. Bu alanların ortak bir paydada buluşarak veriler üzerinde bilgi keşfi yapılmasına veri madenciliği adı verilir [29]-[31]. Kısaca veri madenciliği büyük veri topluluklarından ileriye dönük tahmin yapılmasını sağlayacak bağıntı ve kuralların bilgisayarlar aracılığıyla yapılmasıdır [32].

3.1. VERİ MADENCİLİĞİNİN TARİHİ GELİŞİMİ

Veri madenciliği yöntemlerine yönelik ilk çalışmalar matematikçiler tarafından 1950'de bilgisayar bilimleri alanında yapılmıştır. Bu çalışmaların sonucu olarak yapay zekâ ve makine öğrenmesini literatüre katmışlardır. 1960'lı yıllarda matematik ve bilgisayar bilimi ile uğraşanlar regresyon analizi, en büyük olabilirlik kestirim algoritması, sinir ağları gibi yeni algoritmaları keşfederek veri madenciliğinin ilk adımlarını oluşturmuşlardır. 1970, 1980 ve 1990'lı yıllarda yeni programlama dillerinin geliştirilmesiyle genetik, kümeleme ve karar ağacı gibi algoritmalar ortaya çıkmıştır. 1990 yılı ile beraber büyük veri topluluklarından bilgi keşfinin ilk adımları oluşturulmuştur. Yeni teknolojilerin gelişmesiyle veri madenciliği, bilgi keşfinde kullanılan süreçler topluluğu haline gelmiştir [31].

3.2. VERİ MADENCİLİĞİNİN ÇALIŞMA ALANLARI

Veri madenciliğinin günümüzde yaygın bir kullanım alanı bulunmaktadır. Pazarlama, bankacılık, sigortacılık ve telekomünikasyon gibi özellikle müşteri tabanlı alanlarda yaygın şekilde kullanılmaktadır. Veri madenciliği Şekil 3.1'de görüldüğü gibi birçok farklı alanı içeren alanlar topluluğudur [30]. Bu alanlar; veri tabanı sistemleri, istatistik, makine öğrenmesi, veri görselleştirme ve diğer disiplinlerdir.



Şekil 3.1. Veri madenciliğinin birçok alanla bileşimi.

3.2.1. Sağlık Alanında Yapılan Çalışmalar

Sağlık alanında birçok veri madenciliği çalışması yapılmıştır. Bu çalışmalarda hastaların sağlık verileri kullanılmaktadır. Sağlık alanında yapılan çalışmalara örnek verilecek olursa; bir kişinin ailesinde olan bir hastalığın kendisinde ya da diğer aile üyelerinde olup olmadığına yönelik tahminsel çalışma, ölüm oranları ve salgın hastalıkların tahmin edilmesi gibi örnek çalışmalar yapılmaktadır [33].

Bu çalışmaların ortak amacı olumlu sonuç elde ederek hastalık ihtimali olan kişileri bilinçlendirmek ve tedaviye yönlendirmektir. Harleen Kaur ve arkadaşları sınıflandırma yöntemlerinden karar ağacı ile model kurarak göğüs kanseri riskini tahmin etmeye çalışmışlardır. Bunun için hastaların yaş ve cinsiyet gibi verilerinden yararlanmışlardır [33]. Günümüzde genetik mühendisleri bu çalışmaları geliştirmek amacıyla çalışmalar gerçekleştirmektedirler.

3.2.2. Kamu Alanında Yapılan Çalışmalar

Kamu yöneticileri günümüzde verinin önemine hakimdirler. Müşteriye hizmet sunan özel firmalardaki hizmet anlayışı gibi devlet kurumları da vatandaşlara yönelik aynı kalitede hizmet verebilme arayışı içindedirler. Örnek olarak devletin kamu güvenliğini sağlamak amacıyla güvenlik olaylarını olmadan tahmin, vergi yolsuzluklarını tespit gibi uygulamalar gerçekleştirilmektedir. Güvenlik kurumları için suç istatistiklerine ait web tabanlı raporlama, vatandaşların suça meyillerini tahmine yönelik uygulama, olay anında suçu engelleyebilme gibi uygulamalar gerçekleştirmek üst düzey analitik uygulamalar ile olabilir.Son zamanlarda e-devlet uygulamaları hayatımıza girmiş ve oldukça yaygın kullanılmaktadır. E-devlet hemen hemen her yaş grubu vatandaşın ihtiyacını karşılamaktadır. E-devlet uygulamaları öğrencilerin öğrenci belgesi alması, gerekli yerlere vermek için sabıka kaydı sorgulayıp alma, sigorta prim kontrolü,

belediye ile ilgili hizmetler gibi sayısız hizmet imkanı veren bir uygulama topluluğudur. E-devlet uzmanları da vatandaşların istediği hizmeti alması için bu yeni uygulamaları gerçekleştirme, olan uygulamaları geliştirerek bilgi güvenliklerinden sorumludurlar. Veri madenciliği ile e-devlet uygulamalarında vatandaşlara hizmet anketleri doldurularak ya da mevcut bilgiler ile en çok hangi hizmetler alınıyorsa o hizmetlere yönelik uygulamalar çoğaltılabilir. Vatandaş bilgilerine göre bölgesel ihtiyaçlar belirlenerek çalışmalar yapılabilir [34].

3.2.3. Telekomünikasyon Alanında Yapılan Çalışmalar

Telekomünikasyon sektöründe de abone temelli diğer alanlar gibi abonelerin ayrılması büyük bir sorundur. Telekomünikasyon firmaları ayrılma potansiyeli olan müşterilerine yönelik çalışmalar düzenleyerek ayrılan müşterilerini kaybetmemek amacıyla çeşitli kampanyalar sunarak ayrılma oranını düşürmek isterler. Telekomünikasyon sektörüne yönelik örnekler çoğaltılacak olursa:

- Müşterilerin ayrılma nedenlerinin belirlenmesi,
- Müşterilerin demografik verilerinin ayrılma riskine etkisinin bulunması,
- Yapılan kampanyalara katılım oranının artırılması,
- Mevcut müşterilerinin kaybının engellenmesi ve yeni müşterileri kazanılması,

gibi verilebilir.

3.2.4. Finans Alanında Yapılan Çalışmalar

Finans sektöründe hayati işlerden bazıları; mevcut müşteriyi elde tutarak yeni müşteriler kazanma, maliyet kaybını alt seviyelere indirme, kayıp oranını azaltma, müşteri memnuniyetini sağlama, kaçak oranını düşürme gibi işlerdir. Müşteri grupları neyi tercih ettiği, tercih zamanı ve nedenine yönelik tahmini olan firmalar müşteriye yönelik talep oluşturma ve doğru zamanda bu talebi karşılama noktasında önde olacaklardır. Özellikle bankalar müşteri kaybını engellemek için büyük çaba sarf ederler [34]. Finans verileri çok boyutlu verilerdir. Bu verileri veri madenciliği ile işleyerek firmalara müşteri hakkında yararlanılabilir bilgi verilmesi durumunda maliyet noktasında büyük kazançlar sağlanabilir. Finans sektöründe müşterilerinin ayrılma tahmini dışında güvenliğe yönelik de veri madenciliği uygulamaları yapılmaktadır.

3.3. VERİ MADENCİLİĞİ SÜRECİ

Veri madenciliğini bir süreç olarak değerlendirmek gerekiyor. Bu süreç aşağıda belirtilen adımları içermektedir [29] :

- Veri temizleme,
- Veri bütünleştirme,
- Veri indirgeme,
- Veri dönüştürme,
- Veri madenciliği algoritmalarının uygulanması,
- Sonuçları değerlendirme

Bu bölümde bahsedilecek kavramlar makine öğrenmesi süreci adımları ile benzer adımlar içermektedir. Makine öğrenmesi süreci adımlarından veri hazırlama adımına veri madenciliği sürecinde veri temizleme, veri bütünleştirme, veri indirgeme ve veri dönüştürme adımları karşılık gelmektedir. Makine öğrenmesi süreci adımlarından modelleme adımına veri madenciliği sürecinde veri madenciliği algoritmalarının uygulanması adımı karşılık gelmektedir. Aynı şekilde makine öğrenmesi süreci adımlarından model değerlendirme ve seçim adımına veri madenciliği sürecinden sonuçları değerlendirme adımı karşılık gelmektedir. Yukarıda bahsedilen makine öğrenmesi ve veri madenciliği süreçlerinin birbirine karşılık geldiği adımları aşağıda yer alan Tablo 3.1'de daha iyi anlaşılmaktadır.

Tablo 3.1. Makine öğrenmesi ve veri madenciliği süreçlerinin birbirine karşılık geldiği adımlar.

Veri Madenciliği Adımları	Makine Öğrenmesi Adımları
Veri Temizleme	Veri Hazırlama
Veri Bütünleştirme	
Veri İndirgeme	
Veri Dönüştürme	
Veri Madenciliği Algoritmalarının Uygulanması	Modelleme
Sonuçları Değerlendirme	Model Değerlendirme ve Seçim Adımı

Bu bölümde veri madenciliği süreci hakkında kısa bilgi verilecek olup veri seti üzerinde yapılacak işlemlerde takip edilecek olan makine öğrenmesi süreci adımlarına ilişkin daha detaylı bilgi Bölüm 4.1'de anlatılacaktır.

Veri Temizleme; veri setlerine çeşitli nedenlerle yanlış ve eksik girilmiş olan verilere gürültü denir. Veri temizleme kaynaklarda önişleme olarak da geçer. Bu adımda hatalı veriler ve eksik verilerin çözümüne yönelik çalışma gerçekleştirilir. Veri setindeki gürültüyü çözmede eksik değeri olan veriler çıkarılabilir, kayıp veriler yerine herhangi bir sabit değer yazılabilir, nümerik verilerin ortalaması alınarak kayıp veriler yerine bu değerler yazılabilir, kategorik verilerin yerine en çok tekrar eden veri yazılabilir, eksik verilerin yerine üst ya da alt kaydın verisi yazılabilir. Ayrıca verilere uygun tahmin yapmak amacıyla karar ağacı, regresyon analizi gibi yöntemler ile de eksik veri çözümlenebilir.

Veri Bütünleştirme; farklı kaynaklardan elde edilen verilerin ortak bir çalışmada olabilmesi için aynı türe dönüştürülmesi gerekir. Bu işleme veri bütünleştirme denir. Örnek olarak cinsiyet değişkeninin bir veri tabanında 0/1 gibi nümerik iken başka veri tabanlarında Erkek/Kadın veya E/K şeklinde karakter veri tipinde olması analizde tutarsızlığa neden olur. Bu tutarsız değerler analizin başarısız olmasına neden olur. Analizde başarılı olmak için veri türleri aynı türe dönüştürülür.

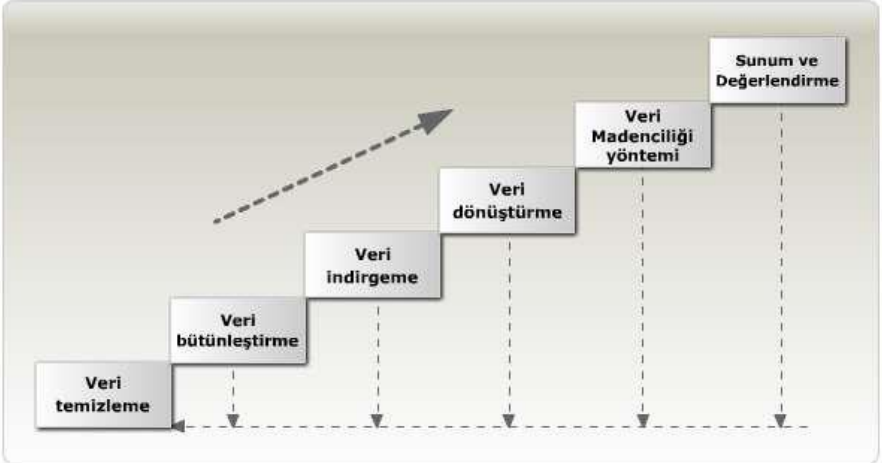
Veri İndirgeme; büyük sayıdaki verinin analizi zahmetlidir. Daha hızlı ve doğru sonuç almak için veriler indirgenebilir. Veriler azaltıldığında ilk hali ile sonuç değişmiyorsa büyük veri setini kullanarak donanımsal yük yapmaya gerek yoktur. Bundan dolayı veri indirgeme önemlidir.

Veri Dönüştürme; verilerin içeriği korunarak kullanılacak modele göre şeklini değiştirme işlemidir. Değişkenlerin varyansları birbirlerinden farklı olduğu takdirde varyans oranı büyük olan değişkenler diğer değişkenlerin önemini yitirmesine sebep olabilir. Bunu koruyarak veri dönüşüm yapılmalıdır.

Veri Madenciliği Algoritmasının Uygulanması; veri temizleme, veri indirgeme, veri bütünleştirme, veri dönüştürme adımlarının gerekli ise hepsi uygulandıktan sonra elde edilen veri setine veri madenciliği ve makine öğrenmesi algoritmaları ile modeller uygulanan adımdır.

Sonuçları Değerlendirme; bu adım modeller oluşturulduktan sonra bu modellerin karşılaştırılması ve değerlendirilmesinin yapıldığı adımdır.

Yukarıda kısaca anlatılan bu adımların Şekil 3.2'de görüntüsü yer almaktadır [29]. Bölüm 4.1'de bu adımlar makine öğrenmesi süreci adımlarında daha detaylı anlatılacaktır.

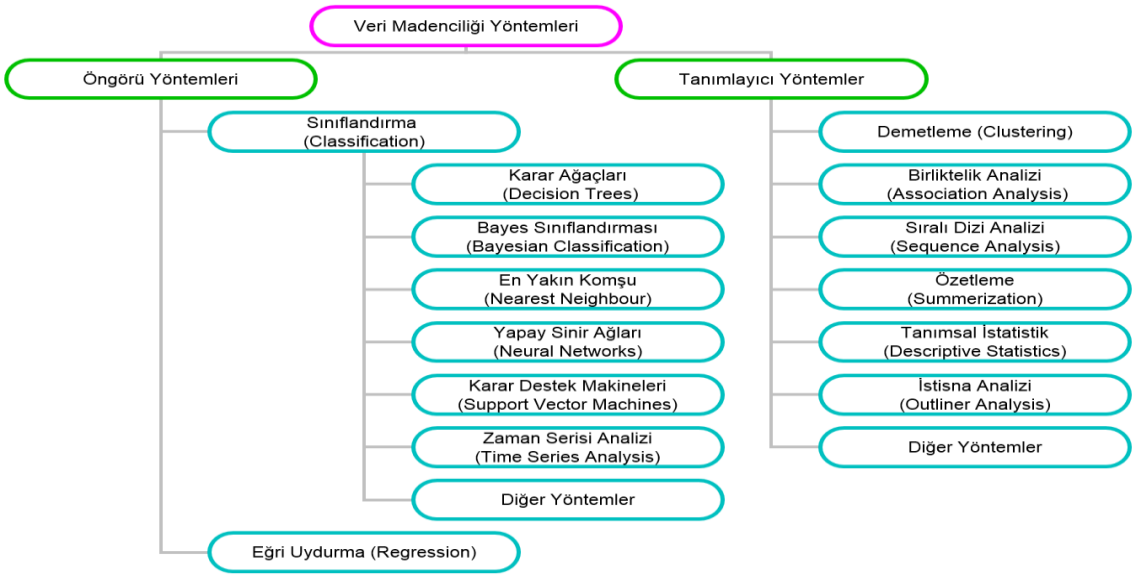


Şekil 3.2. Veri madenciliği süreci adımları.

3.4. VERİ MADENCİLİĞİ YÖNTEMLERİ

Veri madenciliği yöntemleri verilerin farklı boyutlarını kullanarak analiz edilmesi, kategorileştirilmesi, özetlenmesi ve bağıntıların belirlenmesi amacıyla kullanılan yöntemlerdir. Veri madenciliği yöntemleri ile örneğin bir kişinin telefon numarasına ulaşması gibi bir işlem gerçekleştirilmez. Fakat akıllı telefon kullananlar diğer cihazları kullananlara göre daha çok operatör değiştirir şeklinde bir analiz sonucuna ulaşılabilir. Buna benzer analizleri yapmak için istatiksel yöntemler, akıllı sistem algoritmaları, örüntü tanıma, makine öğrenmesi gibi birçok farklı yaklaşım izlenmektedir. Bu yöntemler tahminsel veya açıklayıcı olmak üzere iki gruba ayrılır. Tahminsel yöntemler ile bilinen bazı değişkenleri kullanarak bilinmeyen bir değişkenin değerini bulmak için kullanılan yöntemlerdir. Açıklayıcı yöntemler ise, verinin anlaşılabilirliğini artırmak, bilgi keşfini kolaylaştırmak amacıyla eldeki veriler örüntüleri keşfetmeye yönelik olarak geliştirilen yöntemlerdir [35].

Veri madenciliğinde kullanılan yöntemler, öngörü ve tanımlayıcı olmak üzere ikiye ayrılır. Veri madenciliği yöntemleri Şekil 3.3'teki gibi görülmektedir [2].



Şekil 3.3. Veri madenciliği yöntemleri.

Bu çalışmada veri madenciliği yöntemleri sınıflandırma, kümeleme, birliktelik kuralı olmak üzere üç kısımda anlatılacaktır.

3.4.1. Sınıflandırma Yöntemi

Sınıflandırma, veri madenciliğinde en çok tercih edilen yöntemdir. Mevcut veri setinin belirli bir bölümü eğitim verisi olarak ayrılarak sınıflandırma için kurallar oluşturulur. Bu kurallar sayesinde yeni bir durum görüldüğünde ne yapılacağına dair karar verilebilir.

Veri madenciliği sınıflandırma yöntemlerinden en çok karar ağacı tercih edilmektedir. Karar ağacına ek olarak lojistik regresyon, sinir ağları da çok sık tercih edilmektedir. Veri analizcilerinin çoğu veri setlerini gruplamak için sınıflandırma yöntemleri kullanırken aslında hem veri madenciliğinde model kurmada hem de veri hazırlama da kullanılmaktadır [28].

Genellikle tahmin üzerine olan çalışmalarda sınıflandırma algoritmaları ile modeller oluşturulmaktadır. Bu çalışmada müşteri kayıp analizine yönelik tahmin çalışması olduğu için modeller, sınıflandırma algoritmalarıyla oluşturulmuştur.

3.4.2. Kümeleme Yöntemi

Verilerin aralarındaki benzerlikleri dikkate alınarak gruplanma işlemine kümeleme denir. Veriler arası uzaklık kümeleme yöntemlerinin hemen hemen hepsinde kullanılır. Genelde veriler sıralı ise en yakın komşu ve en uzak komşu algoritmaları kullanılır.

Belirli bir sırası olmayan veriler üzerinde k-ortalamalar yöntemi kullanılır.

Küme, büyük boyutlu yakın nesnelerin uzayda oluşturdukları bulutlar seklinde tanımlanabilmektedir [36]. Kümeleme kavramının ortaya çıkması 1984 yılında Londra'da bir kolera salgını sırasında olduğu söylenebilir. Bu felakette çok ciddi sayıda ölüm vakaları kaydedilmiştir. İngilizler salgında ölen kişilerin yerlerini haritada işaretlemişler ve işaretlenmiş olan bazı bölgelerde yoğunluk fark etmişlerdir. Haritada yoğunluğun olduğu bölgelerde probleme su pompalarının sebep olduğunu bulmuşlar ve koleranın yayılması engellemişlerdir [37].

3.4.3. Birliktelik Kuralları Yöntemi

Birliktelik Analizi, belirli bir veri setinde aşırı derecede birliktelik görülmesine ilişkin kural çıkarmaya denir. Bu ilişkinin sonucu birliktelik kuralları şeklinde gösterilir. Birliktelik kuralları en yaygın şekilde market sepeti uygulamasında görülmektedir. Market sepet uygulamasında, müşterilerin alışverişlerinde aldıkları ürünler arasındaki birliktelikler bulunarak müşterilerin satın alma alışkanlıkları belirlenir [2].

3.5. Veri Madenciliği Programları

Veri Madenciliği uygulamalarını hayata geçirmek için bilgisayar aracılığı ile geliştirilen veri madenciliği programları kullanılmaktadır. Bilgisayar teknolojilerinin gelişmesi ile veri madenciliği artık daha hızlı ve etkili şekilde uygulanabilir hale gelmiştir.

Anlık olarak sürekli artan veriler yüzünden her an binlerce veri topluluğuna maruz kalmaktayız. Bu verilerin zararlı olabileceğine bakmadan her yerden ve anlık olarak her yaş grubunun elinde bulunduğu cep telefonu, tablet gibi taşınabilir cihazlarla erişilebilinmektedir. Bilimsel veri, sosyal medyadan gelen anlık veri, uydu ile gelen veriler gibi veri toplulukları arasından verileri işlemek ve ona göre verileri kullanmak gereklidir. Bu da veri madenciliği sayesinde olmaktadır. [38].

Veri Madenciliği uygulamalarını gerçekleştirmek için bir sürü program geliştirilmiştir. Bu programlara örnek olarak Statistical Package for the Social Sciences (SPSS), SAS gibi bir ticari program ile RapidMiner (YALE), Orange, Waikato Environment for Knowledge Analysis (WEKA), R, Konstanz Information Miner (KNIME) gibi açık kaynak programlar verilebilir. Bu bölümde kısaca en çok kullanılan veri madenciliği programları anlatılacaktır.

3.5.1. Knime

Konstanz Information Miner, Konstanz Üniversitesi veri madenciliği grubu tarafından geliştirilen veri madenciliği programıdır [39]. KNIME kullanıcılara bir yazılım geliştirme IDE'si imkanı sağlar. Bu IDE ile kullanıcılar kendi modüllerini geliştirebilmektedir. Programın kurulum şartı yoktur. Knime ile veri almak için verilerin .txt uzantılı metin belgesi veya .arff formatında olması gereklidir [40].

3.5.2. Orange

Slovenya'da bulunan Ljubljana Üniversitesi bünyesinde yer alan Bilgisayar ve Enformatik Bilimleri Bölümü Yapay Zekâ ekibi tarafından geliştirilmiş bir programdır. Orange yazılımı C++ dili ile geliştirilmiştir. Yalnızca metin belgesinden veri alır [40].

3.5.3. RapidMiner (Yale)

RapidMiner, Ralf Klinkenberg, Ingo Mierswa ve Simon Fischer tarafından Dortmund Teknoloji Üniversitesi Yapay Zeka Biriminde yapılmış olan bir programdır. Yale ise Yale üniversitesinde Java dili ile geliştirilmiş bir programdır [41]. Yale 2007 yılından itibaren RapidMiner [42] adı altında kullanılmaya başlanmıştır. Diğer programlardan büyük farkı 22 adet dosya formatından veri alabilmesidir. Veri Madenciliği ve Makine Öğrenme Algoritmalarını da kapsayan RapidMiner, Weka gibi oldukça fazla algoritmaya sahiptir. Veri analizi, önişleme, veri madenciliği yöntemleri gibi işlemleri içermektedir. Oracle, MS SQL Server, MySQL, IBM DB2 başta olmak üzere birçok veri tabanını ve metin dosyalarını desteklemektedir [40]. Bu açıdan en kapsamlı yazılımlardan biridir. Excel dosyalarıyla bağlantı kurulabilmektedir. MS Windows, Linux, Mac Os X işletim sistemlerinde çalışabilmektedir.

3.5.4. Weka

Waikato Environment for Knowledge Analysis kelimelerinin kısaltılmasıdır [38]. Waikato Üniversitesinde, Java platformu üzerinde geliştirilmiş ve GNU genel kamu lisansı altında olan açık kaynak kodlu bir veri madenciliği programıdır. Java Database Connectivity (JDBC) kullanarak SQL veri tabanına ulaşır [15]. İçerisinde tüm veri madenciliği ve makine öğrenmesi algoritmaları vardır. Veri analizi, önişleme, veri madenciliği yöntemleri gibi işlemleri içermektedir. WEKA'ya özel tasarlanan .arff (Attribute Relationship File Format) dosya formatı üzerinde çalışır.

3.5.5. SAS

Anthony Barr, James Goodnight, John Sall ve Jane Helwig isimli dört kişi tarafından 1976 yılında Statistical Analysis System ismi ile kurulmuştur. Günümüzde SAS borsaya açık olmayan dünyanın en büyük yazılım şirketlerinden biridir. SAS, IBM, Microsoft, Oracle gibi firmalar ile ciddi rekabet içindedir. 2009 yılında IBM'in SPSS firmasını satın alması ile SAS, IBM ile rakip olmuştur [43].

3.5.6. SPSS

Statistical Package for the Social Sciences, 1968 yılında piyasaya çıkmış istatistiksel analize yönelik bir bilgisayar programıdır. SPSS, 2009'da IBM şirketine satılmıştır. SPSS özellikle sosyal bilimler alanında istatistiksel analiz için kullanılmaktadır. Pazarlama şirketleri, sağlık araştırmacıları, anket şirketleri, devlet kurumları, eğitim araştırmacıları tarafından da kullanılmaktadır [44].

3.5.7. R

R, istatistiksel hesaplama için geliştirilen bir bilgisayar programı olup aynı zamanda programlama dilidir. İçerisinde binlerce paket vardır. Bu paketler ile veri madenciliği, veri görselleştirmek için grafik oluşturma gibi bir sürü işlem yapılabilir. Yeni Zelanda Auckland Üniversitesinden Ross Ihaka ve Robert Gentleman tarafından geliştirilmiş olan R paketlerinin ihtiyaca göre yazılarak artırılması sebebiyle sürekli gelişmektedir [45]. S yazılımına alternatif olması amacıyla açık kaynak kodlu olarak geliştirilmiştir. İstatistikçiler arasında standart haline gelmiştir. R, istatistiksel yazılım geliştirme ve veri analizi alanında kullanılmaktadır. Genel Kamu Lisansı (GNU) altında olup her işletim sistemi için sürümü mevcuttur.

3.5.8. Veri Madenciliği Programları Karşılaştırılması

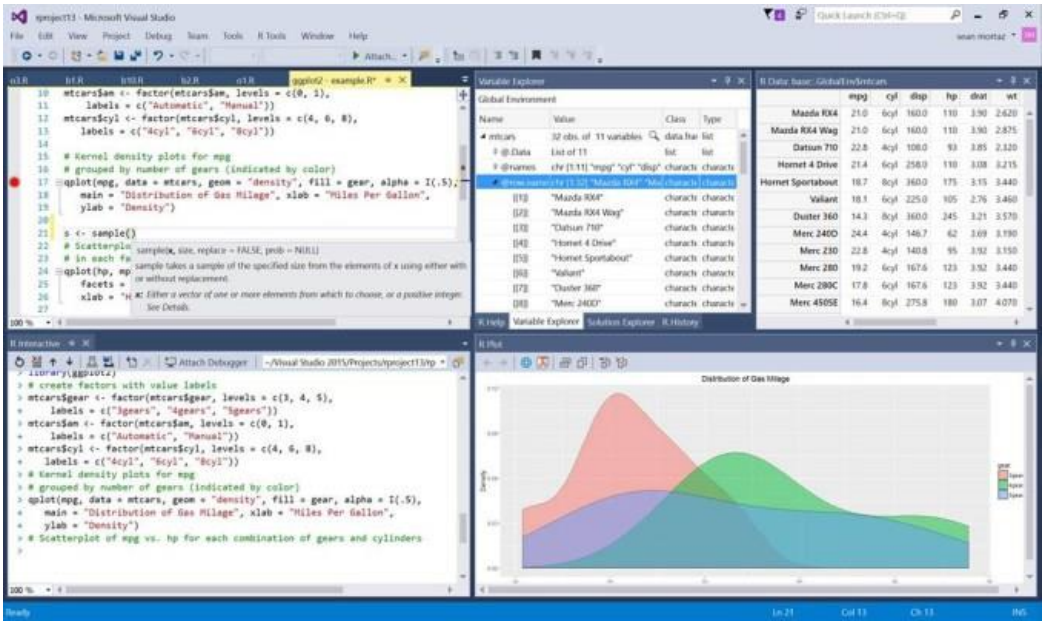
Veri madenciliği ile yapılan büyük verilerin anlamlı ve kullanılabilir hale getirilmesi işlemi çok önemlidir. Veri madenciliği özellikle abone temelli kurumlarda ticari yönden avantaj sağlaması nedeniyle yatırım yapılan bir alan haline gelmiştir. Artık şirketlerin hizmeti müşterinin ayağına götürmesi de yeterli olmamaktadır. Şirketlerin müşterilerini ve davranışlarını tanımaları çok önemlidir. Bunları yapabilmenin yolu da sürekli olarak artan veri kaynaklarını işleyip iş süreçlerine aktarabilmekten geçmektedir.

Dünyanın en büyük 500 şirketi sıralamasına bakıldığında büyük değişim görülmektedir.

Enerji, sağlık, doğal kaynaklar gibi alanların dışında kalan finans, sigortacılık, telekomünikasyon gibi alanlardaki şirketlerin özellikle müşteri davranışlarını öngörerek son zamanlarda bu listelere girmesi gerçekleşmektedir. Bundan dolayı veri bilimi son zamanlarda hiç olmadığı kadar gelişme göstermiştir.

Veri madenciliği alanında büyük firmalar yazılım programları hazırlamışlardır. Büyük yazılım firmaları arasında SAS firması yaptığı yazılım ile ön plandadır. Bir diğer önemli firma olan SPSS 2009 yılında IBM'e satılmasından sonra SAS'ın arkasında kalmıştır. Teknolojinin gelişmesi veri bilimini de yakından etkilemiştir. SPSS gibi firmalar bu gelişime ayak uyduramamışlardır. Eski versiyonlarını ücretsiz veren RapidMiner, Zurich Üniversitesi'nde bir topluluk tarafından geliştirilen açık kaynak versiyonu da olan KNIME, Kaliforniya merkezli 700 den fazla müşterisi olan Alterix'in de veri madenciliği üzerine programları vardır. 2013 yılında SAS'ın KXEN'i, 2014 yılında DELL'in Statsoft (Statistica)'yı satın alması, Microsoft'un ticari olarak R'ı müşterilerine sunan Revolution Analytics'i 2015 yılında bünyesine katması veri biliminde yarışı hızlandırmıştır [45].

Veri madenciliği konusuna ülkemizde de son zamanlarda oldukça önem verilmeye başlanmıştır. Finans, telekomünikasyon gibi müşteri tabanlı sektörlerde veri departmanları yer almaktadır. Üniversitelerin yanısıra eğitim kurumları yukarıda adı geçen programların eğitimlerini vermektedir. Bu sayede veri bilimi alanında yetişen eleman sayısı artmaktadır.



Şekil 3.4. Microsoft'un visual studio 2016 programında R entegresi.

Şekil 3.4'te Microsoft'un Visual Studio 2016 da R ile programlama yapılabilmesi için yaptığı geliştirmenin görüntüsü vardır.

R programı "Hızlı, Kolay ve Gelişmiş" kısaca bu şekilde anlatılabilir. Platform bağımsız olması, açık kaynak olması, hali hazırda sayısız paket desteği olması ve forum, blog gibi internet ortamlarında her türlü sorulara sürekli cevap bulanabilmesi sebebiyle son zamanlarda R dilinin kullanım oranları sürekli artmaktadır.

Tablo 3.2'de de görüldüğü üzerine R dilinin seçilen diğer programlardan geri kalmadığı görülmektedir [46]. Ayrıca R paketleri sayesinde istenen işleme göre genişleyebilir bir program olduğu ve istatiksel hesaplamanın yanında paket çeşitliliği sayesinde diğerlerine göre daha avantajlı olduğu görülmektedir.

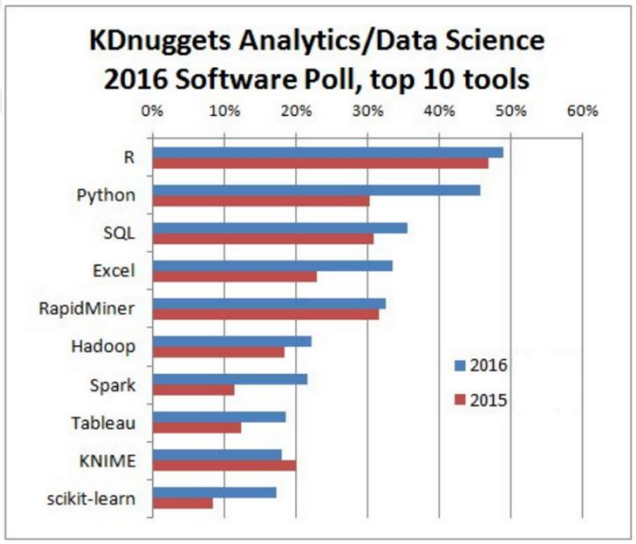
Tablo 3.2. Açık kaynak kodlu veri madenciliği programlarının karşılaştırılması.

	Keel	Knime	Orange
Veri Madenciliği Algoritmaları	Var	Var	Var
Makine Öğrenmesi Paketleri	Var	Var	Var
İstatiksel Hesaplama	Var	Var	Var
Veri Analizi	Var	Var	Var
Önişleme	Var	Var	Var
Değişken Seçimi	Var	Var	Var
Görselleştirme	Var	Var	Var
GKA	İyi	İyi	İyi
Genişletilebilirlik	Evet	Evet	Evet
Esneklik	Evet	Evet	Evet
Kullanım Kolaylığı	Evet	Evet	Evet
Hatasız Çalışma	Evet	Evet	Evet
Dokümantasyon	Var	Var	Var
Script Yazma	Var	Var	Var
Eklenebilir Paketler	Evet	Evet	Evet
Veri Alma/Verme	Var	Var	Var
Excel de çalışabilme	Evet(import ile)	Hayır	Hayır
Veritabanlarıyla Çalışabilme	Var	Var	Var
Desteklenen Dosya Formatları	.dat, .arff, .csv, .xml, .txt, .prn, .xls, .dif, .html	.arff, .csv	. arf (sadece okuma) .basket, .csv (sadece okuma) .data, .txt, .names, .xls, .tab,

Tablo 3.3'te Veri madenciliği programlarının 2016 yılındaki en çok kullanılan ilk 10 program ve kullanım oranı görülmektedir [47].

Tablo 3.3. Veri madenciliği programlarını tercih oranı.

Veri Madenciliği Programları	Kullanım Oranı (%)
R	%49
Python	%45.8
SQL	%35.5
Excel	%33.6
RapidMiner	%32.6
Hadoop	%22.1
Spark	%21.6
Tableau	%18.5
KNIME	%18.0
Scikit-learn	%17.2



Şekil 3.5. 2015'e kıyasla 2016'daki en iyi 10 açık kaynak kodlu veri madenciliği programları.

R programı, Platform bağımsız olması, açık kaynak olması, hali hazırda sayısız paket desteği olması gibi sebeplerle R dili tercih edilme oranını her geçen gün artmaktadır. Buda Tablo 3.2, Tablo 3.3 ve Şekil 3.5'te görülmektedir [48]. Bundan dolayı bu çalışmada R veri madenciliği programı seçilmiş ve tüm uygulamalar R'da gerçekleştirilmiştir.

4. MAKİNE ÖĞRENMESİ

Makine öğrenmesi, eldeki verilerden yorum çıkarmak için matematik ve istatistik gibi bilimleri kullanarak bilinmeyen bir değişkeni tahmin etme şeklinde tanımlanabilir. Makine öğrenmesindeki bazı terimler aşağıda sırasıyla açıklanmaktadır.

- Gözlemler: Öğrenmek için kullanılan her bir veri gözlem olarak adlandırılır.
- Özellikler: Bir gözlemi ifade eden veriler o gözlemin özellikleridir.
- Etiketler: Gözlemlere atfedilen kategorilerdir.
- Eğitim Verisi: Algoritma ile kurulan modelin öğrenmesi için oluşturulan gözlem topluluğuna denir.
- Test Verisi: Algoritmanın eğitim veri seti aracılığı ile şekillendirdiği modelin gerçeğe yakın olduğunu test etmek için kullandığı veri setidir. Eğitime katılmaz ve eğitim bittikten sonra etiketsiz olarak algoritmaya verilerek algoritmanın tahminde bulunması beklenir.

Gözetimli ve gözetimsiz öğrenme olmak üzere iki tür öğrenme türü vardır. Gözetimli öğrenme etiketli gözlemlerden bir sonuç çıkarmaya yönelik öğrenme sürecine denir. Veri madenciliği ve makine öğrenmesi metotlarından sınıflandırma ve regresyon bu öğrenmeye girer. Gözetimsiz öğrenme etiketsiz gözlemlerden öğrenme sürecine denir. Gözetimli öğrenmeden farkı herhangi bir bilgi girişi olmadan veriler verildiğinde bu verilerden çeşitli sonuçların çıkarılması beklenen öğrenme türüdür.

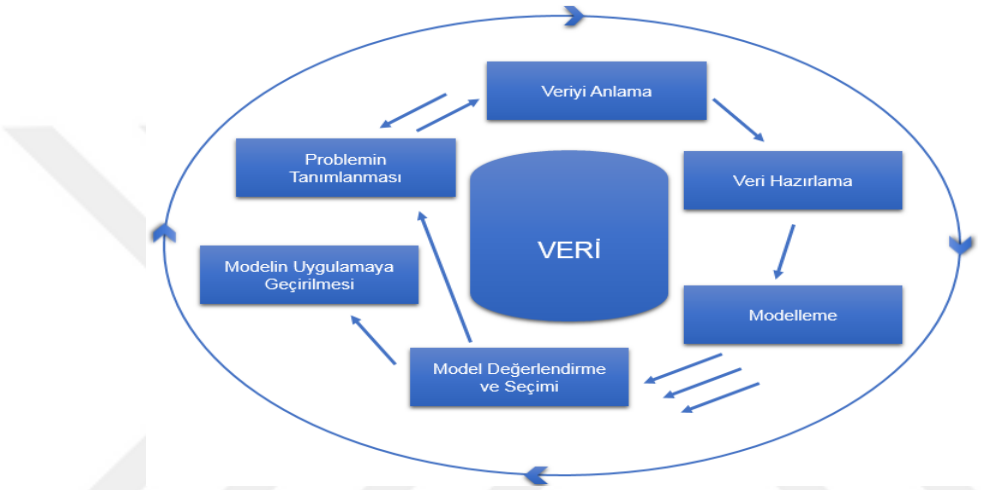
4.1. MAKİNE ÖĞRENMESİ SÜRECİ ADIMLARI

Makine öğrenmesi ile veri madenciliği arasında yakın bir bağ vardır. Veri madenciliği olarak bilinen alanda, veri madenciliği ve makine öğrenmesi algoritmaları ile donanım bakım kayıtları, kredi başvuruları, finansal işlemler, tıbbi kayıtlar gibi verileri içeren büyük ticari veri tabanlarından değerli bilgiyi keşfetmek için sıklıkla kullanılmaktadır [49]. Veri madenciliği ve makine öğrenmesi süreci için literatürlerde verilen adımlar birbiriyle benzerdir [50]-[52]. Bu çalışmada veri madenciliği için çapraz endüstri standart süreç modeli (CRISP) modeli örnek model olarak ele alınmış ve bu sürecin

adımları açıklanmıştır. CRISP modeli aşağıdaki altı aşamadan oluşur.

1. Problemin Tanımlanması
2. Veriyi Anlama
3. Veriyi Hazırlama
4. Model Kurma
5. Model Değerlendirilmesi ve Model Seçimi
6. Modelin Uygulaması

Şekil 4.1'de bu altı aşamanın görüntüsü görülmektedir [53].



Şekil 4.1. CRISP modeli.

4.1.1. Problemin Tanımlanması

Herhangi bir problemi çözmek için öncelikle problem çözme konusunda uzman olmak gereklidir. Bu problemin çeşidine göre değişebilir. Bir sayısal problemi çözmek için sayısal problemi çözme yetkinliğinin olmasının yanısıra problemi iyice anlamak gerekir. Ayrıca problem de iyi tanımlanmış olmalıdır. Benzer şekilde diğer problemleri de çözmede yetkinliğin yanısıra problemi iyi tanımlamalı ve sonuca öyle gidilmelidir.

Problemin amacı belirlenmeli, makine öğrenmesi ile problemin çözümünde katkısının ne olacağı belirlenmelidir. Kısaca bir yol haritası çıkarılmalıdır. Problem tanımlamadan sonraki aşama veri ön işleme aşamasıdır. Veri analizinin kararlı bir sonuç vermesi için veri ön işleme çok önemlidir. Veri ön işleme sürecini iyi planlamak ve analizden kararlı bir sonuç çıkarmak için veriyi iyi tanımak ve anlamak gereklidir. Veri önişleme Bölüm 4.1.3'teki veri hazırlama kısmında detaylı olarak anlatılacaktır.

4.1.2. Veriyi Anlama

Kaynaklarda enformasyon, bilgi ve bilgelik arasındaki bağdan söz edilerek veri açıklanır. Kısaca veri; enformasyon, bilgi ve bilgeliğin oluşturduğu zincirin ilk basamağıdır. Veri enformasyon haline işlenerek getirilir. Enformasyona yorum katılarak bilgi elde edilir ve bu bilgiyi de kullanarak bilgelik haline getirilmiş olur. Veri, enformasyon, bilgi ve bilgelikten oluşan veri zinciri Şekil 4.2'de görülmektedir [54].

Şekil 4.2. Veri, enformasyon, bilgi ve bilgelik zinciri.

Günümüzde verinin bilgiye dönüşümü veri madenciliği ve makine öğrenmesinin konusudur. Üzerinde çalışılan veri setini iyi analiz etmek gerekir. Öncelikle bu veri setinin kaynağı ve hangi alan ile alakalı olduğu da bilinmelidir. Veri kaynağı güvenilir olmadığı zaman veri ile elde edilen sonuçların faydalı olması tartışılabilir. Veri seti, çalışan kişinin alanı dışında olsa bile veri setinin değişkenlerine hakim olmalı ve bu değişkenlerin tiplerini bilmesi gerekmektedir. Örneğin sağlık alanında bir veri setinin analizi yapan biri değişkenlerin akademik olarak işlevlerini bilmese de o değişkenlerin tiplerini, aralarındaki korelasyonu bilmelidir. Veri seti üzerinde değişkenlerin özet bilgisi, ortalaması, frekans gibi tanımlayıcı değerleri ile değişkenlerin arasındaki ilişkileri grafiksel ve çeşitli yöntemlerle incelenmelidir. Bir hastanın meslek niteliğinin kategorik, ikili, tamsayı tiplerinden hangisi olduğunu bilmesi gerekir. Hastalığın belirtilerini bilmese de veri setinin analizini gerçekleştirdikten sonra bir uzmana danışarak analizi hakkında yorum alabilir. Kısaca veriyi anlama aşamasında veri setinin analize hazırlanması için yapılacak eksik değer tespiti ve tamamlanması, aykırı değer tespiti ve düzeltilmesi, veri dönüşümü, normalizasyon gibi işlemlere ihtiyaç duyulup duyulmadığına bakılır.

4.1.3. Veriyi Hazırlama

Veri hazırlama veri setine hâkimiyeti sağlandıktan sonraki ön işleme adımıdır. Bu adım analiz öncesi çok önemlidir. Veri madenciliği çalışmalarında en çok zaman harcanan veya harcanması gereken ön işleme adımıdır. Bu adımda veri seti incelendikten sonra eksik değer ve aykırı değer içermesi durumunun tespiti ve çözümlenmesi, veri bütünleştirme ve veri dönüşümü gerekli ise bu işlemlerin uygulanmasına yönelik adımlar yapılır. Bu bölümde tezde yapılan işlemlere yönelik kayıp veriler, aykırı değerler ve veri dönüşümü adımlarından bahsedilecektir.

4.1.3.1. *Kayıp Veriler*

Veri setinde bazı durumlarda bazı değişkenlere ait ölçülemeyen veriler mevcut olabilir. Böyle verilere kayıp veri (missing data) denir [54]. Eksik verileri tamamlama ve tespitte birçok yöntem kullanılmaktadır. Bunlardan bazılarını gösterilecek olursa [55]-[61]:

- Hedef değişkenin olmadığı veya aynı örneğe ait kayıp veri sayısı çok olması durumlarda kayıt silinebilir.
- Kayıp veriler analizde önemli değilse kayıp değerlerin hepsine, "N/A" atanabilir.
- Nümerik veri türüne sahip kayıp verilerin yerine numerik değişkenlerin ortalaması yazılabilir.
- Kategorik veri türüne sahip kayıp verilerin yerine en fazla tekrar eden değer yazılabilir.
- Veri madenciliği ve makine öğrenmesi algoritmaları ile maksimum beklenti (expectation-maximization) gibi bazı özel eksik veri tamamlama algoritmalarının kullanılmasıyla kayıp değerler doldurulabilir.
- Kayıp veriler regresyon analizi gibi yöntemlerle tamamlanabilir.

4.1.3.2. *Aykırı Veriler*

Aykırı veriler (outliers), uç noktalar olarak da adlandırılır. Aykırı veriler, yer aldığı örneklemde diğerlerinden farklı olan gözlemler olarak adlandırılır [62],[63]. Aykırı veriler aşağıda verilen şekillerdeki işlemler olursa karşımıza çıkabilir [62],[64]:

- Veriler girilirken veya kod hatası gibi işlemsel hatalardan kaynaklanabilir,
- Olağandışı bir durumda ortaya çıkabilir,

- Veri girişi yapanların da açıklayamadığı olağandışı bir gözlemle karşılaşılabilir,
- Düzgün veri girişi sağlansa bile değişken kaynaklı aykırı veriler oluşabilir.

Aykırı değerleri düzenlemek için kullanılan temel yöntemler aşağıda verilmiştir.

- Kutulama (Binning) ile
- Kümeleme ile
- Eğri Uydurma ile
- İnsan Denetimi ile

4.1.3.3. *Normalizasyon*

Veri setindeki nümerik değerlerin değişim aralıkları çok ise bu değerlerin normalize edilmesi gerekmektedir. Normalizasyon için min-max, z-score, ondalık ölçekleme gibi yöntemler kullanılır. Her veri seti için gerçekleştirilen bir süreç değildir. Normalizasyonun gerekli olup olmadığına karar vermek için veri setinin normalize edildikten veya edilmeden elde edilen sonuçlar ile tekrar değerlendirmek gerekir. Normalizasyon yöntemlerinden min-max Denklem (4.1)'de, z-score Denklem (4.2)'de, ondalık ölçekleme de Denklem (4.3)'te gösterilmiştir [29]:

Min-max:

$$v' = \frac{v - min_A}{max_A - min_A}\left(yeni_{max_A} - yeni\_min_A\right) + yeni_{min_A} \tag{4.1}$$

Z-score:

$$v' = \frac{v - \bar{A}}{\sigma_A} \tag{4.2}$$

Ondalık ölçekleme:

$$v' = \frac{v}{10^j} \tag{4.3}$$

Max($|v'| < 1$ koşulunu uyan en küçük tam sayı olsun)

4.1.3.4. *Veri Bütünleştirme*

Farklı veritabanlarından veya veri kaynaklarından elde edilen verileri bir arada kullanmak için yapılan işlemlerdendir. Bu çalışmada veriler belirli ve bir kaynaktan geldiği için veri bütünleştirme işlemi yapılmamıştır.

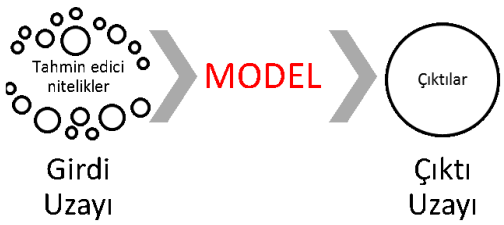
4.1.3.5. *Veri Dönüştürme*

Bu adım, veriler ile daha sağlıklı sonuçlar elde etmek için veri tipleri arasındaki dönüşümleri içeren adımdır. Aboneler arasındaki konuşma sürelerinin saat temelli belirli aralıklar ile gruplandırılması veya yaş değerlerinin belirli bir ölçekte aralıklandırılması veri dönüştürmeye örnektir.

4.1.4. Model Kurma

Model kurma adımı, veri seti hakkında tam yetki sağlandığı ve ön işleme tam anlamıyla yapıldıktan sonra geçilen adımdır. Bu adımda artık probleme ve öğrenme stratejisine uygun bir algoritma yardımıyla girdilerin istenilen çıktılara dönüştürülmesi amaçlanmaktadır. Veri madenciliği ve makine öğrenmesinde kullanılan Basit Bayes Algoritması, ID3, C4.5 ve Gini Karar Ağacı Algoritmaları, *k*-en yakın komşu algoritmaları ile ilgili bilgiler bu bölümde anlatılacaktır. Modeli oluşturmak problemi çözmek demek değildir. Elde edilen modellerin değerlendirilmek en iyisini seçmek gerekir. Çeşitli model değerlendirme yöntemi ve süreçleri ile en iyi model bulunur.

Tahmin edici değişkenlerden oluşan bir girdi uzayından ilgili değişkenler ile oluşturulan modelde, modellerin doğru işlem yapması veri madenciliği ve makine öğrenmesi algoritmaları sayesinde olmaktadır [67]. Şekil 4.3'te tahmin edici değişkenler ile model oluşturularak çıktı elde etme ile ilgili görüntü yer almaktadır [68].



Şekil 4.3. Makine öğrenmesi modeli.

Bütün modelleri temsil eden küme $M = \{ M1, M2, \ldots, Mn \}$ olsun. Model kurmadaki amaç bu modeller içinden performansı en iyi olanı seçmektir. Yani test verisinin, kurulan modellerden hangisinde iyi sonuç verdiğini belirlemektir. Bütün modelleri temsil eden M kümesi aşağıdakilerden oluşabilir [66]:

- Bir model farklı şekillerde de denenebilir. Örneğin, *k*-En Yakın Komşu Algoritmasıyla oluşturulan modelin birden çok *k* değeri ile denenmesi,
- Farklı modeller kullanılabilir. Örneğin, Bayes ağları, C4.5 karar ağacı gibi.

Bu tez kapsamında M modeller kümesini oluşturulurken kullanılan algoritmalar Bölüm 4.1.4.1, Bölüm 4.1.4.2 ve Bölüm 4.1.4.3'te anlatılmıştır.

4.1.4.1. *Naive Bayes Algoritması*

Naive Bayes algoritması adını İngiliz matematikçi Thomas Bayes'ten almıştır. Bayes algoritmaları istatiksel sınıflandırma teknikleri arasında yer alır ve istatiksel Bayes teoremine dayanır. Bayes sınıflandırıcı tahmin edici model olup basit uygulanabilir bir yöntemdir. Naive Bayes, bağımsız değişkenler ile hedef değişken arasındaki ilişkiyi gösteren bir sınıflandırma algoritmasıdır [67]. Naive Bayes, sürekli değere sahip veri ile değil kesikli veri ile çalışır. Bundan dolayı sürekli olan bağımlı veya bağımsız değişkenler kategorik veriye dönüştürülür. Örneğin, nümerik değerli yaş bağımsız değişkeni sürekli değerken, "16-25", "25-44","45-65","65+" gibi kesikli hale getirilmelidir [68].

X= { $x_1, x_2, x_3, \ldots, x_n$ } örnek kümesi, $C_1, C_2, C_3, \ldots, C_m$ Sınıf kümesi olsun. Sınıfı belirlenecek olan örnek,

$$P(X|C_i) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{4.4}$$

Denklem (4.4)'te görüldüğü gibi olasılıkları hesaplanır. Her bir sınıf için hesaplanan bu olasılıkların, olasılığı en büyük veren sınıfın veri örneği o sınıfa aittir [29].

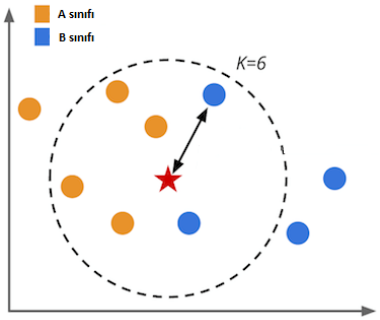
4.1.4.2. *k-En Yakın Komşu Algoritması*

k-en yakın komşu algoritmasındaki bütün veriler örüntü uzayında tutulur. *k*-en yakın komşu algoritması, bilinmeyen verilerin hangi sınıfa ait olduğunu bulmak için örüntü uzayına bakarak bilinmeyen veriye en yakın olan *k* örnekleri bulunur. Uzaklık bulunurken Öklid, Manhattan gibi uzaklık hesaplama yöntemleri ile hesaplanarak komşular arası uzaklık bulunur. Bilinmeyen veriler, *k* en yakın komşuya en fazla benzerlik gösterdiği sınıf değerine atanır. *k*-en yakın komşu algoritmasında sırasıyla aşağıdaki adımlar yapılır[69]:

- Verilen bir noktaya en yakın komşuların sayısı olan *k* belirlenir.
- Verilen nokta ile diğer bütün noktalar arasındaki uzaklıklar hesaplanır.
- Bir önceki işlemde hesap edilen uzaklıklara göre kayıtların sıralaması yapılır ve bunlar arasında en küçük *k* alınır.
- Seçilen kayıtlar bulunarak en fazla tekrar eden kategori seçilir.

- Seçilen kategori tahmin edilecek gözlemin kategorisi kabul edilir.

Şekil 4.4'te *k* en yakın komşu yapısı genel anlamda görülmektedir [70].



Şekil 4.4. *k*-en yakın komşu algoritma görüntüsü.

4.1.4.3. *Karar Ağacı Algoritması*

Karar ağacı, ağaç görünümünde tahmin edici bir algoritmadır. Kolay anlaşılır kurallar oluşturulabilir ağaç yapısı sayesinde çok yaygın kullanılan sınıflama tekniği olmuştur. Karar ağacı karar düğümleri, dallar ve yapraklardan oluşur. Bu görüntüsüyle bir ağaca benzemektedir. Karar ağacının düğümlerinde test ve dallara ayrılma işlemleri sırayla gerçekleşmektedir. Her bir dal tek başına sınıflamayı bitirebilir. Bir dalın uç kısmında sınıflama gerçekleşemiyorsa, o dalın sonucunda bir karar düğümü oluşur. Dalın sonunda belli bir sınıf oluşursa, o yapraktır. Karar ağacı işleme yukarıdan kök düğüm ile başlayıp aşağıya doğru yaprağa ulaşana kadar ardışık düğümleri takip ederek gerçekleşir [71].

Karar ağacının önemli sorunlarından birisi herhangi bir kökten hangi kıstasa göre dallanmanın olacağıdır. Genelde dallanma için entropiye dayalı algoritmalar kullanılır. Karar ağacına örnek ID3, C4.5 ve Gini gibi algoritmalar verilebilir.

ID3 karar ağacı algoritması; J.R. Quinlan tarafından 1986 yılında geliştirilen karar ağacı algoritmasıdır. Eğitim verilerinin ayrılma oranın ne kadar iyi olduğunu ölçmek için entropi ve bilgi kazancından faydalanılır. Entropi ve bilgi kazancı aşağıda anlatılmıştır [49],[29]:

D, eğitim kümesini, C_i;i=1, … ,m sınıf sayısını, $C_{i,D}$; D'deki C_i; sınıfının gözlem kümesini, $|D|$ ve $|C_{i,D}|$ de sırası ile D ve $C_{i,D}$'deki gözlemlerin sayısını göstersin. $P_i=\frac{C_{i,D}}{D}$, D'deki C_i sınıfına ait rasgele bir gözlemin olasılığını göstersin. D'deki bir gözlemi

sınıflandırmak için olası bilginin hesaplanması Denklem (4.5)'te gösterilmiştir:

$$Bilgi(D) = Entropi(D) = -\sum_{i=1}^{m} p_i * log_2(p_i) \quad (4.5)$$

Değişken A $\{ a_1, a_2, a_3, \dots, a_v \}$ gibi v farklı değerden oluşsun. A niteliği D'yi $\{ D_1, D_2, , D_3, \dots, D_v \}$ şeklinde v parçaya ayırabilir. (Dj'de A değişkeni aj değerini alan gözlemler yer alır); fakat amaç en iyi sınıflandırmayı bulmaktır. Bundan dolayı öncelikle iyi sınıflandırma için ayırma sonrasında hala ne kadar bir bilginin gerektiği Denklem (4.6)'daki gibi bulunur.

$$Bilgi_A(D) = \sum_{j=1}^{v} \frac{D_j}{|D|} * Bilgi(D_i) \quad (4.6)$$

Daha sonrasında da bilgi kazanımı Denklem (4.7)'de görüldüğü gibi Denklem (4.5) ve Denklem (4.6)'nın farkından hesaplanmaktadır.

$$Kazanım(D) = Bigi(D) - Bilgi_A(D) \quad (4.7)$$

Böylece karar ağacı oluşturulurken bölme işlemi için A değişkeni kullanıldığında ne kadar bilgi kazancı elde edileceği görülmüş olur [72].

C4.5 karar ağacı algoritması; bilgi kazancı performansını olumsuz etkilese de değişken seçmede farklı değeri olan değişkenler üzerine eğilimi vardır [29],[73],[74]. Bu durumu aşmak için ID3 algoritmasında yer alan bilgi kazancı yöntemi C4.5 algoritmasında yerini kazanım oranına bırakmıştır [29],[75]. Bundan dolayı C4.5, ID3'ün gelişmiş hali olarak söylenilebilir. Bilgi temelli bir ölçü olan kazanım oranı, C4.5 algoritmasında varsayılan olarak kullanılan bölme kriteridir [75]. Bilgi kazancına, ayrım bilgisi adı verilen normalizasyon uygulanır [29],[75]:

$$AyrımBilgisi_A(D) = \sum_{j=1}^{v} \frac{D_j}{|D|} * log_2(\frac{|D_j|}{D}) \quad (4.8)$$

Denklem (4.8)'deki $\text{AyrımBilgisi}_A(D)$, eğitim verisinin A değişkeninin aldığı v farklı değere bağlı olarak v parçaya ayrıldığında oluşturan bilgiyi temsil etmektedir. En yüksek kazanım oranını veren değişken ayrım yapılacak değişken olarak seçilir. Kazanım oranı Denklem (4.9)'daki gibi hesaplanır.

$$Kazanım\ Oranı(A) = \frac{Kazanım\ (A)}{AyrımBilgisi\ (A)} \quad (4.9)$$

C4.5 algoritması kategorik ve nümerik değişkenlere uygulanabilmektedir [76].

Gini karar ağacı algoritması; İkili bölünmelere dayalı bir diğer sınıflandırma yöntemi Gini algoritması olarak isimlendirilmektedir. Bölünme temelli olması CART algoritmasından geliştirilmesinden dolayıdır. Gini algoritması ile işlem yapılırken;

- Bütün değişkenlerin sürekli olduğu varsayılır.
- Her değişken için birden çok ayrımı olduğu varsayılır.
- Değişkenlerin ayrım noktaları için gruplama gibi diğer araçlara ihtiyaç duyulabilir.
- Kategorik değişkenler için kullanıldığında değiştirilmelidir.

Eğer bir T veri seti n farklı sınıftan N örnek içeriyorsa, gini indeks, gini(T) aşağıdaki Denklem (4.10)'daki gibi hesaplanır, pj, j sınıfının T içindeki izafi sıklığını ifade eder

$$gini\ (T) = 1 - \sum_{j=1}^{n} p_j^2 \tag{4.10}$$

Eğer T veri seti T1 ve T2 olarak sırasıyla N1 ve N2 büyüklüğünde ikiye ayrılırsa, ayrılan veri için gini indeksi Denklem (4.11)'de görüldüğü gibidir.

$$gini_{ayrım}T = \frac{N_1}{N} gini(T_1) + \frac{N_2}{N} gini(T_2) \tag{4.11}$$

En düşük gini değerini veren ayrıma sahip degişken seçilir [77].

4.1.5. Model Performans Değerlendirme ve Seçim Süreci

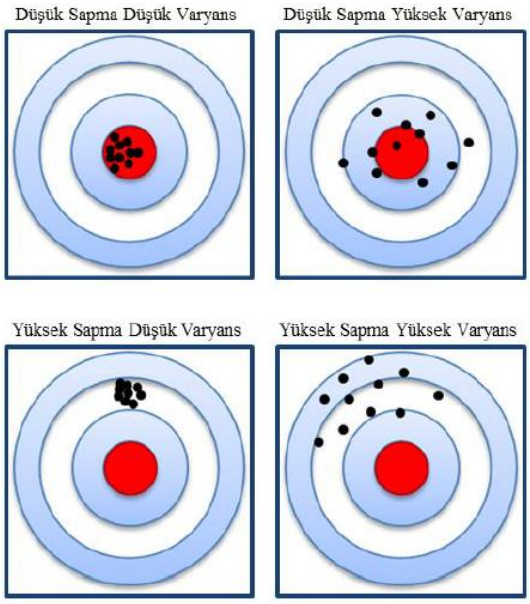
Model değerlendirme ve seçim sürecinde, Bölüm 4.1.4 model kurma adımında anlatılan algoritmalar ile kurulan modellerin performansları değerlendirilerek en iyi model belirlenir. Çeşitli model değerlendirilmesi ve seçimi yöntemleri bu bölümde anlatılacaktır.

4.1.5.1. *Performans Değerlendirme ve Model Seçimi*

Bir modelin değerlendirilmesi ve model seçimi, kullanılan veri madenciliği ve makine öğrenme algoritmasının yanı sıra hatalı sınıflandırılmasına veya eğitim ile test verilerinin büyüklüğüne de bağlıdır. Bu çalışmada hold-out ve k-kat çapraz geçerleme model değerlendirilmesi ve seçimi yöntemleri kullanılmıştır. Bu bölümde hold-out ve k-kat çapraz geçerleme model değerlendirilmesi ve seçimi yöntemleri kısaca anlatılacaktır[72].

Hold-out: Hold-out yönteminde veri seti eğitim ve test verisi olarak ikiye ayrılır. İkiye ayrılan veri setinden eğitim veri seti, eğitim sürecinde model oluşturulma için seçilen sınıflandırma algoritmasının eğitiminde kullanılmaktadır. Model için en iyi değerler ile en iyi performans kıstası burada belirlenir. Veri setinin ikinci parçası olan test veri seti model performansını test için kullanır. Elde az sayıda veri olma durumu ile eğitim ve test verisinin tek seferlik ayrımı bu yöntemin dezavantajlarıdır [72].

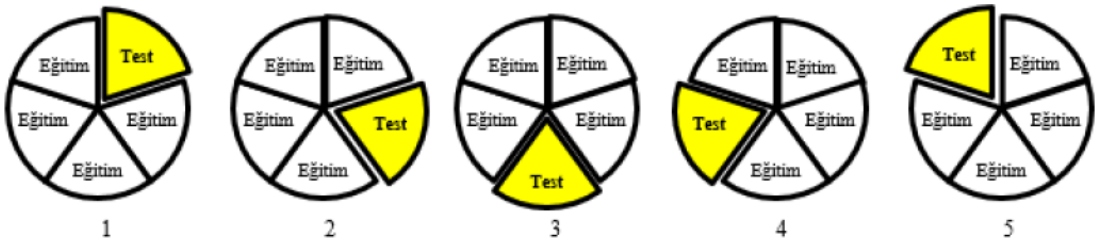
Çapraz Geçerleme: K-kat çapraz geçerleme (cross validation) olarak bilinen bu yöntem iki farklı şekilde kullanılır. Birincisi bir algoritma ile kurulan model için ikincisi ise iki veya daha fazla algoritmanın performansını ölçmede kullanılmasıdır [78]. K-kat çapraz geçerlemede veri seti k adet eşit parçaya ayrılarak her defada bir tanesi test, k-1 tanesi de eğitim için kulanılır. Sonuçta k adet hata oranı ile tüm tahminin hatasını hesaplamak için hataların ortalaması alınır. Kullanılan k-kat sayısının büyük olması halinde tahmin daha doğru olacağından gerçek hata tahmininin sapması küçük; varyans ve hesaplama zamanı ise büyüktür [79],[80]. Yani k-kat sayısı küçük seçilirse, tahminin sapması küçük ve sapması gerçek hata tahmininden büyük olacaktır [79],[81]. Sapma kaynaklı hata, gerçek değer ile tahmin edilen değerin arasındaki farka eşittir. Varyans kaynaklı hata, verilen bir nokta için model tahminindeki değişikliktir. Şekil 4.5'te sapma ve varyans arasındaki ilişki görülmektedir [82].



Şekil 4.5. Sapma ve varyans ilişkisi.

k'nın kullanılan en yaygın değeri 10'dur [81],[83]-[86]. Kohavi [87] model seçimi için 10-kat çapraz geçerlemeyi kendi çalışmalarında uygulamış ve bu çalışmasının

sonucunda 10-kat çapraz geçerlemeyi önermektedir. 5-kat çapraz geçerlemenin de Şekil 4.6'da görüntüsü görülmektedir [88].



Şekil 4.6. 5 - Kat çapraz geçerleme.

Bu çalışmada model değerlendirilmesi ve seçimi yöntemlerinden hold-out ile 4 kat çapraz geçerleme, 5 kat geçerleme ve 10 kat çapraz geçerleme kullanılmıştır.

4.1.5.2. *Model Performans Değerlendirme Ölçütleri*

Sınıflandırma algoritmaları ile oluşturulan modelin neye göre değerlendirileceği çeşitli yöntemlerle yapılır. Bunlardan biri kontenjans tablosu (confusion matrix) oluşturulmaya dayanır [89]. Gerçek değerler ile model aracılığı ile tahmin edilen değerler Tablo 2.3'te gösterilmiştir. Tablo 2.3'te göre sınıflandırma algoritmalarının değerlendirilmesine yönelik performans değerlendirme ölçütleri aşağıda verilmiştir [90]-[97].

Tablo 4.1. Kontenjans tablosu.

		Gerçek		
		Pozitif	Negatif	Toplam
Tahmin	Pozitif	Doğru Pozitif (dp)	Yanlış Pozitif (yp)	tPoz
	Negatif	Yanlış Negatif (yn)	Doğru Negatif (dn)	tNeg
	Toplam	poz	neg	m

Tablo 4.1'e göre sınıflandırma algoritmaları ile oluşturulan modelin doğruluğu Denklem (4.12)'de, hata oranı ise Denklem (4.13)'te görüldüğü şekilde hesaplanmaktadır [98]:

$$Doğruluk(ACC) = \frac{d_p+d_n}{m} \quad (4.12)$$

$$Hata\ oranı\ (ERR) = 1 - ACC \quad (4.13)$$

Doğru sınıflandırılan pozitif örneklerin toplam pozitif örneklerine oranına duyarlılık denir. Denklem (4.14)'te duyarlılık formülü görülmektedir [98].

$$Duyarlılık\ (TPR) = \frac{d_p}{poz} = \frac{d_p}{d_p+y_n} \quad (4.14)$$

Doğru sınıflandırılan negatif örneklerin toplam negatif örnek sayısına oranına belirleyicilik denir. Denklem (4.15)'te belirleyicilik formülü görülmektedir [98].

$$Belirleyicilik\ (SPC) = \frac{d_n}{neg} = \frac{d_n}{d_n+y_p} \quad (4.15)$$

Gerçekte negatif olan ancak pozitif olarak sınıflandırılmış örneklerin, tüm negatif etiketli örneklere oranı yanlış pozitif oranı denir. Gerçekte pozitif olan ancak negatif olarak sınıflandırılmış örneklerin, tüm pozitif etiketli örneklere oranı da yanlış negatif oranı olarak adlandırılmaktadır. Sırasıyla Denklem (4.16) ve Denklem (4.17)'de yanlış pozitif ve yanlış negatif oranları görülmektedir [98].

$$Yanlış\ pozitif\ oranı\ (FPR) = \frac{y_p}{neg} = \frac{y_p}{y_p+d_n} \quad (4.16)$$

$$Yanlış\ negatif\ oranı(FNR) = 1 - TPR = \frac{y_n}{poz} = \frac{y_n}{y_n+d_p} \quad (4.17)$$

Doğru sınıflandırılan pozitif örneklerin toplam pozitif tahmin edilen örneklere oranına kesinlik ya da pozitif öngörü değeri denir. Doğru sınıflandırılan negatif sınıf etiketine sahip örneklerin toplam negatif tahmin edilen örneklere oranına negatif öngörü değeri denir. Sırasıyla Denklem (4.18) ve Denklem (4.19)'da pozitif öngörü değeri ve negatif öngörü değeri görülmektedir [98].

$$Pozitif\ öngörü\ değeri\ (PPV) = \frac{d_p}{tPoz} = \frac{d_p}{d_p+y_p} \quad (4.18)$$

$$Negatif\ \ öngörü\ değeri\ (NPV) = \frac{d_n}{tNeg} = \frac{d_n}{d_n+y_n} \quad (4.19)$$

F-ölçüsü (F-measure), kesinlik ve duyarlılık performans değerlendirme ölçütlerinin

harmonik ortalamasıdır. Denklem (4.20)'de F-ölçüsü formülü görülmektedir [98].

$$F - \text{Ölçüsü } (F) = \frac{2*PPV*TPR}{PPV+TPR} \quad (4.20)$$

Tahmin sonucunun gerçekte pozitif sınıf etiketinin varlığında pozitif çıkma olasılığının, negatif sınıf etiketi varlığında pozitif çıkma olasılığına oranına pozitif olabilirlik oranı (positive likelihood ratio) denir. Tahmin sonucunun gerçekte pozitif sınıf etiketinin varlığında negatif çıkma olasılığının, negatif sınıf etiketi varlığında negatif çıkma olasılığına oranına negatif olabilirlik oranı denmektedir. $L+$ ne kadar yüksekse gerçekte pozitif sınıf etiketi almış örnekler o kadar iyi sınıflandırılmakta; $L-$ ne kadar küçükse, gerçekte negatif sınıf etiketine sahip örnekler o kadar iyi ayrılmaktadır. Sırasıyla Denklem (4.21) ve Denklem (4.22)'de pozitif olabilirlik oranı ve negatif olabilirlik oranı görülmektedir [98].

$$Pozitif\ olabilirlik\ oranı\ (LR+) = \frac{TPR}{FPR} = \frac{TPR}{1-SPC} \quad (4.21)$$

$$Negatif\ olabilirlik\ oranı\ (LR-) = \frac{FNR}{TNR} = \frac{1-TPR}{SPC} \quad (4.22)$$

Pozitif ve negatif olabilirlik oranlarını içine alan Tanısal Üstünlük Oranı (Diagnostic Odds Ratio), tahmin edilen pozitif sınıfın üstünlüğünün, negatif sınıfa üstünlüğüne oranına denir [94]. Denklem (4.23)'te Tanısal Üstünlük Oranı görülmektedir [98].

$$Tanısal\ üstünlük\ oranı\ (DOR) = \frac{LR+}{LR-} \quad (4.23)$$

4.1.6. Modelin Uygulaması

Model seçimine geçmeden önce problem tanımı iyi yapılmalı ve kullanılacak veri seti iyi anlaşılmalıdır. Veri setlerinde eksik değer ve aykırı değer varsa bunlar çözülerek analize hazır hale getirilmelidir. Veri seti üzerinde kurulan en iyi model, performans değerlendirilmesi ve seçimi yöntemleri ile performans değerlendirme ölçütleri sayesinde seçilir. Makine öğrenmesi sürecindeki adımlar tamamlandıktan sonra bir eksiklik yoksa modeli uygulamaya geçilir.

Bu tez kapsamında veri madenciliği ve makine öğrenmesi ile yapılan modellere yönelik Bölüm 4.1.5.1'de anlatılan model değerlendirilmesi ve seçimi yöntemleri ve Bölüm 4.1.5.2'de anlatılan model performans değerlendirme ölçütleri göz önünde

bulundurularak modellerden hangisinin iyi sonuç verdiği görülerek en iyi model seçilmiştir. Seçilen en iyi modelin dinamik şekilde de nasıl bir sonuç vereceğine yönelik R paketlerinden Shiny ile de web ortamında bir uygulama geliştirilmiştir.

5. MATERYAL VE YÖNTEM

5.1. R PROGRAMIYLA MÜŞTERİ AYRILMA TAHMİN UYGULAMASI

Bu bölümünde makine öğrenmesi ve veri madenciliği teknikleri kullanılarak müşteri kayıp analizine yönelik uygulamaya yer verilmiştir. Müşteri kayıp analizi üzerine yapılan çalışmada Bölüm 4.1'de verilmiş olan makine öğrenme süreci adımlarını içeren CRISP modeli izlenerek gerçekleştirilen uygulama adım adım anlatılmıştır.

5.2. PROBLEMİN TANIMLANMASI

Abone tabanlı çoğu sektörde olduğu gibi telekomünikasyon sektöründe de en önemli problem müşteri kaybıdır. Abone tabanlı kurumlar mevcut hangi abonelerinin kaybedebileceklerini bilmedikleri için abone kaybını engellemek veya en az zararla atlatmak için çeşitli adımlar atabilirler. Bu çalışmanın ana amacını telekomünikasyon müşteri veri seti üzerinde abonelerin ayrılıp ayrılmayacağını tahmin oluşturmaktadır. Müşteri kaybını tahmin etmek için sınıflandırma algoritmaları kullanılarak modeller oluşturulmuştur. Bu modellerden en iyi sonuç veren model belirlenmiştir.

5.3. VERİYİ ANLAMA

Çalışma yapılan Telekomünikasyon veri setinde 8000 müşteri verisi ve 171 değişken bulunmaktadır. Telekomünikasyon veri setine ilişkin tüm değişkenler, gösterim biçimleri ve türleri Tablo 5.1'de verilmektedir. Ayrıca Tablo 5.1'de ingilizce olan değişken değerlerinin türkçe açıklamaları da verilmiştir. Tablo 5.1 incelendiğinde telekomünikasyon veri setinin değişkenlerinin sayısal, kategorik ve ikili veri tipindeki değişkenlere sahip olduğu görülmektedir. Bu veri setindeki telekomünikasyon verilerinin 26 tanesi aylık değişkenler diğer geri kalanlar ise müşterilerin kişisel değişken verileridir. Aylık değişkenler 6 aylık değerler olup veri setinde toplam 156 tane aylık değişken yer almaktadır. Aylık değişkenler dışında kalan 15 tane değişken müşterilerin kişisel verilerini tutan değişkenlerdir.

Tablo 5.1. Telekomünikasyon veri setine ilişkin tüm değişkenler, gösterim biçimleri ve türleri.

TAHMİN İÇİN KULLANILAN DEĞİŞKENLER			
	DEĞİŞKEN	AÇIKLAMASI	VERİ TİPİ
1	number	Abone telefon numarası	NUMERİK
2	gender_flag	Cinsiyet	NUMERİK
3	age	Yaş	NUMERİK
4	age_of_line	Müşterililik süresi	NUMERİK
5	tariff_type	Tarife tipi (Postpaid-faturalı, Prepaid-kontörlü)	NOMİNAL
6	device_type	Cihaz tipi, Smartphone, Laptop vs.	NOMİNAL
7	last_reload_date	Yükleme yapılan son tarih (Faturalı aboneler için)	NUMERİK
8	last_reload_amount	En son yapılan yükleme miktarı (Prepaid-Kontörlü aboneler için)	NUMERİK
9	expiry_date	Son kullanma tarihi	NUMERİK
10	hotline_date	Hat başlangıç tarihi	NUMERİK
11	crm_segment	CRM segmenti	NOMİNAL
12	value_segment	Müşteri değer türü	NOMİNAL
13	lifestyle_segment	Tarife Türü	NOMİNAL
14-19	mmo_count_07_12	Aylık kendi aboneleriyle konuşma sayısı (arama)	NUMERIK
20-25	mmo_duration_07_12	Aylık kendi aboneleriyle konuşma süresi (arama)	NUMERIK
26-31	mmo_non_count_07_12	Aylık diğer operatör aboneleriyle konuşma sayısı (arama)	NUMERIK
32-37	mmo_non_duration_0 7_12	Aylık diğer operatör aboneleriyle konuşma süresi (arama)	NUMERIK
38-43	mmt_count_07_12	Aylık kendi aboneleriyle konuşma sayısı (aranma)	NUMERIK
44-49	mmt_duration_07_12	Aylık kendi aboneleriyle konuşma süresi (aranma)	NUMERIK
50-55	mmt_non_count_07_12	Aylık diğer operatör aboneleriyle konuşma sayısı (aranma)	NUMERIK
56-61	mmt_non duration_0 7_12	Aylık diğer operatör aboneleriyle konuşma süresi (aranma)	NUMERIK
62-67	mmo_total_count_07_12	Aylık kendi operatörleri ile konuşma sayısı (arama)	NUMERIK
68-73	mmo_total_duration_07_12	Aylık kendi operatörleri ile konuşma süresi (arama)	NUMERIK
74-79	mmt_total_count_07_12	Aylık diğer operatörleri ile konuşma sayısı (aranma)	NUMERIK

Tablo 5.1 (devam). Telekomünikasyon veri setine ilişkin tüm değişkenler, gösterim biçimleri ve türleri.

80-85	mmt_total_duration_07_12	Aylık diğer operatörleri ile konuşma süresi (aranma)	NUMERIK
86-91	non_percent_07_12	Aylık konuştuğu kişilerden diğer operatörden olanların kendi abonesi olanlara oranı (arama+aranma)	NOMİNAL
92-97	gprs_duration_07_12	Aylık gprs kullanım süresi	NUMERIK
98-103	call_distinct_07_12	Aylık konuşulan farklı kişi sayısı	NUMERIK
104	is_home_changed	Aylık olarak ev lokasyonu değişme durumu	NUMERIK
105	is_work_changed	Aylık olarak iş lokasyonu değişme durumu	NUMERIK
106-111	msmo_count_07_12	Aylık sms atma sayısı	NUMERIK
112-122	home_lat_07_12	Aylık ev lokasyonu (enlem)	NOMİNAL
113-123	home_lon-07_12	Aylık ev lokasyonu (boylam)	NOMİNAL
124-129	cd_total_07_12	Aylık Call-Drop (Çağrı bırakma) Sayıları	NUMERIK
130-135	cd_distinct_cell_07_12	Aylık call-drop (Çağrı bırakma) yaşadığı farklı baz istasyonu sayısı	NUMERIK
136-141	dealer_dist_07_12	Aylık ev lokasyonunun en yakın bayisine olan uzaklığı	NOMİNAL
142-147	callcenter_count_07_12	Aylık Şikayet sayısı	NUMERIK
148-153	payment_07_12	Aylık ödeme miktarı, postpaid (faturalı) için fatura, prepaid (kontörlü) için toplam reload	NOMİNAL
154-159	unpaid_07_12	Zamanında ödenmeyen fatura sayısı	NUMERIK
160-165	payment_type_07_12	Fatura ödeme şekli	NOMİNAL
HEDEF DEĞİŞKEN			
166-171	churn_2013_07_12	Müşteri ayrılma durum bilgisi (churn)	İKİLİ

Veri anlama aşamasında veri seti ile ilgili Şekil 5.1'deki gibi çeşitli görseller ile veri seti hakkında bilgi edinilebilmektedir. Şekil 5.1'deki veri seti özet bilgisi incelendiğinde nümerik değerlerin minimum, maksimum, medyan, ortalama, 1. ve 3. kartil değerlerinin verildiği görülmektedir. Kategorik değişkenlerin aldığı değerlere ait frekanslar da görülmektedir. Ayrıca bu özet bilgide veri seti üzerinde N/A atanmış eksik verilerin ve age (yaş) niteliğinin alt sınırının aykırı değerde olması gibi modellemeye uygun olmayan durumlar görülmektedir. Eksik verilerin tamamlandığı, aykırı verilerin analize hazır hale getirildiği adım olan veri hazırlama adımından sonra, veri seti modellemeye en uygun haline gelmektedir. Veri hazırlama adımından sonra veri setinin eksik verilerin tamamlandığı, aykırı verilerin analize hazır hale getirildiği hali gösterilecektir.

```
> summary(telekomNA)
     number       gender_flag          age          age_of_line      tariff_type     device_type
 Min.   :    2   Min.   :0.000   Min.   :-985.00   Min.   :   0.0   Min.   :1.000   Min.   : 0.000
 1st Qu.: 2513   1st Qu.:1.000   1st Qu.:  28.00   1st Qu.: 266.0   1st Qu.:1.000   1st Qu.: 3.000
 Median : 5014   Median :1.000   Median :  36.00   Median : 705.0   Median :1.000   Median : 3.000
 Mean   : 5013   Mean   :1.277   Mean   :  37.02   Mean   : 984.4   Mean   :1.317   Mean   : 5.664
 3rd Qu.: 7505   3rd Qu.:2.000   3rd Qu.:  46.00   3rd Qu.:1476.0   3rd Qu.:2.000   3rd Qu.:10.000
 Max.   :10000   Max.   :2.000   Max.   : 114.00   Max.   :4894.0   Max.   :2.000   Max.   :15.000

 last_reload_date   last_reload_amount  expiry_date        hotline_date       crm_segment
 Min.   :19000101   0      :2330       Min.   :19000101   Min.   :19000101   Min.   :0.000
 1st Qu.:19000101   20     :1596       1st Qu.:19000101   1st Qu.:19000101   1st Qu.:0.000
 Median :20131228   10     :1228       Median :20141003   Median :20140701   Median :1.000
 Mean   :19805744   12     : 774       Mean   :19750913   Mean   :19747007   Mean   :1.018
 3rd Qu.:20140523   30     : 748       3rd Qu.:20150217   3rd Qu.:20141119   3rd Qu.:1.000
 Max.   :20140614   25     : 617       Max.   :20150311   Max.   :20141211   Max.   :4.000
                    (Other): 707
 value_segment lifestyle_segment mmo_duration_07  mmo_duration_08    mmo_count_10      mmo_count_11
 Min.   :0.0   Min.   :0.000     Min.   :   0.0   Min.   :   0.00   Min.   :   0.00   Min.   :   0.00
 1st Qu.:3.0   1st Qu.:2.000     1st Qu.:   1.0   1st Qu.:   1.00   1st Qu.:   1.00   1st Qu.:   1.00
 Median :5.0   Median :2.000     Median :  17.0   Median :  17.00   Median :  19.00   Median :  18.00
 Mean   :3.7   Mean   :2.163     Mean   :  44.2   Mean   :  43.67   Mean   :  45.09   Mean   :  44.26
 3rd Qu.:5.0   3rd Qu.:3.000     3rd Qu.:  59.0   3rd Qu.:  59.00   3rd Qu.:  62.00   3rd Qu.:  61.00
 Max.   :7.0   Max.   :4.000     Max.   :1150.0   Max.   :1100.00   Max.   :1027.00   Max.   :1033.00
                                 NA's   :2649     NA's   :2649      NA's   :2649      NA's   :2649
  mmo_count_12   mmo_duration_07.1 mmo_duration_08.1 mmo_duration_09  mmo_duration_10
 Min.   :  0.0   Min.   :     0    Min.   :     0    Min.   :     0   Min.   :     0
 1st Qu.:  2.0   1st Qu.:     0    1st Qu.:     4    1st Qu.:     2   1st Qu.:     9
 Median : 19.0   Median :  1825    Median :  2282    Median :  2140   Median :  2369
 Mean   : 44.9   Mean   : 10091    Mean   : 10591    Mean   : 10576   Mean   : 10706
 3rd Qu.: 62.0   3rd Qu.:  8760    3rd Qu.:  9819    3rd Qu.: 10074   3rd Qu.:  9963
 Max.   :960.0   Max.   :646633    Max.   :580523    Max.   :796495   Max.   :713371
 NA's   :2649    NA's   :2649      NA's   :2649      NA's   :2649     NA's   :2649
 mmo_duration_11    mmo_duration_12     mmt_count_07      mmt_count_08      mmt_count_09
 Min.   :     0.0   Min.   :     0.0   Min.   :   0.00   Min.   :   0.00   Min.   :   0.00
 1st Qu.:    18.5   1st Qu.:    44.5   1st Qu.:   1.00   1st Qu.:   2.00   1st Qu.:   2.00
 Median :  2350.0   Median :  2601.0   Median :  25.00   Median :  29.00   Median :  27.00
 Mean   : 11218.8   Mean   : 11238.9   Mean   :  63.65   Mean   :  70.06   Mean   :  65.31
 3rd Qu.: 10497.5   3rd Qu.: 10620.0   3rd Qu.:  82.00   3rd Qu.:  95.00   3rd Qu.:  86.00
 Max.   :539675.0   Max.   :641843.0   Max.   :1155.00   Max.   :1396.00   Max.   :1487.00
 NA's   :2649       NA's   :2649       NA's   :2407      NA's   :2407      NA's   :2407
  mmt_count_10      mmt_count_11      mmt_count_12     mmt_duration_07  mmt_duration_08
 Min.   :   0.00   Min.   :   0.00   Min.   :   0.00   Min.   :     0   Min.   :     0
 1st Qu.:   2.00   1st Qu.:   3.00   1st Qu.:   3.00   1st Qu.:    35   1st Qu.:    65
 Median :  31.00   Median :  29.00   Median :  32.00   Median :  2513   Median :  2928
 Mean   :  69.04   Mean   :  66.09   Mean   :  67.47   Mean   : 10727   Mean   : 11504
 3rd Qu.:  93.00   3rd Qu.:  89.00   3rd Qu.:  90.00   3rd Qu.: 13538   3rd Qu.: 15358
 Max.   :1532.00   Max.   :1678.00   Max.   :2301.00   Max.   :322551   Max.   :344747
 NA's   :2407      NA's   :2407      NA's   :2407      NA's   :2407     NA's   :2407
```

Şekil 5.1. Veri önişleme öncesi veri setinin özet bilgisi.

```
cd_distinct_cell_10 cd_distinct_cell_11 cd_distinct_cell_12 dealer_dist_07 dealer_dist_08
Min.   : 0.000      Min.   : 0.000      Min.   : 0.000             :2189          :2189
1st Qu.: 0.000      1st Qu.: 0.000      1st Qu.: 0.000      0      :1129   0      :1103
Median : 0.000      Median : 0.000      Median : 0.000      0,1    :  34   0,1    :  33
Mean   : 0.993      Mean   : 0.967      Mean   : 0.937      0,49   :  29   0,49   :  28
3rd Qu.: 1.000      3rd Qu.: 1.000      3rd Qu.: 1.000      0,51   :  28   0,54   :  27
Max.   :20.000      Max.   :15.000      Max.   :14.000      0,54   :  28   0,51   :  25
NA's   :6058        NA's   :6058        NA's   :6058        (Other):4563   (Other):4595
dealer_dist_09 dealer_dist_10 dealer_dist_11 dealer_dist_12 callcenter_count_07 callcenter_count_08
       :2189          :2189          :2189          :2189   Min.   : 0.00       Min.   : 0.000
0      :1093   0      :1061   0      : 996   0      :1005   1st Qu.: 0.00       1st Qu.: 0.000
0,1    :  31   0,1    :  31   0,49   :  36   0,59   :  37   Median : 0.00       Median : 0.000
0,49   :  31   0,59   :  31   0,68   :  34   0,83   :  35   Mean   : 0.99       Mean   : 0.983
0,69   :  30   0,83   :  30   0,75   :  33   0,4    :  34   3rd Qu.: 1.00       3rd Qu.: 1.000
0,59   :  29   0,4    :  29   0,1    :  32   0,1    :  33   Max.   :57.00       Max.   :73.000
(Other):4597   (Other):4629   (Other):4680   (Other):4667   NA's   :5283        NA's   :5283
callcenter_count_09 callcenter_count_10 callcenter_count_11 callcenter_count_12   payment_07
Min.   :  0.000     Min.   :  0.000     Min.   :  0.000     Min.   :  0.000            :2297
1st Qu.:  0.000     1st Qu.:  0.000     1st Qu.:  0.000     1st Qu.:  0.000     0      :1859
Median :  0.000     Median :  0.000     Median :  0.000     Median :  0.000     20     : 708
Mean   :  0.946     Mean   :  0.982     Mean   :  1.308     Mean   :  1.131     10     : 434
3rd Qu.:  1.000     3rd Qu.:  1.000     3rd Qu.:  1.000     3rd Qu.:  0.000     30     : 302
Max.   :132.000     Max.   :127.000     Max.   :208.000     Max.   :121.000     25     : 154
NA's   :5283        NA's   :5283        NA's   :5283        NA's   :5283        (Other):2246
  payment_08     payment_09     payment_10     payment_11     payment_12     unpaid_07
       :2297          :2297          :2297          :2297          :2297   Min.   :0.000
0      :1866   0      :1811   0      :1774   0      :1801   0      :1968   1st Qu.:0.000
20     : 729   20     : 672   20     : 645   20     : 643   20     : 719   Median :0.000
10     : 421   10     : 441   10     : 423   10     : 457   10     : 435   Mean   :0.267
30     : 325   30     : 325   30     : 363   30     : 353   30     : 356   3rd Qu.:0.750
25     : 146   25     : 172   25     : 156   25     : 162   25     : 158   Max.   :2.000
(Other):2216   (Other):2282   (Other):2342   (Other):2287   (Other):2067   NA's   :7910
  unpaid_08       unpaid_09       unpaid_10       unpaid_11       unpaid_12     payment_type_07
Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   : 0.00
1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.: 2.00
Median :0.000   Median :0.000   Median :0.000   Median :0.000   Median :0.000   Median :17.00
Mean   :0.267   Mean   :0.267   Mean   :0.267   Mean   :0.267   Mean   :0.267   Mean   :12.87
3rd Qu.:0.750   3rd Qu.:0.750   3rd Qu.:0.750   3rd Qu.:0.750   3rd Qu.:0.750   3rd Qu.:19.00
Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :29.00
NA's   :7910    NA's   :7910    NA's   :7910    NA's   :7910    NA's   :7910    NA's   :6063
payment_type_08 payment_type_09 payment_type_10 payment_type_11 payment_type_12 churn_2013_07
Min.   : 0.00   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00   Min.   :0.00000
1st Qu.: 2.00   1st Qu.: 2.00   1st Qu.: 2.00   1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.:0.00000
Median :17.00   Median :17.00   Median :17.00   Median : 9.00   Median : 9.00   Median :0.00000
Mean   :12.71   Mean   :13.02   Mean   :13.17   Mean   :11.08   Mean   :11.08   Mean   :0.01787
3rd Qu.:19.00   3rd Qu.:19.00   3rd Qu.:19.00   3rd Qu.:19.00   3rd Qu.:19.00   3rd Qu.:0.00000
Max.   :29.00   Max.   :29.00   Max.   :29.00   Max.   :29.00   Max.   :29.00   Max.   :1.00000
NA's   :6063    NA's   :6063    NA's   :6063    NA's   :6063    NA's   :6063
churn_2013_08    churn_2013_09    churn_2013_10   churn_2013_11   churn_2013_12
Min.   :0.00000   Min.   :0.00000   Min.   :0.000   Min.   :0.000   Min.   :0.00000
1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.00000
Median :0.00000   Median :0.00000   Median :0.000   Median :0.000   Median :0.00000
Mean   :0.01937   Mean   :0.01862   Mean   :0.017   Mean   :0.018   Mean   :0.01975
3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:0.00000
Max.   :1.00000   Max.   :1.00000   Max.   :1.000   Max.   :1.000   Max.   :1.00000
```

Şekil 5.1 (devam). Veri önişleme öncesi veri setinin özet bilgisi.

Şekil 5.2'de veri önişlemeye tabi tutulmamış telekomünikasyon veri setine ilişkin veri türleri ve genel dağılımı görülmektedir. Ayrıca Şekil 5.2'de N/A değerleri de görülmektedir.

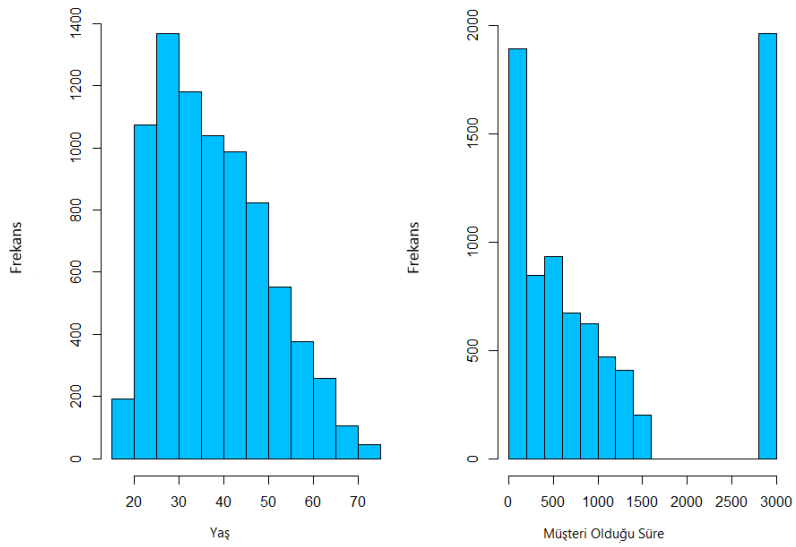
```
> str(telekomNA)
'data.frame':   8000 obs. of  171 variables:
 $ number                : int  9255 1562 1671 6088 6670 5934 8830 7946 3509 2003 ...
 $ gender_flag           : int  1 1 1 2 2 1 2 1 1 1 ...
 $ age                   : int  50 37 29 47 22 55 31 37 53 49 ...
 $ age_of_line           : int  2031 44 1807 2091 37 902 1080 814 529 2139 ...
 $ tariff_type           : int  1 2 1 1 1 1 2 1 1 1 ...
 $ device_type           : int  10 10 0 10 10 3 10 3 3 3 ...
 $ last_reload_date      : int  20140516 19000101 19000101 20140606 20140530 20140415 20140303 20140611 20130918 19000101 ...
 $ last_reload_amount    : Factor w/ 31 levels "0","10","10,75",..: 23 1 1 14 14 14 17 14 2 1 ...
 $ expiry_date           : int  20150210 19000101 19000101 20150303 20150224 20150110 19000101 20150308 20140616 19000101 ...
 $ hotline_date          : int  20141112 19000101 19000101 20141203 20141126 20141012 19000101 20141208 20140615 19000101 ...
 $ crm_segment           : int  1 0 1 0 0 0 1 0 1 1 ...
 $ value_segment         : int  5 0 5 4 3 4 5 5 5 0 ...
 $ lifestyle_segment     : int  4 0 2 3 1 1 2 2 2 2 ...
 $ mmo_duration_07       : num  13 NA NA NA NA NA 114 22 12 NA ...
 $ mmo_duration_08       : num  15 NA NA NA NA NA 44 9 7 NA ...
 $ mmo_count_10          : num  9 NA NA NA NA NA 82 4 0 NA ...
 $ mmo_count_11          : num  134 NA NA NA NA NA 92 199 0 NA ...
 $ mmo_count_12          : num  223 NA NA NA NA NA 248 97 0 NA ...
 $ mmo_duration_07.1     : num  108 NA NA NA NA ...
 $ mmo_duration_08.1     : num  402 NA NA NA NA ...
 $ mmo_duration_09       : num  1586 NA NA NA NA ...
 $ mmo_duration_10       : num  755 NA NA NA NA ...
 $ mmo_duration_11       : num  13676 NA NA NA NA ...
 $ mmo_duration_12       : num  43409 NA NA NA NA ...
 $ mmt_count_07          : num  3 NA NA NA NA NA 136 23 0 NA ...
 $ mmt_count_08          : num  7 NA NA NA NA NA 115 69 198 NA ...
 $ mmt_count_09          : num  4 NA NA NA NA NA 102 48 193 NA ...
 $ mmt_count_10          : num  7 NA NA NA NA NA 101 42 0 NA ...
 $ mmt_count_11          : num  50 NA NA NA NA NA 135 36 0 NA ...
 $ mmt_count_12          : num  206 NA NA NA NA NA 99 73 0 NA ...
 $ mmt_duration_08       : num  10 NA NA NA NA ...
 $ mmo_non_duration_08   : num  98 NA NA NA NA ...
 $ mmt_duration_09       : num  290 NA NA NA NA ...
 $ mmt_duration_10       : num  248 NA NA NA NA ...
 $ mmt_duration_11       : num  1975 NA NA NA NA ...
 $ mmt_duration_12       : num  6748 NA NA NA NA ...
 $ mmt_count_07          : num  54 NA NA NA NA NA 29 19 0 NA ...
 $ mmt_count_08          : num  82 NA NA NA NA NA 69 26 1 NA ...
 $ mmt_count_09          : num  58 NA NA NA NA NA 138 36 13 NA ...
 $ mmt_count_10          : num  68 NA NA NA NA NA 178 29 0 NA ...
 $ mmt_count_11          : num  96 NA NA NA NA NA 225 102 0 NA ...
 $ mmt_count_12          : num  45 NA NA NA NA NA 119 76 0 NA ...
 $ mmt_duration_07       : num  7185 NA NA NA NA ...
 $ mmt_duration_08.1     : num  3991 NA NA NA NA ...
 $ mmt_duration_09.1     : num  5163 NA NA NA NA ...
 $ mmt_duration_10.1     : num  2477 NA NA NA NA ...
 $ mmt_duration_11.1     : num  6090 NA NA NA NA ...
 $ mmt_duration_12.1     : num  3205 NA NA NA NA ...
 $ mmo_total_count_07    : num  82 NA NA NA NA NA 88 10 0 NA ...
 $ mmo_total_count_08    : num  103 NA NA NA NA NA 67 70 101 NA ...
 $ mmo_total_count_09    : num  97 NA NA NA NA NA 46 89 108 NA ...
 $ mmo_total_count_10    : num  114 NA NA NA NA NA 63 158 0 NA ...
 $ mmo_total_count_11    : num  149 NA NA NA NA NA 80 28 0 NA ...
 $ mmo_total_count_12    : num  198 NA NA NA NA NA 121 14 0 NA ...
```

Şekil 5.2. Telekomünikasyon veri setine ilişkin tüm değişkenler, gösterim biçimleri ve türleri.

```
$ home_lon_08        : Factor w/ 1123 levels "","0","25,74",..: 114 1 1 1 1 1 383 975 370 1 ...
$ home_lat_09        : Factor w/ 519 levels "","0","36,04",..: 242 1 1 1 1 1 73 432 40 1 ...
$ home_lon_09        : Factor w/ 1055 levels "","0","26,06",..: 105 1 1 1 1 1 368 900 353 1 ...
$ home_lat_10        : Factor w/ 520 levels "","0","35,9",..: 242 1 1 1 1 1 74 432 445 1 ...
$ home_lon_10        : Factor w/ 1057 levels "","0","26,06",..: 102 1 1 1 1 1 369 901 238 1 ...
$ home_lat_11        : Factor w/ 514 levels "","0","36,04",..: 238 1 1 1 1 1 66 429 2 1 ...
$ home_lon_11        : Factor w/ 1049 levels "","0","26,06",..: 103 1 1 1 1 1 372 913 2 1 ...
$ home_lat_12        : Factor w/ 516 levels "","0","36,03",..: 240 1 1 1 1 1 70 430 2 1 ...
$ home_lon_12        : Factor w/ 1034 levels "","0","26,06",..: 102 1 1 1 1 1 363 899 2 1 ...
$ cd_total_07        : num  NA NA NA NA NA NA NA NA NA NA ...
$ cd_total_08        : num  NA NA NA NA NA NA NA NA NA NA ...
$ cd_total_09        : num  NA NA NA NA NA NA NA NA NA NA ...
$ cd_total_10        : num  NA NA NA NA NA NA NA NA NA NA ...
$ cd_total_11        : num  NA NA NA NA NA NA NA NA NA NA ...
$ cd_total_12        : num  NA NA NA NA NA NA NA NA NA NA ...
$ cd_distinct_cell_07: num  NA NA NA NA NA NA NA NA NA NA ...
$ cd_distinct_cell_08: num  NA NA NA NA NA NA NA NA NA NA ...
$ cd_distinct_cell_09: num  NA NA NA NA NA NA NA NA NA NA ...
$ cd_distinct_cell_10: num  NA NA NA NA NA NA NA NA NA NA ...
$ cd_distinct_cell_11: num  NA NA NA NA NA NA NA NA NA NA ...
$ cd_distinct_cell_12: num  NA NA NA NA NA NA NA NA NA NA ...
$ dealer_dist_07     : Factor w/ 1366 levels "","0","0,01",..: 410 1 1 1 1 1 197 554 2 1 ...
$ dealer_dist_08     : Factor w/ 1425 levels "","0","0,01",..: 403 1 1 1 1 1 197 702 80 1 ...
$ dealer_dist_09     : Factor w/ 1304 levels "","0","0,01",..: 377 1 1 1 1 1 197 517 80 1 ...
$ dealer_dist_10     : Factor w/ 1285 levels "","0","0,01",..: 375 1 1 1 1 1 197 516 59 1 ...
$ dealer_dist_11     : Factor w/ 1216 levels "","0","0,01",..: 358 1 1 1 1 1 197 602 2 1 ...
$ dealer_dist_12     : Factor w/ 1208 levels "","0","0,01",..: 358 1 1 1 1 1 197 602 2 1 ...
$ callcenter_count_07: num  0 NA NA NA NA NA 2 NA 0 0 ...
$ callcenter_count_08: num  1 NA NA NA NA NA 0 NA 1 0 ...
$ callcenter_count_09: num  0 NA NA NA NA NA 1 NA 0 0 ...
$ callcenter_count_10: num  0 NA NA NA NA NA 1 NA 0 0 ...
$ callcenter_count_11: num  0 NA NA NA NA NA 3 NA 0 0 ...
$ callcenter_count_12: num  0 NA NA NA NA NA 0 NA 0 1 ...
$ payment_07         : Factor w/ 508 levels "","0","1","1,00E-06",..: 176 1 1 1 1 1 261 2 2 1 ...
$ payment_08         : Factor w/ 506 levels "","0","0,00E+00",..: 170 1 1 1 1 1 198 170 393 1 ...
$ payment_09         : Factor w/ 527 levels "","0","0,00E+00",..: 166 1 1 1 1 1 201 166 318 1 ...
$ payment_10         : Factor w/ 539 levels "","0","1","1,5",..: 258 1 1 1 1 1 225 2 2 1 ...
$ payment_11         : Factor w/ 547 levels "","0","0,00E+00",..: 245 1 1 1 1 1 211 356 2 1 ...
$ payment_12         : Factor w/ 499 levels "","0","1","1,00E-06",..: 276 1 1 1 1 1 246 2 2 1 ...
$ unpaid_07          : num  NA NA NA NA NA NA NA NA NA NA ...
$ unpaid_08          : num  NA NA NA NA NA NA NA NA NA NA ...
$ unpaid_09          : num  NA NA NA NA NA NA NA NA NA NA ...
$ unpaid_10          : num  NA NA NA NA NA NA NA NA NA NA ...
$ unpaid_11          : num  NA NA NA NA NA NA NA NA NA NA ...
$ unpaid_12          : num  NA NA NA NA NA NA NA NA NA NA ...
$ payment_type_07    : num  NA NA NA NA NA NA NA NA NA NA ...
$ payment_type_08    : num  NA NA NA NA NA NA NA NA NA NA ...
$ payment_type_09    : num  NA NA NA NA NA NA NA NA NA NA ...
$ payment_type_10    : num  NA NA NA NA NA NA NA NA NA NA ...
$ payment_type_11    : num  NA NA NA NA NA NA NA NA NA NA ...
$ payment_type_12    : num  NA NA NA NA NA NA NA NA NA NA ...
$ churn_2013_07      : int  0 0 0 0 0 0 0 0 0 0 ...
$ churn_2013_08      : int  0 0 0 0 0 0 0 0 0 0 ...
$ churn_2013_09      : int  0 0 0 0 0 0 0 0 0 0 ...
$ churn_2013_10      : int  0 0 0 0 0 0 0 0 0 0 ...
$ churn_2013_11      : int  0 0 0 0 0 0 0 0 0 0 ...
$ churn_2013_12      : int  0 0 0 0 0 0 0 0 0 1 ...
```

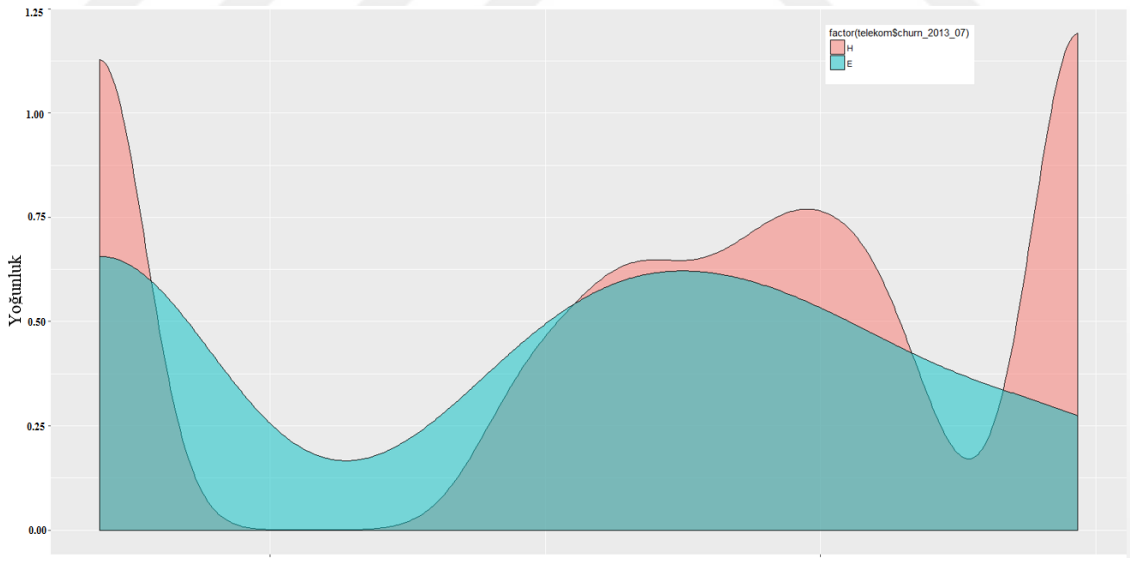
Şekil 5.2 (devam). Telekomünikasyon veri setine ilişkin tüm değişkenler, gösterim biçimleri ve türleri.

Veri setinde yer alan sayısal değişkenlerdeki değişimleri incelemek için çeşitli grafikler kullanılır. Histogram grafikleri de bunlardan biridir. Yaş ve müşterilik süresi niteliğine ait histogramlar Şekil 5.3'te görülmektedir.



Şekil 5.3. Yaş (age) ve müşterililik (age_of_line) histogramı.

Veri setindeki değişkenlerin anlaşılması sadece sayısal türdeki değişkenlerin görselleştirilmesi ile olmaz. Sayısal olmayan, kategorik, sıralı değişkenlerin yoğunluğu/dağılımı da incelenmelidir. Şekil 5.4'te müşterililik süresi ile müşterinin ayrılma bilgisini tutan ayrılma durum bilgisi (Churn) niteliği arasındaki yoğunluk grafiği görülmektedir.



Şekil 5.4. Müşterilik süresi ve müşteri ayrılma durum bilgisi (churn) arası yoğunluk grafiği.

Şekil 5.4 incelendiğinde müşterilik süre yoğunluğu arttıkça müşterilerin ayrılma durumu (H) hayır olarak görülüyor. Ek olarak veri setinden grafikler ile nasıl çıkarım yapılabileceğinin anlatılacağı Bölüm 5'te veri madenciliği yoluyla veri görselleştirme

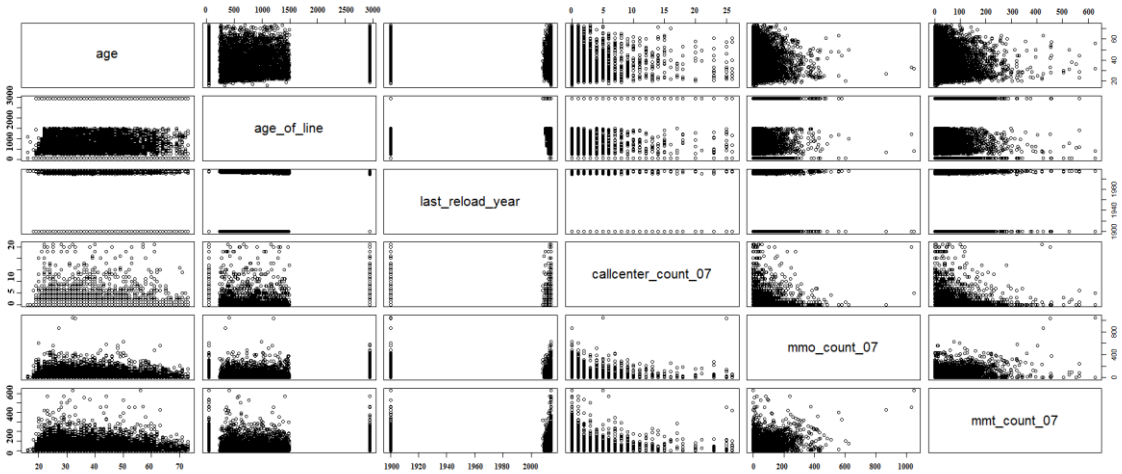
başlığı altında daha detaylı bilgi yer almaktadır.

Hedef değişken ve diğer değişkenler arasındaki ilişkinin görseller aracılığı ile anlaşılması mümkündür. Ancak analiz sürecinin doğru yapılması ve manipülatif sonuçlar elde edilmemesi için gerçek anlamda bir korelasyon olup olmadığının hesaplanması gerekir. Şekil 5.5'de telekomünikasyon veri setine yönelik korelasyon görülmektedir.

```
> cor(telekomcor[,1:9])
                        age  age_of_line  tariff_type device_type last_reload_year mmo_count_07 mmo_duration_07 callcenter_count_07 churn_2013_07
age             1.000000000  0.089173823 -0.007370368 -0.12008242     -0.006120624  -0.04668652     -0.04668652        -0.019016040  -0.004165275
age_of_line     0.089173823  1.000000000  0.038045488  0.02509882      0.135966836   0.10483374      0.10483374        -0.002778591  -0.049057639
tariff_type    -0.007370368  0.038045488  1.000000000  0.17847514     -0.681909414   0.19622496      0.19622496         0.065506032  -0.014967684
device_type    -0.120082425  0.025098816  0.178475138  1.00000000     -0.019273874   0.07033928      0.07033928         0.038224017  -0.052513027
last_reload_year -0.006120624 0.135966836 -0.681909414 -0.01927387     1.000000000  -0.11438419     -0.11438419        -0.037390358  -0.015201417
mmo_count_07    -0.046686523  0.104833736  0.196224960  0.07033928     -0.114384188  1.00000000      1.00000000         0.052099531  -0.010854428
mmo_duration_07 -0.046686523  0.104833736  0.196224960  0.07033928     -0.114384188  1.00000000      1.00000000         0.052099531  -0.010854428
callcenter_count_07 -0.019016040 -0.002778591 0.065506032 0.03822402   -0.037390358  0.05209953      0.05209953         1.000000000   0.038122172
churn_2013_07   -0.004165275 -0.049057639 -0.014967684 -0.05251303     -0.015201417 -0.01085443     -0.01085443        0.038122172   1.000000000
```

Şekil 5.5. Telekomünikasyon veri setindeki sayısal değişkenler arası korelasyon değerleri.

Şekil 5.5'de Hedef değişken (churn) ile yaş (age), müşterililik süresi (age_of_line), tarife tipi (tariff_type), kullanılan cihaz tipi (device_type), son yükleme tarihi (last_reload_year), aylık kendi aboneleriyle konuşma sayısı (mmo_count), aylık kendi aboneleriyle konuşma süresi (mmo_duration) negatif yönde düşüş, şikayet sayısında (callcenter_count) ise pozitif yönde düşüş görülmektedir. Bu yorum beşeri ilimlerde kullanılan korelasyon tablosu kullanılarak yapılmıştır [83]. Şekil 5.6'da Hedef değişken ve diğer değişkenler (sayısal) arasındaki korelasyonu gösteren grafikler görülmektedir.



Şekil 5.6. Hedef değişken ve diğer değişkenler arasındaki korelasyonu gösteren korelasyon grafiği.

5.4. VERİYİ HAZIRLAMA

Verilerin analizlere hazırlanmasında Bölüm 3.3'te özet bilgi olarak yer alan veri madenciliği süreçleri ve Bölüm 4.1'deki adımlar takip edilerek oluşturulan modelleri karşılaştırma çalışması boyunca kullanılacak olan makine öğrenmesi süreci adımlarında anlatılan işlemlerin uygulanması bu bölümde yapılacaktır. Bu çalışmada takip edilen makine öğrenmesi süreçlerinden veri temizleme, veri dönüştürme yapılmış ve bu işlemlerden sonra modellemeye geçilmiştir. Veri önişleme olarak da bilinen veri temizleme adımında aykırı veriler (uç noktalar) ve eksik veriler üzerinde işlemler yapılmıştır. Şekil 5.1'deki veri seti özet bilgisi incelendiğinde eksik veriler ve aykırı veriler görülmektedir. Bu görsele ek olarak aykırı veriler kutu grafikleri (boxplot) ile de görüntülenebilmektedir.

İlk olarak veri temizleme adımında eksik verilerin tamamlanması ile aykırı verilerin düzenlenmesi anlatılmıştır. Bir sonraki adım olan veri dönüştürme adımında aylık konuşma süreleri ve yaş gibi değişkenlerde yapılan veri dönüşümleri anlatılmıştır. Ayrıca veri setinin değişken sayısını azaltmak ve faktör analizi yapmak için değişken seçme yöntemlerinden Temel Bileşen Analizi (PCA) kullanılmıştır.

5.4.1. Veri Temizleme

Veri temizleme, veriler üzerinde işlem yapmak için en kritik ve zaman kaybedilen adımdır. Bu bölümde veri temizleme, eksik verilerin tespiti ve tamamlanması, R paketleri ile aykırı verilerin tespiti ve çözümlenmesi olmak üzere iki alt başlıkta anlatılacaktır.

5.4.1.1. *Eksik Verilerin Tespiti ve Tamamlanması*

Eksik verileri çözümlerken birden çok yöntem kullanılır. Bölüm 4.1.3.1'de, eksik verileri tamamlama yöntemlerinden bahsedilmiştir. Telekomünikasyon veri seti üzerinde R paketlerinden mice ile, expectation-maximization (EM-maksimum beklenti) algoritmasıyla, nümerik değişkenlerin eksikliğinde ortalama, kategorik değişkenlerin eksikliğinde en çok tekrar eden değer atamalarının yapıldığı klasik yöntemlerle ve son olarak eksik verilerin atılması olmak üzere dört farklı yöntem ile eksik veriler tamamlanmış ve bunun sonucunda dört veri seti elde edilmiştir. Bu dört veri setinin Tablo 5.2'de görüldüğü gibi sınıflandırma algoritmalarındaki performansları karşılaştırılmıştır. Bu karşılaştırma sonucunda tüm sınıflandırma algoritmaları ile en iyi

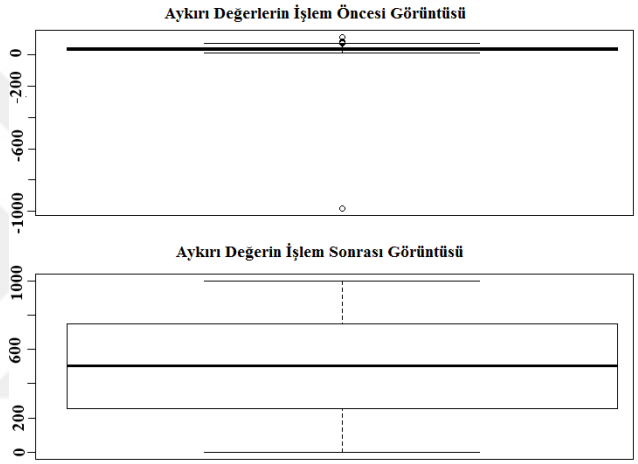
sonucu veren veri seti, R paketlerinden olan mice ile elde edilen veri seti olmuştur. Bundan sonraki tüm çalışma, R paketlerinden mice ile eksik verileri tamamlanmış olan bu veri seti üzerinde gerçekleştirilecektir. Eksik verileri tamamlamaya yönelik yapılan çalışmaların kodları EK-1'de yer almaktadır. Buradaki karşılaştırma, çalışılan telekomünikasyon veri setine özgü bir çalışmadır. Başka veri setleri üzerinde farklı sonuçlar verebilir.

Tablo 5.2. Eksik veri tamamlama yöntemlerinin sınıflandırma algoritmalarıyla karşılaştırılması.

Algoritmalar	Yöntemler	Sonuçlar			
		Doğruluk	Hata	DOR	F-ölçü
C4.5	Mice	0,983	0,016	96,21	0,99
	EM	0,978	0,0212	40,59	0,989
	NA Çıkarma	0,980	0,019	92.11	0,990
	Klasik	0,982	0,0171	84,92	0,9912
ID3	Mice	0,989	0,019	65,45	0,990
	EM	0,982	0,0175	199,695	0,992
	NA Çıkarma	0,980	0,019	65,45	0,990
	Klasik	0,985	0,015	547,936	0,992
Gini	Mice	0,977	0.022	15,65	0,98
	EM	0,975	0.024	15,96	0,987
	NA Çıkarma	0.970	0.029	11,21	0,98
	Klasik	0.974	0.025	25,92	0,98
k-en yakın	Mice	0,983	0,016	30,24	0,9914
	EM	0,9825	0,0175	0	0,9911
	NA Çıkarma	0,9809	0,01	0	0,9903
	Klasik	0,9828	0,0171	0	0,9913
Bayes	Mice	0,7952	0,204	27,49	0,883
	EM	0,777	0,222	13,65	0,873
	NA Çıkarma	0,7924	0,207	11,77	0,882
	Klasik	0,7877	0,212	5.41	0,879

5.4.1.2. *R Paketleri ile Aykırı Verilerin Tespiti ve Çözümlenmesi*

Aykırı veriler, elimizdeki veri seti hakkında tahmin ve tanımlamalarımızda hatalı sonuç vermemize neden olur. Aykırı değerleri tespit ederken çeşitli grafikler kullanılabilir. En temel boxplot grafiği ile tespit edilir. Bölüm 4.1.3.2'de aykırı veri oluşma durumları ve aykırı değerleri çözümleme yöntemlerinin neler olduğundan bahsedilmiştir. Bu çalışmada aykırı değerler kutulama (bining) yöntemi ile çözümlenmiştir. Şekil 3.1'de veri seti özetinde görülen yaş niteliğindeki aykırı değer, Şekil 5.7'de aykırı veriler üzerine kutulama yöntemi uygulanmasının öncesi ve sonrası şeklinde görüntüsü yer almaktadır. Bu çalışma kapsamında aykırı verilerin tespiti ve çözümlemesine yönelik yapılan çalışmanın kodları EK-1'de yer almaktadır.



Şekil 5.7. Yaş niteliğindeki aykırı değerlerin işlemler öncesi ve sonrası.

5.4.2. Veri Dönüştürme

Verilerin daha sağlıklı şekilde kullanılması için yararlanılacak olan algoritmanın çalışabileceği türe dönüştürülmesi işlemine veri dönüşümü denir. Kabaca sürekli değişkenleri kesikli hale getirme işlemine veri dönüştürme denir. Yaş, konuşma süresi gibi değişkenleri çeşitli aralıklar ile sınırlandırarak modelleme aşamasında düzgün sonuç alınmasını sağlar. Ayrıca yorumlama işlemini de kolaylaştırır. Telekomünikasyon veri setinde yaş (age), aylık konuşma süresi (mmo_duration) değişkenlerinde dönüşüm uygulanmıştır. Örnek olarak yaş niteliğinin özet bilgisi Şekil 5.8'de görülmektedir. Şekil 5.8'deki yaş niteliği sürekli değişkenlerden oluşmaktadır. Şekil 5.9'da yaş değişkeninin yeni aralıklar ile kesikli değişkene dönüşümü gösterilmiştir. Yaş niteliği Şekil 5.8'de görüldüğü gibi en düşük 16 ile en yüksek 73 değerleri arasında iken belirlenen yeni sınırlar ile 4 kategoriye indirgenerek kesikli hale getirilmiştir. Böylece

Bölüm 4.4'teki grafiklerde de daha rahat görüleceği gibi daha kullanılabilir hale getirilmiştir.

```
      age
 Min.   :16.00
 1st Qu.:28.00
 Median :36.00
 Mean   :38.13
 3rd Qu.:46.00
 Max.   :73.00
```

Şekil 5.8. Yaş niteliğinin veri dönüşüm öncesi.

```
     age
 16-24: 993
 25-44:4662
 45-64:2159
 65+  : 186
```

Şekil 5.9. Yaş niteliğinin veri dönüşüm sonrası.

Telekomünikasyon veri seti üzerindeki eksik verilerin, aykırı verilerin ve sürekli değişkenlerin kesikli değişkene dönüşümü gibi veri ön işleme adımlarından sonra veri setinin son hali Şekil 5.10'da görülmektedir.

```
> summary(telekom)
     number       gender_flag     age          age_of_line      tariff_type           device_type
 Min.   :    2   U:  13      16-24: 993   Min.   :  49   Kontörlü:5461   Mobil Tel     :3701
 1st Qu.: 2513   K:5758      25-44:4662   1st Qu.: 266   Faturalı:2539   Akıllı Telefon:3254
 Median : 5014   E:2229      45-64:2159   Median : 705                   Bilinmeyen    : 786
 Mean   : 5013               65+  : 186   Mean   :1117                   Usb Modem     : 229
 3rd Qu.: 7505                            3rd Qu.:1476                   Tablet PC     :  17
 Max.   :10000                            Max.   :2935                   Modül         :  11
                                                                         (Other)       :   2
   last_reload_year last_reload_amount     expiry_year       hotline_year      crm_segment
 2014      :3980    0      :2330       Bilinmeyen:2759   Bilinmeyen:2759   Bilinmeyen:2430
 Bilinmeyen:2326    20     :1596       2010       :   1   2010       :   1   Bronze    :3618
 2013      :1189    10     :1228       2013       : 270   2013       : 395   Silver    :1402
 2012      : 275    12     : 774       2014       :1708   2014       :4845   Gold      : 477
 2011      : 113    30     : 748       2015       :3262                      Platinum  :  73
 2010      :  77    25     : 617
 (Other)   :  40    (Other): 707
     value_segment    lifestyle_segment  mmo_count_07      mmo_count_08      mmo_count_10
 Bilinmeyen  :1915   Bilinmeyen : 642   Min.   :   0.00   Min.   :   0.00   Min.   :   0.00
 YeniMusteri1: 367   Segmentsiz : 608   1st Qu.:   0.00   1st Qu.:   1.00   1st Qu.:   1.00
 YeniMusteri2: 390   Kitle      :4249   Median :  15.00   Median :  17.00   Median :  17.00
 Standart    :5075   Genç       :1807   Mean   :  40.63   Mean   :  43.86   Mean   :  43.55
 Premium     : 206   Profesyonel: 694   3rd Qu.:  54.00   3rd Qu.:  59.00   3rd Qu.:  59.00
 PremiumPlus :  47                      Max.   :1049.00   Max.   :1150.00   Max.   :1100.00

 mmo_count_11      mmo_count_12      mmo_duration_07   mmo_duration_08   mmo_duration_09
 Min.   :   0.00   Min.   :   0.00   0-2 saat  :8000   0-2 saat  :7245   0-2 saat  :7173
 1st Qu.:   1.00   1st Qu.:   1.00   2-4 saat  :   0   2-4 saat  : 591   2-4 saat  : 654
 Median :  19.00   Median :  18.00   4-6 saat  :   0   4-6 saat  : 107   4-6 saat  : 118
 Mean   :  44.92   Mean   :  43.63   6-8 saat  :   0   6-8 saat  :  46   6-8 saat  :  34
 3rd Qu.:  62.00   3rd Qu.:  60.00   8-10 saat :   0   8-10 saat :   6   8-10 saat :  12
 Max.   :1027.00   Max.   :1033.00   10-12 saat:   0   10-12 saat:   2   10-12 saat:   3
                                     (Other)   :   0   (Other)   :   3   (Other)   :   6
 mmo_duration_10   mmo_duration_11   mmo_duration_12  mmt_count_07      mmt_count_08
 0-2 saat :7188    0-2 saat  :7139   0-2 saat  :7200  Min.   :  0.00   Min.   :   0.00
 2-4 saat : 616    2-4 saat  : 681   2-4 saat  : 647  1st Qu.:  1.00   1st Qu.:   2.00
 4-6 saat : 155    4-6 saat  : 139   4-6 saat  : 102  Median : 19.00   Median :  24.00
 6-8 saat :  28    6-8 saat  :  30   6-8 saat  :  41  Mean   : 44.44   Mean   :  63.87
 8-10 saat:  10    8-10 saat :   7   8-10 saat :   8  3rd Qu.: 62.00   3rd Qu.:  82.00
 (Other)  :   2    10-12 saat:   2   14-16 saat:   1  Max.   :960.00   Max.   :1155.00
 NA's     :   1    (Other)   :   2   (Other)   :   1
```

Şekil 5.10. Veri önişleme sonrası veri setinin özet bilgisi.

```
cd_distinct_cell_11 cd_distinct_cell_12 dealer_dist_07 dealer_dist_08 dealer_dist_09 dealer_dist_10
Min.   : 0.0000     Min.   : 0.0000     0      :3318   0      :3292   0      :3282   0      :3250
1st Qu.: 0.0000     1st Qu.: 0.0000     0,1    :  34   0,1    :  33   0,1    :  31   0,1    :  31
Median : 0.0000     Median : 0.0000     0,49   :  29   0,49   :  28   0,49   :  31   0,59   :  31
Mean   : 0.9606     Mean   : 0.9277     0,51   :  28   0,54   :  27   0,69   :  30   0,83   :  30
3rd Qu.: 1.0000     3rd Qu.: 1.0000     0,54   :  28   0,51   :  25   0,59   :  29   0,4    :  29
Max.   :15.0000     Max.   :14.0000     0,62   :  27   0,86   :  25   0,62   :  28   0,49   :  29
                                        (Other):4536   (Other):4570   (Other):4569   (Other):4600
dealer_dist_11 dealer_dist_12 callcenter_count_07 callcenter_count_08 callcenter_count_09
0      :3185   0      :3194   Min.   : 0.0000     Min.   : 0.0000     Min.   : 0.0000
0,49   :  36   0,59   :  37   1st Qu.: 0.0000     1st Qu.: 0.0000     1st Qu.: 0.0000
0,68   :  34   0,83   :  35   Median : 0.0000     Median : 0.0000     Median : 0.0000
0,75   :  33   0,4    :  34   Mean   : 0.9441     Mean   : 0.9048     Mean   : 0.7306
0,1    :  32   0,1    :  33   3rd Qu.: 1.0000     3rd Qu.: 1.0000     3rd Qu.: 1.0000
0,4    :  32   0,49   :  32   Max.   :26.0000     Max.   :27.0000     Max.   :28.0000
(Other):4648   (Other):4635
callcenter_count_10 callcenter_count_11 callcenter_count_12   payment_07      payment_08
Min.   : 0.0000     Min.   : 0.0000     Min.   : 0.0000     Min.   :1.000   Min.   :  1.00
1st Qu.: 0.0000     1st Qu.: 0.0000     1st Qu.: 0.0000     1st Qu.:1.000   1st Qu.:  1.00
Median : 0.0000     Median : 0.0000     Median : 0.0000     Median :1.000   Median :  1.00
Mean   : 0.8822     Mean   : 0.8989     Mean   : 0.8756     Mean   :1.088   Mean   : 99.61
3rd Qu.: 1.0000     3rd Qu.: 1.0000     3rd Qu.: 0.0000     3rd Qu.:1.000   3rd Qu.:197.00
Max.   :27.0000     Max.   :28.0000     Max.   :28.0000     Max.   :8.000   Max.   :505.00
                                                            NA's   :3241
  payment_09      payment_10      payment_11      payment_12      unpaid_07       unpaid_08
Min.   :  1.0   Min.   :  1.0   Min.   :  1.0   Min.   :  1.00   Min.   :0.0000   Min.   :0.0000
1st Qu.:  1.0   1st Qu.:  1.0   1st Qu.:  1.0   1st Qu.:  1.00   1st Qu.:0.0000   1st Qu.:0.0000
Median :  1.0   Median :  1.0   Median :  1.0   Median :  1.00   Median :0.0000   Median :0.0000
Mean   :102.9   Mean   :113.8   Mean   :106.7   Mean   : 91.83   Mean   :0.2021   Mean   :0.2021
3rd Qu.:200.0   3rd Qu.:224.0   3rd Qu.:210.0   3rd Qu.:179.00   3rd Qu.:0.0000   3rd Qu.:0.0000
Max.   :526.0   Max.   :538.0   Max.   :546.0   Max.   :498.00   Max.   :2.0000   Max.   :2.0000

  unpaid_09       unpaid_10       unpaid_11       unpaid_12            payment_type_07
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Mobil Tel      :3701
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   Akıllı Telefon:3254
Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000   Bilinmeyen     : 786
Mean   :0.2021   Mean   :0.2021   Mean   :0.2021   Mean   :0.2021   Usb Modem      : 229
3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   Tablet PC      :  17
Max.   :2.0000   Max.   :2.0000   Max.   :2.0000   Max.   :2.0000   Modül          :  11
                                                                    (Other)        :   2
payment_type_08 payment_type_09 payment_type_10 payment_type_11 payment_type_12 churn_2013_07
19     :1555    19     :1809    19     :1762    19     :2768    19     :2768    H:7857
0      :1274    0      :1067    2      :1046    0      :1740    0      :1740    E: 143
20     : 881    2      : 945    0      : 951    20     :1363    20     :1363
2      : 877    20     : 847    20     : 795    25     :1336    25     :1336
17     : 702    17     : 685    17     : 691    2      : 215    2      : 215
8      : 610    8      : 659    8      : 657    8      : 137    8      : 137
(Other):2101    (Other):1988    (Other):2098    (Other): 441    (Other): 441
churn_2013_08 churn_2013_09 churn_2013_10 churn_2013_11 churn_2013_12
H:7845        H:7851        H:7864        H:7856        H:7842
E: 155        E: 149        E: 136        E: 144        E: 158
```

Şekil 5.10 (devam). Veri önişleme sonrası veri setinin özet bilgisi.

5.4.3. Değişken Seçme

Boyut küçültme işlemi elimizdeki veri setinin değişken sayısını küçültme amacıyla yapılmaktadır. Bunun birkaç sebebi olabilir. Birincisi yeterli hesaplama kapasitesi olan donanıma sahip olunmayabilir. İkincisi, veri setindeki değişkenlerden bilgi verici olmayanlarını elemek istenebilir. Üçüncüsü sadece hangi özelliklerin önem taşıdığı görmek istenebilir. Bu çalışmada değişken azaltma yöntemlerinden PCA kullanılmıştır.

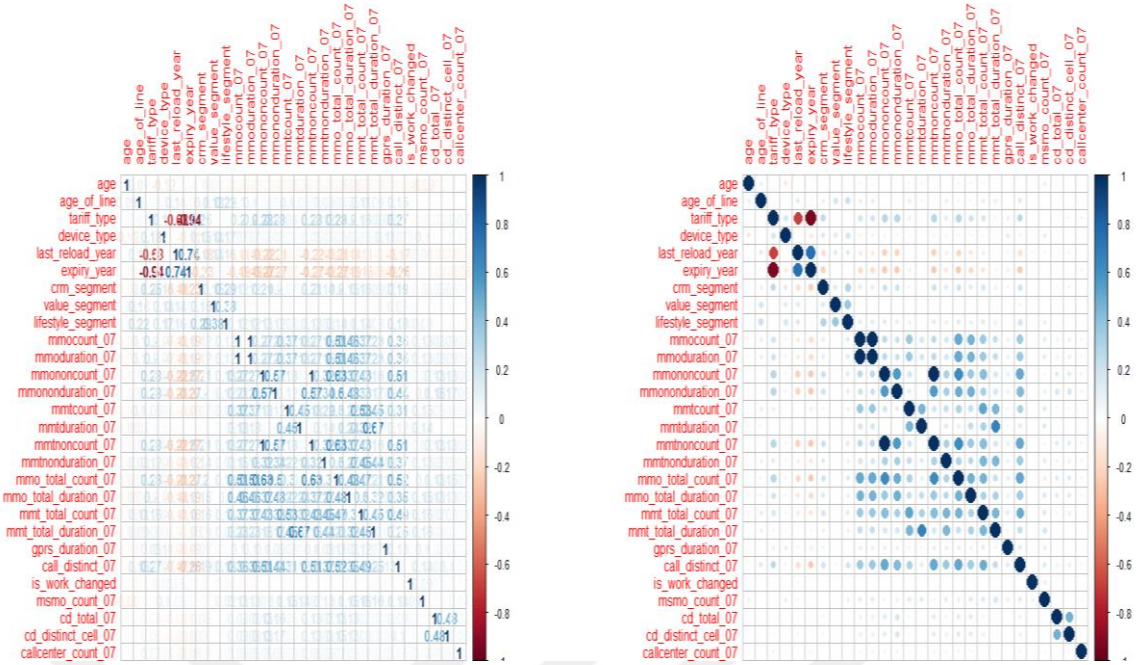
5.4.3.1. *Temel Bileşen Analizi*

Temel Bileşen Analizi (PCA) çok değişkenli nümerik bir veri seti içindeki bilgiyi daha az değişkenler ile fakat minimum bilgi kaybı ile açıklamanın matematiksel tekniğidir. Bir veri setindeki bilgi buradaki toplam değişken ile açıklanmaktadır. PCA büyük boyutlu verisetlerindeki boyutsallığı azaltır. Simetrik matrislerin spektral analiz

yöntemlerine dayanan PCA kullanıcıların ileri düzeyde matematik, istatistik ve liner cebir bilgisini gerektirir. Fakat R veri madenciliği paketleri ve görselleştirme yetenekleri kullanıcılardan böyle bir bilgi talep etmeden analize ve sonuçlara odaklanabilmelerine imkan sağlamaktadır.

PCA'nın boyut indirgemesi veri setindeki mevcut aralarında korelasyon olan değişkenleri bazı lineer dönüşümlerle aynı sayıda ama aralarında korelasyon olmayan değişkenlere dönüştürmeye dayanmaktadır. Bu yeni değişkenler mevcut değişkenlerin bir lineer kombinasyonu olup asal bileşenler olarak anılmaktadır. Tekniğin matematik yapısı gereği veri setindeki toplam enformasyon az sayıdaki asal bileşende yoğunlaşmış hale gelmektedir.

PCA'nın temel yapısı; PCA analizinin temeli veri setindeki değişkenler arası kovaryans veya korelasyon matrisinin spektral özelliklerine dayanmaktadır. Yapısı gereği pozitif ve simetrik olan bu matrisin özdeğerleri (eigenvalues) pozitiftir ve verilerin varyansları ile özdeştir. PCA veri setlerinin kovaryans veya korelasyon matrislerinin özdeğerlerini ve özvektörlerini bulma problemidir. Yazılımlar gelişmeden önce bu teknik oldukça zor olmaktaydı. R paketleri sayesinde PCA hem kolay hem de son derece görsel hale gelmiştir. PCA, R paketleri sayesinde tek komutla gerçekleşmesinin yanı sıra değişkenleri seçme işlemi grafiklerle de görsel olarak gösterilebilmektedir [99]. Örnek olarak 7. aya ait telekomünikasyon bilgilerinin PCA analizi yapılarak değişken azaltma anlatılacaktır. PCA ilk adım olarak kovaryans/korelasyon matris hesabı gerektirir. Şekil 5.11'de 7. aya ait telekomünikasyon veri setinin korelasyon hesabını şekilsel olarak görülmektedir.



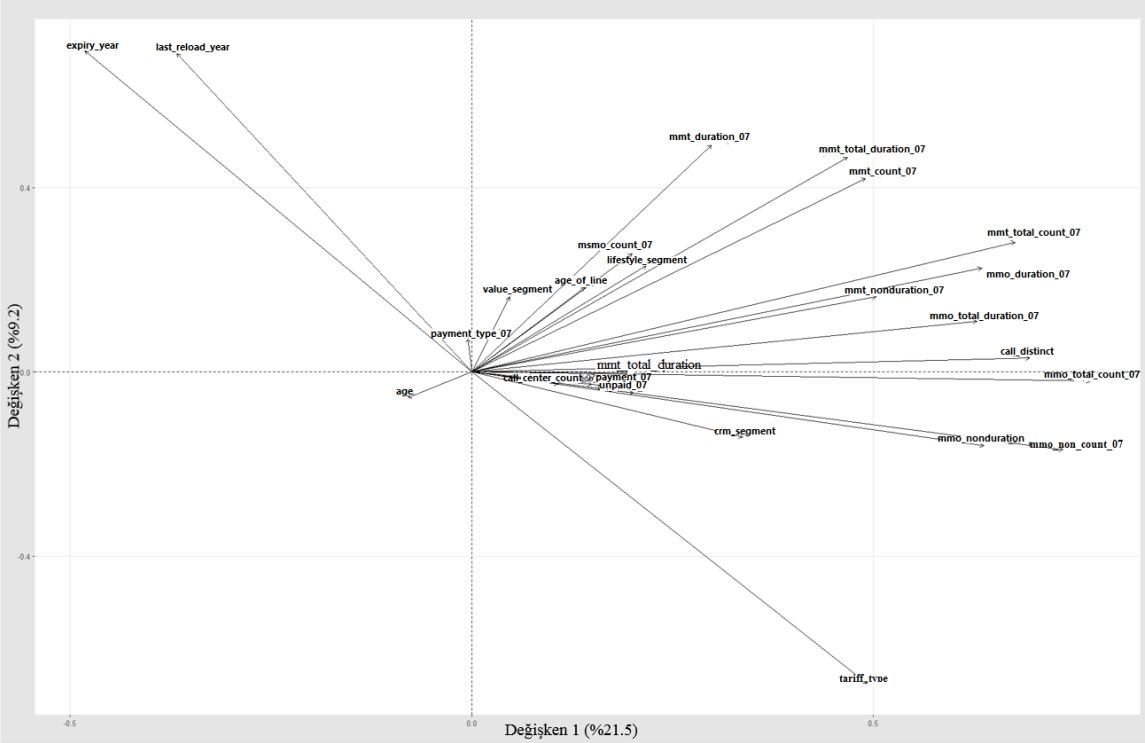
Şekil 5.11. Korelasyonun şekilsel gösterimi.

Şekil 5.12'de 7. aya ait telekomünikasyon veri setinin age (yaş), churn (ayrılma durumu), age_of_line (müşterililik süresi), tariff_type (tarife tipi) ve call_center_count (müşteri şikayet bilgisi), hedef değişken olan churn (müşteri ayrılma durumu) ve gender_flag (cinsiyet) içeren 5 değişken arasında korelasyon özet tablosu ve maksimum korelasyon oranı görülmektedir.

```
> fsv<- facto_summarize(res.pca, "var", axes = 1:5)
>
>  print(fsv)
                                    name        Dim.1        Dim.2        Dim.3       Dim.4       Dim.5       coord        cos2     contrib
gender_flag                   gender_flag -0.026076769  0.065978029 -0.1312350828  0.09389572  0.14523061 0.052164082 0.052170603   5.2164082
age                                 age -0.073771237 -0.059203975  0.0641298294 -0.01014488 -0.12038093 0.027654427 0.027657884   2.7654427
age_of_line                 age_of_line  0.132390113  0.149634610 -0.0164149522  0.41678137 -0.11348085 0.226771720 0.226800070  22.6771720
tariff_type                 tariff_type  0.591096927 -0.706282290 -0.2039526060  0.14033937  0.01574440 0.909769940 0.909883675  90.9769940
device_type                 device_type  0.154418090 -0.027181032 -0.0524871107  0.42220449  0.07971024 0.211949010 0.211975507  21.1949010
last_reload_year       last_reload_year -0.452710787  0.686004017  0.1769431197  0.12786019 -0.04472970 0.725206409 0.725297071  72.5206409
expiry_year                 expiry_year -0.587788798  0.741219228  0.2293222139 -0.09052673 -0.02165910 0.956154497 0.956274032  95.6154497
hotline_year               hotline_year -0.587933384  0.740817134  0.2290576733 -0.09202542 -0.02145424 0.955872070 0.955991569  95.5872070
crm_segment                 crm_segment  0.347525598 -0.087488675  0.0978623022  0.42034976  0.07141089 0.319798772 0.319838752  31.9798772
value_segment             value_segment  0.034403039  0.150608018  0.1214248978  0.68800614 -0.08892157 0.519869848 0.519934840  51.9869848
lifestyle_segment     lifestyle_segment  0.191200552  0.251565846  0.1694461668  0.64723654 -0.05447578 0.550437785 0.550506598  55.0437785
mmo_count_07               mmo_count_07  0.610814671  0.263933924 -0.2117885116 -0.04371265 -0.62598861 0.881382582 0.881492768  88.1382582
mmo_duration_07         mmo_duration_07  0.610814671  0.263933924 -0.2117885116 -0.04371265 -0.62598861 0.881382582 0.881492768  88.1382582
mmo_noncount_07         mmo_noncount_07  0.721588242  0.043371858  0.4997464901 -0.11094525  0.13722756 0.803457514 0.803557959  80.3457514
mmo_nonduration_07   mmo_nonduration_07  0.630263505  0.008039082  0.3761035726 -0.07600344  0.13679136 0.563239009 0.563309423  56.3239009
mmt_count_07               mmt_count_07  0.457859206  0.382716648 -0.4139084032 -0.04766517  0.03793289 0.531138125 0.531204526  53.1138125
mmt_duration_07         mmt_duration_07  0.267830569  0.392875172 -0.5773846271 -0.04873286  0.25997689 0.629419998 0.629498686  62.9419998
mmt_noncount_07         mmt_noncount_07  0.721588242  0.043371858  0.4997464901 -0.11094525  0.13722756 0.803457514 0.803557959  80.3457514
mmt_nonduration_07   mmt_nonduration_07  0.480462403  0.224931553 -0.0005490238 -0.03941140  0.35369252 0.408090284 0.408141302  40.8090284
mmo_total_count_07   mmo_total_count_07  0.751177400  0.136160885  0.2266121102 -0.11591788 -0.15119554 0.670457368 0.670541185  67.0457368
mmo_total_duration_07 mmo_total_duration_07  0.607885852  0.197286472  0.0045356063 -0.11517849 -0.12627118 0.437678228 0.437732945  43.7678228
mmt_total_count_07   mmt_total_count_07  0.643709402  0.335232769 -0.1218273583 -0.05135167  0.15535062 0.568355519 0.568426572  56.8355519
mmt_total_duration_07 mmt_total_duration_07  0.433847378  0.411121742 -0.4764594848 -0.05611954  0.34543202 0.706730960 0.706819313  70.6730960
gprs_duration_07       gprs_duration_07  0.194150712  0.020525855 -0.0388071286  0.02735900  0.11008985 0.052490092 0.052496654   5.2490092
call_distinct_07       call_distinct_07  0.678979975  0.142192360  0.1459107216  0.04550466  0.03209470 0.505623157 0.505686368  50.5623157
is_home_changed         is_home_changed  0.030694584  0.037358362 -0.0046657742  0.14121809  0.30360230 0.114476481 0.114490792  11.4476481
is_work_changed         is_work_changed -0.007698572  0.069544551  0.0061232330  0.15018343  0.32017578 0.130000799 0.130017051  13.0000799
msmo_count_07             msmo_count_07  0.175828839  0.249133155 -0.1238844821  0.03279747  0.17676193 0.140650930 0.140668513  14.0650930
cd_total_07                 cd_total_07  0.181800595  0.044167133  0.2558100551 -0.15395015  0.03730220 0.125533081 0.125548774  12.5533081
cd_distinct_cell_07 cd_distinct_cell_07  0.192981095  0.035712716  0.2546701202 -0.17262592  0.02361583 0.133731385 0.133748104  13.3731385
callcenter_count_07 callcenter_count_07  0.109016511 -0.009335423  0.0081145382  0.01500852  0.05663689 0.015470589 0.015472523   1.5470589
unpaid_07                     unpaid_07 -0.013185778 -0.025440185  0.0193251312 -0.00721641  0.05309990 0.004066205 0.004066713   0.4066205
payment_type_07         payment_type_07  0.057449612  0.065432695  0.0351333901 -0.08157749 -0.02957824 0.016346010 0.016348053   1.6346010
churn_2013_07             churn_2013_07 -0.013428898 -0.015824849 -0.0213730671 -0.31732693  0.06202217 0.105430702 0.105443882  10.5430702
> fsvmax<-max(fsv$cos2)
> print(fsvmax)
[1] 0.956274
```

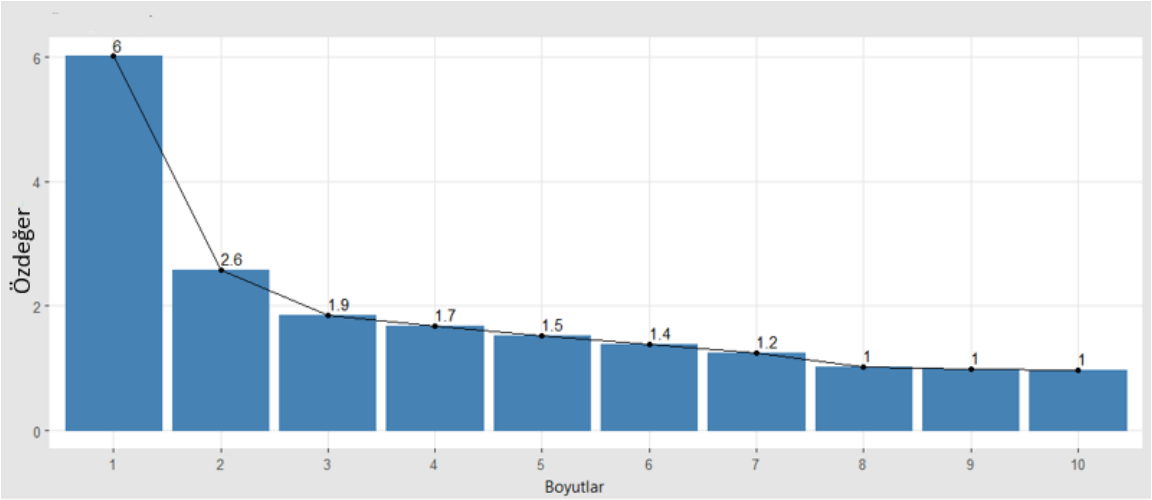
Şekil 5.12. 5 Değişken arasındaki korelasyon özet tablosu.

PCA() fonksiyonundan elde edilen sonuç **k**orelasyon çemberi olarak da Şekil 3.13'te görülmektedir**.** PCA bileşenleri orijin bileşenler ile aynı sayıdadır fakat birbirleri ile 0 (ortogonal) korelasyonludurlar. İlk beş bileşen veri setindeki varyasyonun büyük bir yüzdesini yansıtmaktadır. Şekil 5.13'te ilk iki asal bileşene karşı değişkenlerin pozisyonu görülmektedir. Şekil 5.13'teki okların kısalığı ve çembere uzaklığı değişkenler ile bileşenler arasındaki korelasyon ilişkisinin zayıf olduğunu göstermektedir.



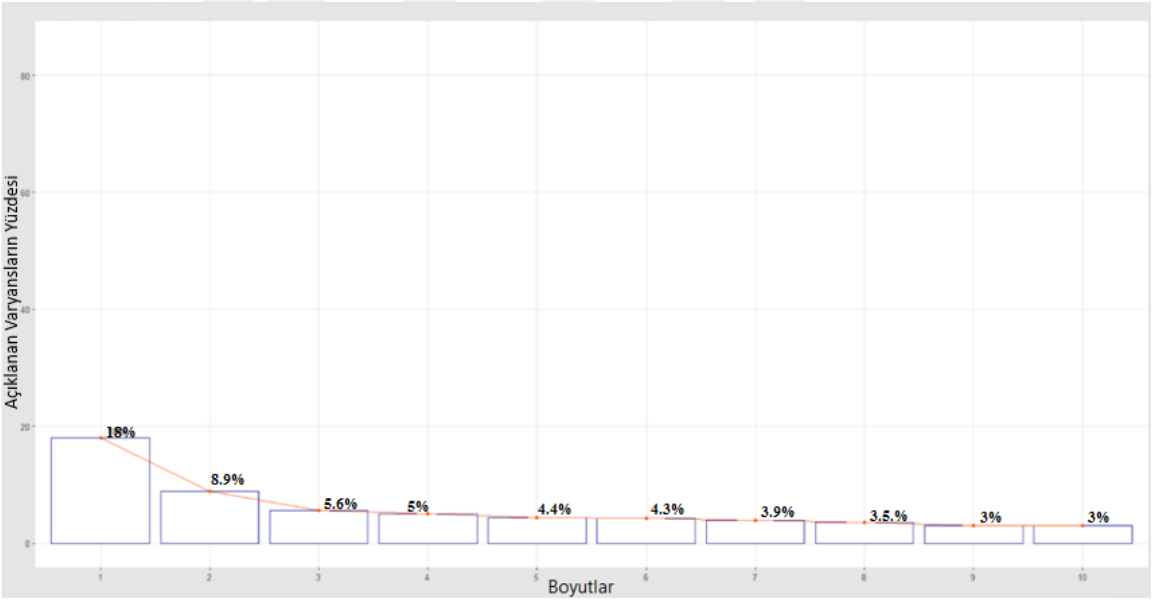
Şekil 5.13. Korelasyon çemberi.

PCA analizinde değişkenler ile bileşenleri arasındaki ilişkiyi veri setinin Kovaryans/Korelasyon matrisleri belirlemektedir. Şekil 5.14'te 7. aya ait telekomünikasyon veri setinin özdeğerlerinin açıklandığı barplot grafiği görülmektedir.



Şekil 5.14. Özdeğerler.

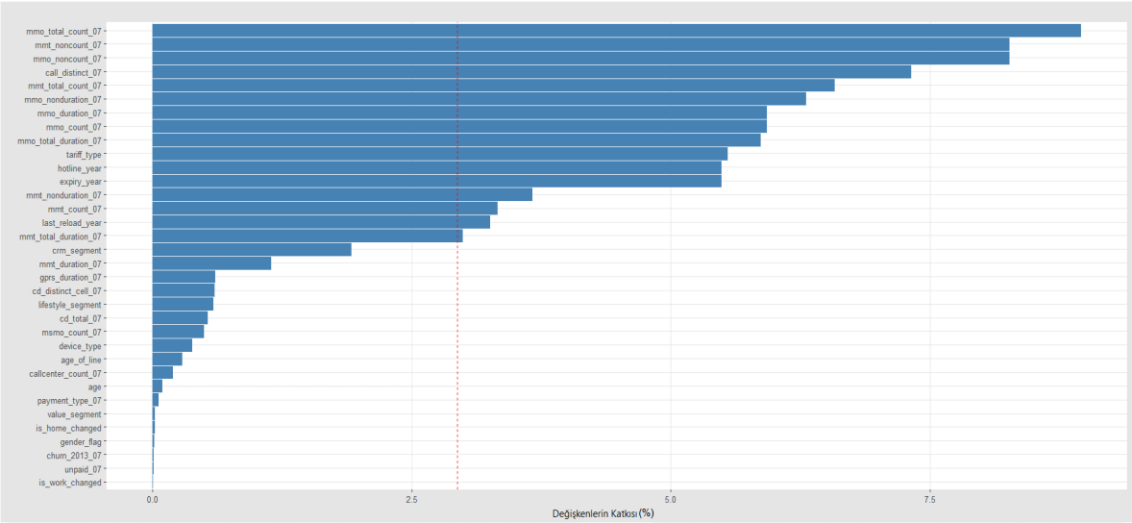
Şekil 5.15'taki grafikte 7. aya ait telekomünikasyon veri seti değişkenlerinin varyans yüzde grafiği görülmektedir. Seçilecek bileşen sayısı özdeğerler grafiğinden bulunmaktadır.



Şekil 5.15. Bileşenlerin varyans yüzde grafiği.

Kaiser kriterine göre bu değişkenlerin özdeğerleri birden büyük olmalıdır [100]. Şekil 3.15'te bu kritere uyan on bileşen vardır. Bu bileşenler sistemdeki toplam varyansın (%18 + %8.9 + %5.6 + %5 + %4.4 + %4.3 + %3.9 + %3.5 + %3 + %3 = %59.6) yüzde 59.6'lık kısmını açıklayabilmektedir. Veri setimizdeki değişkenler kategorik olan müşteri kişisel bilgileri ve aylık telekomünikasyon verileri şeklinde PCA analizine tabi tutulmuştur. Örnek olarak veri setinde hangi değişkenin hangi bileşenle nasıl ilişkisi

olduğunu Şekil 5.16'daki gibi birinci değişkeni etkileyen diğer değişkenler görülmektedir.



Şekil 5.16. Birinci bileşen olan cinsiyeti (gender_flag) etkileyen değişkenler.

Şekil 5.16'da örnek olarak 7. Aya ait telekomünikasyon veri setinin birinci bileşeni olarak alınan yaşa (age) göre gprs niteliği, evde ya da işte konuşma süresini (is_home_changed, is_work_changed) tutan değişkenler gibi değişkenleri arasında ilişki diğer değişkenlere göre daha az olduğu görülmektedir. Bundan dolayı bu değişkenlerin veri setinden çıkarılabileceği anlaşılmaktadır. Ayrıca bu değişkenlerle ve bu değişkenlersiz oluşturulacak modeller arasında nasıl bir fark olacağına bakarak çıkarılan değişkenin analize katkısı kontrol edilmiştir.

PCA tekniği sistemdeki varyasyonu neyin açıkladığını bulma yaklaşımıdır. Burada amaç telekomünikasyon veri setinde olmadığı durumda oluşturulacak modelleri etkilemeyecek bileşenleri görmektir. Bu amaç doğrultusunda ve yukarıdaki anlatımda geçtiği gibi telekomünikasyon veri setine PCA ile değişken azaltma uygulanmış ve expiry_date, hotline_date gibi müşteri kişisel bilgilerine ait değişkenler ile non_percent_07_12 (aylık konuştuğu kişilerden diğer operatörden olanların kendi abonesi olanlara oranı), gprs_duration_07_12 (aylık gprs kullanım süresi), call_distinct_07_12 (aylık konuşulan farklı kişi sayısı), is_home_changed (aylık olarak ev lokasyonu değişme durumu), is_work_changed (aylık olarak iş lokasyonu değişme durumu), home_lat_07_12 (aylık ev lokasyonu (enlem)), home_lon_07_12 (aylık ev lokasyonu (boylam)), cd_total_07_12 (aylık call-drop sayıları), cd_distinct_cell_07_12 (aylık call-drop yaşadığı farklı baz istasyonu sayısı), dealer_dist_07_12 (aylık ev

lokasyonunun en yakın bayisine olan uzaklığı) adlı 12 adet aylık telekomünikasyon verisine ait değişkenlerin PCA'ya göre analize katkısının az olduğu gözlenmiştir. Sınıflandırma algoritmaları ile kurulan modellerde bu 12 değişkenin dahil olması veya çıkarılması analiz sonuçlarını değiştirmemiştir. Çıkarılan 12 değişkenden 4 tanesi tekil geriye kalan 8 taneside 6 aylık telekomünikasyon verisini tutmaktadır yani toplam 52 değişken sınıflandırma algoritmaları ile kurulan modellere ve veri görselleştirmeye katılmamıştır. Sınıflandırma algoritmaları ile oluşturulan modeller 119 (171-52) değişken ile yapılmıştır.

5.5. MODELLLEME

Bu tezde müşteri kaybını tahmin için sınıflandırma modelleri oluşturulmuştur. Bu sınıflandırma modellerinin telekomünikasyon veri seti üzerinde kurulmasında ise Bölüm 4.1.4'te anlatılan Naive Bayes Algoritması, k-En Yakın Komşu Algoritması, C4.5 Karar Ağacı ve ID3 Algoritmalarından faydalanılmıştır.

Algoritmalar ile oluşturulan modellerin hangisinin daha iyi bir sonuç verdiğini karşılaştırmak için Bölüm 4.1.5.1'de anlatılan hold-out ve çapraz geçerleme performans değerlendirme ve model seçim yöntemleri göz önünde bulundurularak telekomünikasyon veri seti ile yapılan analizlerde iki farklı yöntem kullanılmıştır. Bu yöntemlerden birincisinde 4-kat çapraz geçerleme, 5-kat çapraz geçerleme ve 10-kat çapraz geçerleme yöntemi; ikincisinde ise hold-out yöntemi ile %60-%40, %75-%25, %80-%20 ayrım oranlarına sahip sırasıyla eğitim-test veri seti ayrımları yapılmıştır. Böylece iki farklı performans değerlendirme ve model seçim yöntemi kullanılmış olup performans değerlendirme sonuçlarının birbiriyle tutarlı sonuçlar vermesi incelenmiştir. Bu sonuçlar elde edilirken PCA işlemi ile değişkenler azaltılmıştır. Fakat PCA işlemine maruz kalan veri seti ile PCA uygulanmamış veri seti ayrı ayrı modellere uygulanarak yapılan değişken azaltma işleminin veri seti hakkında manüpulatif sonuç verip vermediği kontrol edilmiştir. Değişken azaltmaya yönelik işlemin karşılaştırılması Tablo 5.3'te görülmektedir. Bu karşılaştırılmada örnek olarak hold-out yöntemi ile %60 eğitim ve %40 test veri seti ayrımı yapılarak performans değerlendirmesi yapılmıştır.

Tablo 5.3. PCA ile değişken azaltılma işlemi uygulanan ve uygulanmayan veri seti karşılaştırılması.

	PCA Uygulanan Veri seti		PCA Uygulanmayan Veri seti	
Modeller	DOĞRULUK	HATA	DOĞRULUK	HATA
C4.5	0,980	0,019	0,9799	0,0191
ID3	0,9740	0,025	0,9740	0,025
Gini	0.970	0.029	0.9699	0.0291
NB	0,787	0,212	0,7822	0,2168
k-nn	0,980	0,019	0,9775	0,0215

Tablo 5.3'te görüldüğü gibi PCA işleminin uygulanması ve uygulanmaması ile oluşturulan modeller arasındaki sonuçlar benzer hatta aynıdır. Bundan dolayı PCA ile verisetinin azaltılmış hali çalışmamızda kullanılmıştır.

Çapraz geçerleme performans değerlendirme ve model seçim yönteminden 4-kat çapraz geçerleme, 5-kat çapraz geçerleme ve 10-kat çapraz geçerleme performans değerlendirme ölçütlerine tabi tutulmuş ve sonuçta elde edilen değerlerin ortalama doğruluk değeri esas alınmıştır. Aynı şekilde hold-out yöntemi ile %60-%40, %75-%25, %80-%20 ayrım oranlarına sahip sırasıyla eğitim-test veri setine ayrılmış ve test veri setini tahmin etmede performans değerlendirme ölçütlerine tabi tutulmuş ve sonuçta doğruluk, hata, tanısal üstünlük oranı, F-measure değerlerine bakılmıştır. Bölüm 4.1.5.2'de verilen model performans değerlendirme ölçütlerinden tüm algoritmaların performansını birbiriyle karşılaştırabilmek için yapılan işlemler anlatılmaktadır. Kurulan modeller ve performans değerlendirme çalışmalarının kodları EK-2' de yer almaktadır.

Bu tez çalışması R aracılığı ile gerçekleştirilmiştir. R, istatistiksel hesaplamalara imkan sağlayan ücretsiz bir programdır [101]. Bu tez kapsamında arayüzü esnek ve ayarlanabilen RStudio IDE'si kullanılmıştır. RStudio, açık kaynak kodlu veya ticari olmak üzere kullanılabilir. Kullanıcılara rahat bir çalışma ortamı sunmaktadır [102]. Tez çalışmasında en iyi performans veren modelin dinamik web tabanlı uygulaması olarak RStudio projesi olan Shiny ile de bir çalışma yapılmıştır. RStudio da Shiny ile R kodlarını internete aktarmaya imkan veren bir R paket uygulamasıdır [103]. Yapılan uygulamaları internet üzerinde paylaşmanın yolu shiny server sayfasında (shinyapps.io) yayınlamaktır [104].

6. BULGULAR

Bulgular bölümünde, Naive Bayes algoritması, *k*-en yakın komşu algoritması, C4.5, ID3 ve Gini karar ağacı algoritmalarının telekomünikasyon veri seti üzerinde uygulanması ve elde edilen modellerin karşılaştırma sonuçlarına yer verilmiştir. Her model için hedef değişken, performans değerlendirme ve model seçim yöntemi ve kullanılan R kütüphaneleri sırasıyla Tablo 6.1, Tablo 6.2, Tablo 6.3 ve Tablo 6.4'te gösterilmiştir. Bu tablolarda belirtilen veriler ile modellerin analizinden sonra oluşan sonuçlar Bölüm 6.3.1'de yer alan model performans karşılaştırma tablosunda gösterilmektedir.

6.1. NAIVE BAYES ALGORİTMASI İLE MODEL KURMA

Tablo 6.1. Naive Bayes algoritma model özeti.

Hedef Değişken	Churn (E/H-Evet/Hayır)
Performans Değerlendirme ve Model Seçim Yöntemi	• 4-kat çapraz geçerleme, 5-kat çapraz geçerleme ve 10-kat çapraz geçerleme • %60-%40, %75-%25, %80-%20 oranlarında hold-out
Kullanılan R Kütüphaneleri	• csv dosyasından veri okuma, • TunePareto [105]: Veri setini çapraz geçerleme için parçalara ayırmak için gerekli paket, • e1071 [106] Naive Bayes algoritmasını uygulayabilmek için gerekli paket.

Naive Bayes modeli oluşturulurken Bölüm 4.1.5.1'de yer alan performans değerlendirme ve model seçim yöntemlerinde de anlatılan hold-out ve k-kat çapraz geçerleme kullanılmıştır.

Tablo 6.1'de görüldüğü gibi hold-out yöntemde eğitim ve test veri setinin %60-%40, %75-%25, %80-%20 şeklinde sırayla ayrımı yapılmıştır. k-kat çapraz geçerlemedende 4-kat çapraz geçerleme, 5-kat çapraz geçerleme ve 10-kat çapraz geçerleme kullanılmıştır. Modellerin karşılaştırılması Bölüm 6.3.1'de yer alan model performans karşılaştırması başlığı altında gösterilmektedir. Performans değerlendirme ve model seçim yöntemlerine ait kodlar EK-2'de yer almaktadır.

6.2. EN YAKIN KOMŞU ALGORİTMASI İLE MODEL KURMA

Tablo 6.2. *k*-en yakın komşu model özeti.

Hedef Değişken	Churn (E/H-Evet/Hayır)
Performans Değerlendirme ve Model Seçim Yöntemi	• 4-kat çapraz geçerleme, 5-kat çapraz geçerleme ve 10-kat çapraz geçerleme • %60-%40, %75-%25, %80-%20 oranlarında hold-out
Kullanılan R kütüphaneleri	• csv dosyasından veri okuma, • TunePareto [105]: Veri setini çapraz geçerleme için parçalara ayırmak için gerekli paket, • knnGarden [107]: *k*-En Yakın Komşu algoritmasını uygulayabilmek için gerekli paket.

k-En Yakın Komşu Algoritması ile kurulan model performansını diğer algoritmalarla kıyaslamak için önce en iyi performansı gösteren k değeri bulunarak seçilmiştir. En iyi k değeri elde edebilmek için Tablo 6.2'de görüldüğü gibi 4-kat çapraz geçerleme, 5-kat çapraz geçerleme ve 10-kat çapraz geçerleme ile algoritma $k=4$, 5 ve 10 için ayrı ayrı uygulanmıştır. *k*-En yakın komşu algoritması ile oluşturulan modelin performans değerlendirme yöntemlerinden çapraz geçerlemede en iyi k değeri Bölüm 6.3.1'de yer alan model performans karşılaştırma tablosunda da görüldüğü gibi 10-kat çapraz geçerlemede vermiştir. Ayrıca hold-out yöntemi ile de performans değerlendirilmesi için eğitim ve test veri seti sırasıyla %60-%40, %75-%25, %80-%20 şeklinde ayrım yapılmıştır.

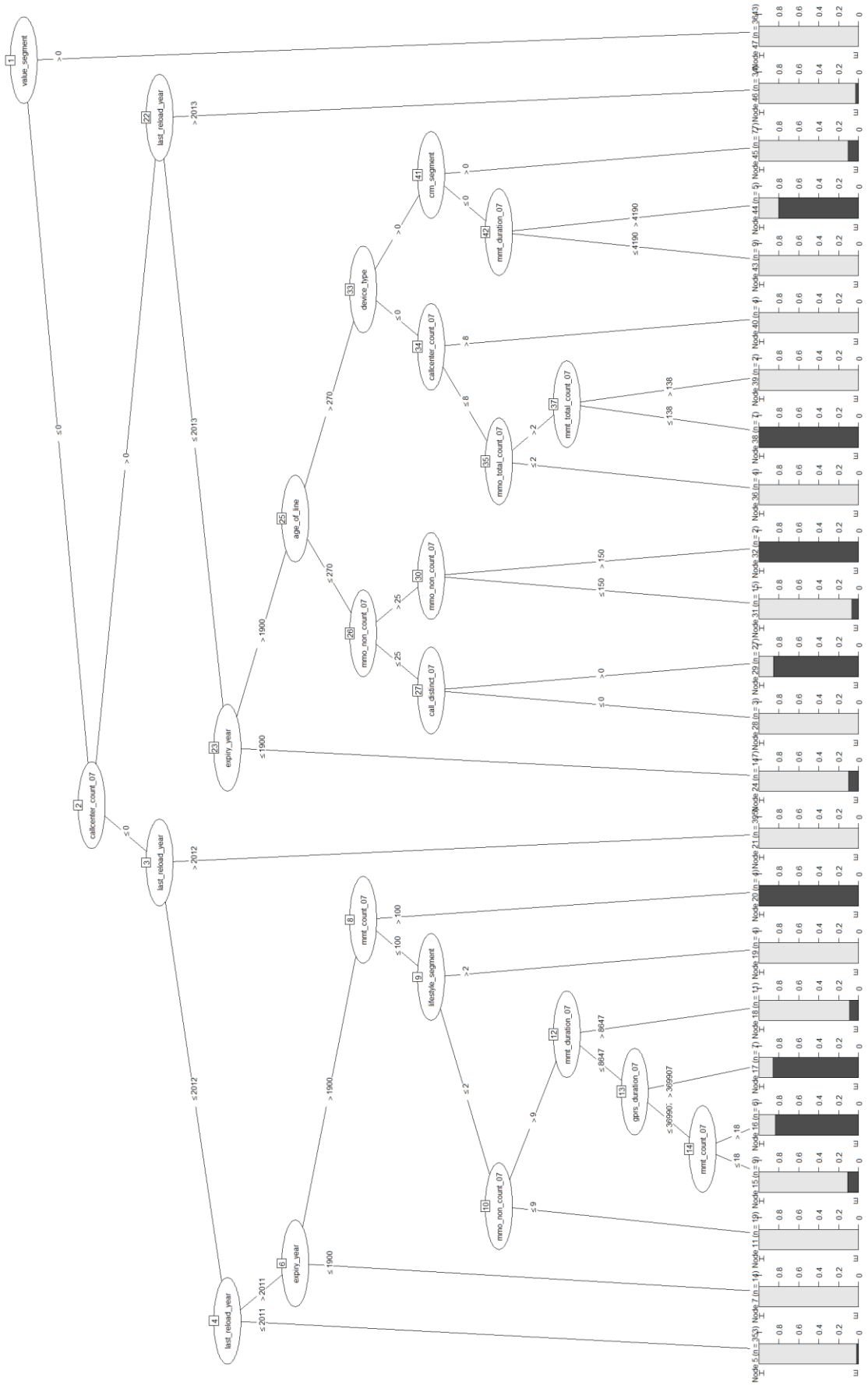
Modellerin karşılaştırılması Bölüm 6.3.1'de yer alan model performans karşılaştırması başlığı altında gösterilmektedir. K-kat çapraz geçerleme ve hold-out performans değerlendirme ve model seçim yöntemlerine ait kodlar EK-2'de yer almaktadır.

6.3. KARAR AĞACI ALGORİTMALARI İLE MODEL KURMA

Tablo 6.3. Karar ağacı algoritmaları model özeti.

Hedef Değişken	Churn (E/H-Evet/Hayır)
Algoritma	C4.5, ID3 ve Gini
Performans Değerlendirme ve Model Seçim Yöntemi	• 4-kat çapraz geçerleme, 5-kat çapraz geçerleme ve 10-kat çapraz geçerleme • %60-%40, %75-%25, %80-%20 oranlarında hold-out
Kullanılan R kütüphaneleri	• csv dosyasından veri okuma, • TunePareto [105]: Veri setini çapraz geçerleme için parçalara ayırmak için gerekli paket, • RWeka [108],[109], partykit [110]: ID3 ve C4.5 karar ağacı ile çalışır. • Rpart [113] gini karar ağacı ile çalışır. • Rpart.plot [114] gini karar ağacında ağaç elde etmek için kullanılır.

ID3, C4.5 ve Gini karar ağacı algoritmaları ile telekomünikasyon veri seti üzerine kurulan modellerin performansları Tablo 6.3'te görüldüğü gibi çapraz geçerleme ve hold-out performans değerlendirme yöntemleri ile belirlenmiştir. 4-kat çapraz geçerleme, 5-kat çapraz geçerleme ve 10-kat çapraz geçerleme performans değerlendirme yöntemi ile hold-out performans değerlendirme yönteminde eğitim ve test veri seti ayrımı olarak sırasıyla %60-%40, %75-%25, %80-%20 eğitim-test veri seti ayrımları uygulanmıştır. Bu algoritmalar için de özel bir parametre seçimi yapılmamıştır. Bu çalışmada oluşturulan modellerden en iyi sonucu C4.5 karar ağacı algoritması vermiştir. Şekil 6.1'de %60-%40 eğitim-test verisi hold-out ayrımıyla elde edilen C4.5 karar ağacının ekran görüntüsü görülmektedir. Modellerin karşılaştırılması Bölüm 6.3.1'de yer alan model performans karşılaştırması başlığı altında gösterilmektedir. K-kat çapraz geçerleme ve hold-out performans değerlendirme ve model seçim yöntemlerine ait kodlar EK-2'de yer almaktadır.



Şekil 6.1. %60 ayrımla elde edilen karar ağacının görüntüsü.

Şekil 6.1'deki ağaç şekline bakarak modelin yazdırılması pek mümkün görülmemektedir. Şekil 6.2'deki gibi ağacın yazdırılması ile elde edilen ayrımların ekran görüntüsü verilmektedir. Şekil 6.2'deki ağaç modelinin ekran görüntüsünde (353.0/8.0) şeklinde parantez içinde görülen değerler doğru (353) ve yanlış (8) sınıflandırılan gözlemleri göstermektedir.

```
J48 pruned tree
------------------

value_segment <= 0
|   callcenter_count_07 <= 0
|   |   last_reload_year <= 2012
|   |   |   last_reload_year <= 2011: H (353.0/8.0)
|   |   |   last_reload_year > 2011
|   |   |   |   tariff_type <= 1
|   |   |   |   |   mmt_count_07       <= 101
|   |   |   |   |   |   mmo_non_count_07  <= 9: H (24.0)
|   |   |   |   |   |   mmo_non_count_07  > 9
|   |   |   |   |   |   |   mmt_duration_07   <= 8647
|   |   |   |   |   |   |   |   mmt_duration_07   <= 956: H (5.0)
|   |   |   |   |   |   |   |   mmt_duration_07   > 956
|   |   |   |   |   |   |   |   |    mmt_total_duration_07  <= 17691: H (11.0/4.0)
|   |   |   |   |   |   |   |   |    mmt_total_duration_07  > 17691: E (8.0)
|   |   |   |   |   |   |   mmt_duration_07   > 8647: H (13.0/1.0)
|   |   |   |   |   mmt_count_07       > 101: E (4.0)
|   |   |   |   tariff_type > 1: H (9.0)
|   |   last_reload_year > 2012: H (395.0)
|   callcenter_count_07 > 0
|   |   last_reload_year <= 2013
|   |   |   tariff_type <= 1
|   |   |   |   age_of_line <= 270
|   |   |   |   |   mmo_non_count_07  <= 25
|   |   |   |   |   |   mmo_non_duration_07 <= 3: H (3.0)
|   |   |   |   |   |   mmo_non_duration_07 > 3
|   |   |   |   |   |   |   callcenter_count_07 <= 1: E (23.0/2.0)
|   |   |   |   |   |   |   callcenter_count_07 > 1
|   |   |   |   |   |   |   |   mmt_count_07       <= 5: H (2.0)
|   |   |   |   |   |   |   |   mmt_count_07       > 5: E (4.0/1.0)
|   |   |   |   |   mmo_non_count_07  > 25
|   |   |   |   |   |   payment_type_07 <= 25: H (17.0/1.0)
|   |   |   |   |   |   payment_type_07 > 25: E (2.0)
|   |   |   |   age_of_line > 270
|   |   |   |   |   device_type <= 0
|   |   |   |   |   |   last_reload_year <= 1900: H (4.0)
|   |   |   |   |   |   last_reload_year > 1900
|   |   |   |   |   |   |   callcenter_count_07 <= 8
|   |   |   |   |   |   |   |   mmt_count_07       <= 63
|   |   |   |   |   |   |   |   |   mmt_total_count_07      <= 138: E (7.0)
|   |   |   |   |   |   |   |   |   mmt_total_count_07      > 138: H (2.0)
|   |   |   |   |   |   |   |   mmt_count_07       > 63: H (3.0)
|   |   |   |   |   |   |   callcenter_count_07 > 8: H (3.0)
|   |   |   |   |   device_type > 0
|   |   |   |   |   |   crm_segment <= 0
|   |   |   |   |   |   |   mmt_duration_07   <= 4190: H (9.0)
|   |   |   |   |   |   |   mmt_duration_07   > 4190: E (5.0/1.0)
|   |   |   |   |   |   crm_segment > 0: H (81.0/8.0)
|   |   |   tariff_type > 1
|   |   |   |   gender_flag <= 2
|   |   |   |   |   lifestyle_segment <= 1: H (36.0/2.0)
|   |   |   |   |   lifestyle_segment > 1
|   |   |   |   |   |   crm_segment <= 1
|   |   |   |   |   |   |   payment_type_07 <= 18: H (15.0/1.0)
|   |   |   |   |   |   |   payment_type_07 > 18
|   |   |   |   |   |   |   |   mmt_count_07       <= 62
|   |   |   |   |   |   |   |   |   mmo_total_duration_07   <= 44201: E (7.0)
|   |   |   |   |   |   |   |   |   mmo_total_duration_07   > 44201: H (2.0)
|   |   |   |   |   |   |   |   mmt_count_07       > 62: H (6.0)
|   |   |   |   |   |   crm_segment > 1
|   |   |   |   |   |   |   age <= 47: H (27.0)
|   |   |   |   |   |   |   age > 47
|   |   |   |   |   |   |   |   age <= 50: E (3.0)
|   |   |   |   |   |   |   |   age > 50: H (3.0)
|   |   |   |   gender_flag > 2: H (38.0/1.0)
|   |   last_reload_year > 2013: H (34.0/1.0)
value_segment > 0: H (3643.0)

Number of Leaves  :     34
```

Şekil 6.2. Modelin yazdırılması ve elde edilen ayrımların görüntüsü.

Şekil 6.2'nin incelenmesi ile aşağıda örnek olarak verilmiş kurallar görülmektedir. Bu şekilde kurallar oluşturulabilir.

Kural-1: Eğer Value_segment<=0 VE callcenter_count_07<=0 VE last_reloaded_year<=2012 VE last_reload_year<=2011 İSE müşteri ayrılmaz (Churn durumu: H)

Kural-2: Eğer Value_segment<=0 VE callcenter_count_07<=0 VE last_reloaded_year<=2012 VE last_reload_year>2011 VE tariff_type<=1 VE mmt_count_07<=101 VE mmo_non_count_07<=9 İSE müşteri ayrılmaz (Churn durumu: H)

Kural-3: Eğer Value_segment<=0 VE callcenter_count_07<=0 VE last_reloaded_year<=2012 VE last_reload_year>2011 VE tariff_type<=1 VE mmt_count_07<=101 VE mmo_non_count_07>9 VE mmt_duration_07<=8647 VE mmt_duration_07<=956 İSE müşteri ayrılmaz (Churn durumu: H)

Kural-4: Eğer Value_segment<=0 VE callcenter_count_07<=0 VE last_reloaded_year<=2012 VE last_reload_year>2011 VE tariff_type<=1 VE mmt_count_07<=101 VE mmo_non_count_07>9 VE mmt_duration_07<=8647 VE mmt_duration_07>956 VE mmt_total_duration_07<=17691 İSE müşteri ayrılmaz (Churn durumu: H)

Kural-5: Eğer Value_segment<=0 VE callcenter_count_07<=0 VE last_reloaded_year<=2012 VE last_reload_year>2011 VE tariff_type<=1 VE mmt_count_07>101 İSE müşteri ayrılır (Churn durumu: E)

Kural-6: Eğer Value_segment<=0 VE callcenter_count_07<=0 VE last_reloaded_year<=2012 VE last_reload_year>2011 VE last_reload_year>2012 İSE müşteri ayrılmaz (Churn durumu: H)

Kural-7: Eğer Value_segment<=0 VE callcenter_count_07>0 VE last_reload_year<=2013 VE tariff_type<=1 VE age_of_line<=270 VE mmo_non_count_07<=25 VE mmo_non_duration_07<=3 İSE müşteri ayrılmaz (Churn durumu: H)

Kural-8: Eğer Value_segment<=0 VE callcenter_count_07>0 VE last_reload_year<=2013 VE tariff_type<=1 VE age_of_line<=270 VE mmo_non_count_07<=25 VE mmo_non_duration_07>3 VE call_center_count_07<=1 İSE müşteri ayrılır (Churn durumu: E)

Kural-9: Eğer Value_segment<=0 VE callcenter_count_07>0 VE

last_reload_year<=2013 VE tariff_type<=1 VE age_of_line>270 VE device_type<=0 VE last_reload_year<=1900 İSE müşteri ayrılmaz (Churn durumu: H)

Kural-10: Eğer Value_segment>0 İSE müşteri ayrılmaz (Churn durumu: H)

6.3.1. Model Performans Karşılaştırması

Bu bölümde tüm analizlerden elde edilen sonuçlar iki alt başlık halinde tablosal olarak gösterilmiştir. İlk olarak telekomünikasyon veri setinin 4-kat çapraz geçerleme, 5-kat çapraz geçerleme ve 10-kat çapraz geçerleme ile elde edilen sonuçlar görülmektedir.

İkinci olarak da %60-%40, %75-%25, %80-%20 oranlarında hold-out uygulanarak elde edilen sonuçlar görüntülenmektedir.

6.3.1.1. *4-kat Çapraz Geçerleme, 5-kat Çapraz Geçerleme ve 10-kat Çapraz Geçerleme Performans Değerlendirme ve Model Seçim Yöntemi ile Elde Edilen Sonuçlar*

Bu çalışmada, birden fazla model ile oluşturulmuş ve algoritmaların performans değerlerini ölçmek için 4-kat çapraz geçerleme, 5-kat çapraz geçerleme ve 10-kat Çapraz Geçerleme yöntemleri kullanılmıştır. Tablo 6.4'te, 4-kat Çapraz Geçerleme, 5-kat Çapraz Geçerleme ve 10-kat çapraz geçerleme ile elde edilen sonuçlar gösterilmektedir. Tüm tablolarda en iyi sonucu veren model olarak C4.5 karar ağacı ile elde edilen model olmuştur. Ayrıca ID3 modeli ile elde edilen değerlerin C4.5 algoritması ile elde edilen modele yakın sonuç vermesi C4.5 algoritmasının ID3 algoritmasının gelişmiş hali olduğunu doğrulamaktadır.

Tablo 6.4. 4-Kat, 5-Kat ve 10-Kat çapraz geçerleme performans değerlendirme sonuçları.

Modeller	4-Kat Çapraz Geçerleme		5-Kat Çapraz Geçerleme		10-Kat Çapraz Geçerleme	
	Doğruluk	Hata	Doğruluk	Hata	Doğruluk	Hata
Bayes	0,8050	0.149	0,8050	0.149	0,8056	0.1984
C4.5	0,9810	0.018	0,9831	0.0159	0,9838	0.0152
ID3	0,9806	0.0184	0,9819	0.0171	0,9832	0.0228
Gini	0,977	0.022	0.9810	0.018	0.9825	0.165
k-nn	0,9738	0.1952	0,9800	0.019	0,9826	0.0134

6.3.1.2. *Hold-Out Performans Değerlendirme ve Model Seçim Yöntemi ile Elde edilen Sonuçlar*

Bu çalışma sonucunda oluşturulan modelin eğitim ve test kümeleri sırasıyla %60-%40, %75-%25, %80-%20 ayrımlar ile karşılaştırılması Tablo 6.5'te gösterilmektedir. Sadece modeli doğruluk değişkeni üzerinden değerlendirmek yerine Doğruluk, Hata, Üstünlük oranı, F-ölçü değerlerine yer verilmiştir. Tablo 6.5'te görüldüğü üzere hold-out ayrım değerlerinin en optimum ayrım değeri olarak çıkan %75 eğitim ve %25 test verisi olan ayrımdır. Tablo 6.5'te dört değişkene göre en iyi sonucu veren model C4.5 karar ağacı ile elde dilen model olmuştur.

Tablo 6.5. Telekomünikasyon müşteri veri seti hold-out ayrımlarına ilişkin doğruluk, hata, tanısal üstünlük oranı, F-ölçü değerleri.

	DOĞRULUK			HATA			DOR (TANISAL ÜSTÜNLÜK ORANI)			F-ÖLÇÜ		
Yüzde	60	75	80	60	75	80	60	75	80	60	75	80
C4.5	0,980	0,980	0,983	0,019	0,019	0,016	92,11	97,05	165,62	0,990	0,990	0,991
ID3	0,974	0,979	0,978	0,025	0,020	0,021	16,20	63,77	41,57	0,987	0,989	0,989
Gini	0.970	0.975	0.975	0.029	0.025	0.024	11.21	15.96	25.92	0.98	0.987	0.98
NB	0,787	0,803	0,800	0,212	0,196	0,199	5,411	9,324	7,438	0,879	0,889	0,887
k-nn	0,980	0,981	0,980	0,019	0,018	0,019	0	0	0	0,990	0,990	0,990

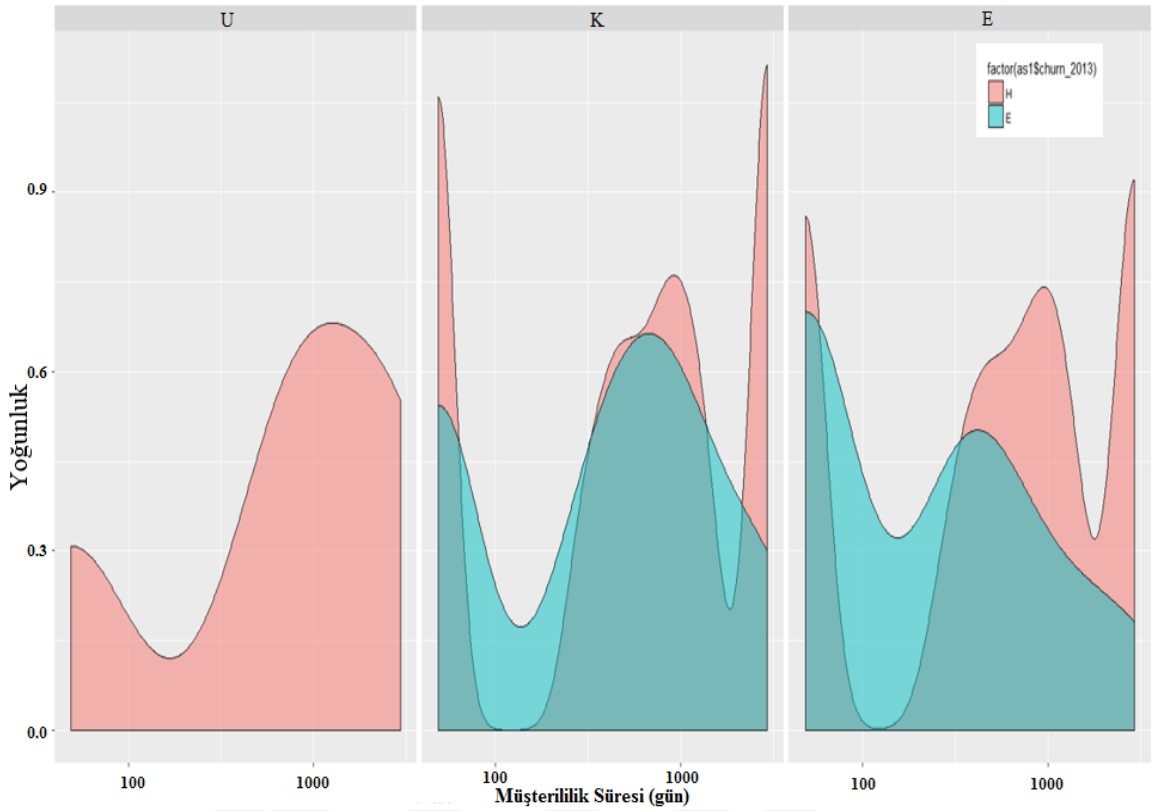
7. VERİ MADENCİLİĞİ YOLUYLA VERİ GÖRSELLEŞTİRME

Bu bölümde, R programında yer alan bazı paketler yardımı ile bir Telekomünikasyon veri seti üzerinde çeşitli yorumlar yapılabilecek grafikler oluşturulması anlatılmıştır. Çalışma kapsamında elde edilen grafikleri yorumlayarak çıkarım yapmada üst düzey matematik yetenekleri gerektirmemekle beraber grafiğin çıkarıldığı sektör ve veri seti hakkında ilgili bilgisi olan herkesin elindeki veri setine ilişkin buna benzer grafikleri çıkarma ve elde ettiği sonuçlarla ilgili analiz ve yorum yapabilme imkanı vardır.

Bu bölümde R grafik paketlerinden density ve violin grafikleri üzerinde yapılan bir çalışma anlatılacaktır [111]. Bu grafikler hazırlanırken bir müşterinin bir telekomünikasyon operatöründe müşteri olduğu süre ile aynı operatörde olan ve olmayan diğer telekomünikasyon müşterileri arasındaki aylık konuşma süreleriyle çeşitli kategorik değişkenlerle gruplandırılarak yorumlamak amacıyla grafikler üretilmiştir.

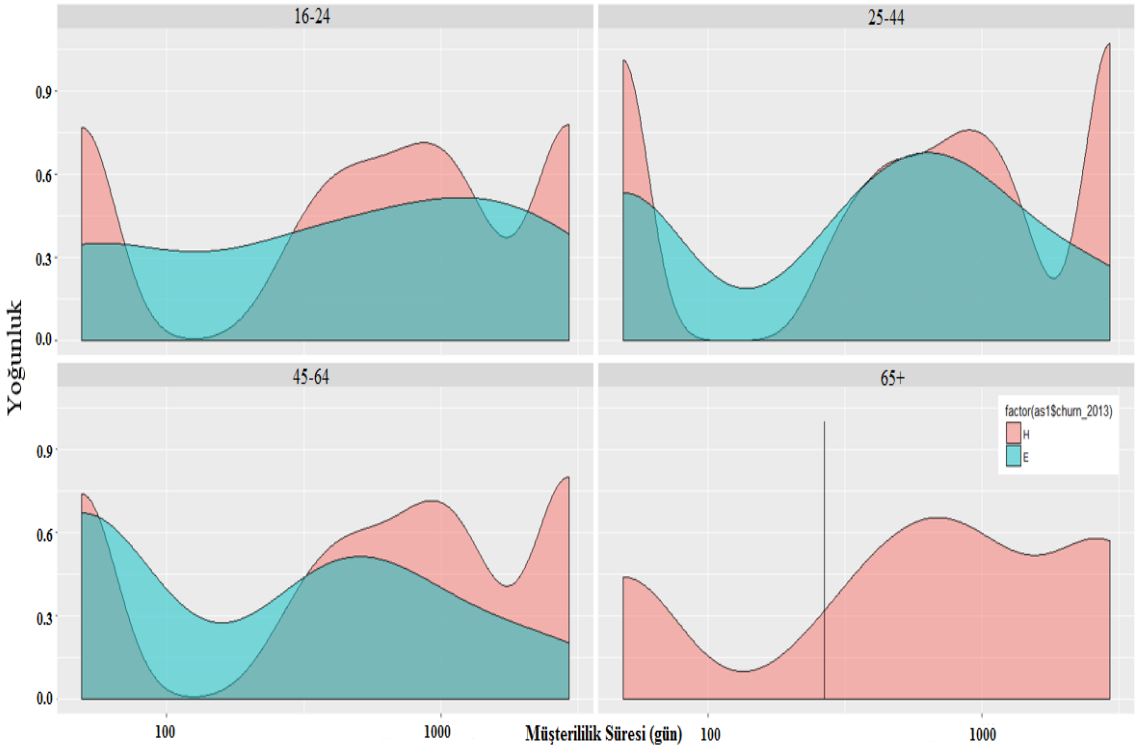
7.1. DENSITY GRAFİĞİ İLE ELDE EDİLEN GRAFİKLER

Bu bölümde R paketlerinden olan ggplot2 paketi [111] ile elde edilen density (yoğunluk) grafiği üzerine yorumlar yapılacaktır. Örnek olarak Şekil 7.1'de Müşterililik sürenin cinsiyete ile gruplandırılarak müşteri ayrım oranı görülmektedir. Ayrıca veri görselleştirme için yapılan çalışmanın kodları Ek-3'te yer almaktadır.



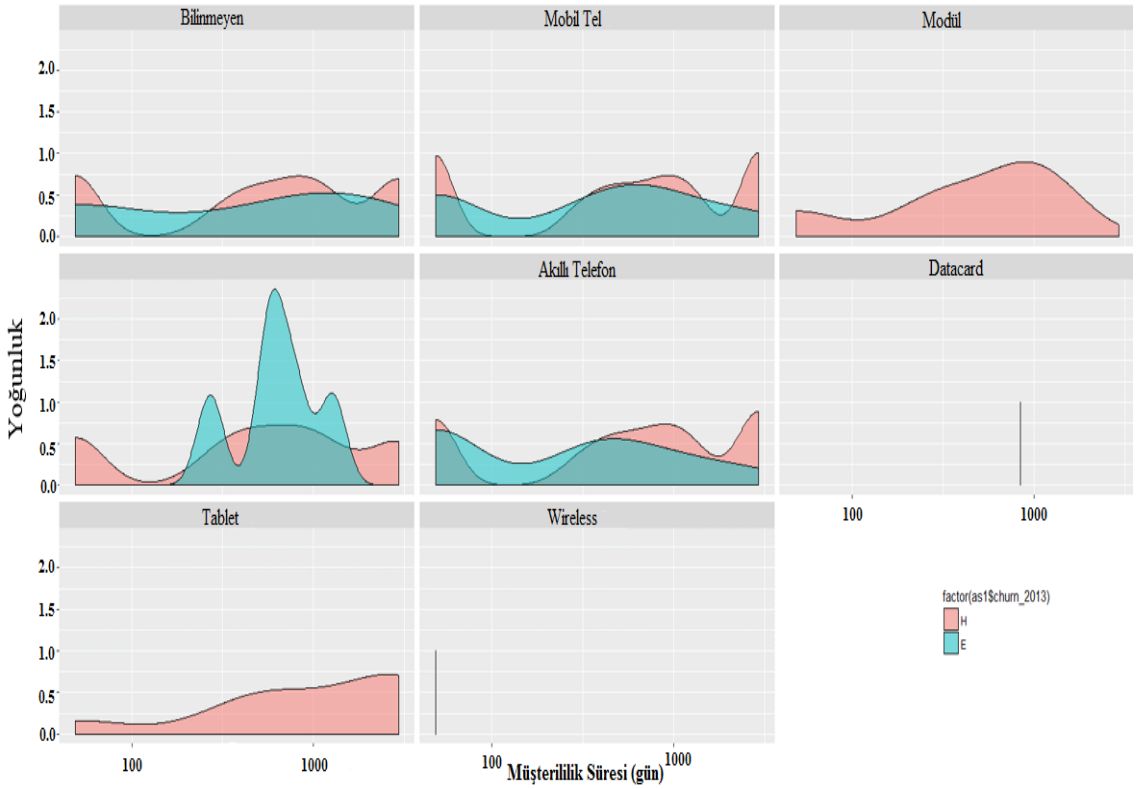
Şekil 7.1. Müşterilik süresi ile ayrılma (churn) değişkenlerinin cinsiyete göre gruplandırılması.

Şekil 7.1'de görüldüğü gibi kadın, erkek ve cinsiyet niteliği çeşitli nedenlerden dolayı girilmemiş ve bilinmeyen (U) olarak gösterilen müşterilerin ayrılma durumu görülmektedir. Şekil 1'deki grafiğe göre Erkekler (E), kadınlara (K) göre ayrılıp (E) ayrılmama (H) konusunda daha kararsız olduğu görülmektedir. Cinsiyeti çeşitli nedenlerden dolayı girilmemiş olanların hiçbirisinin ayrılmadığı görülmektedir.



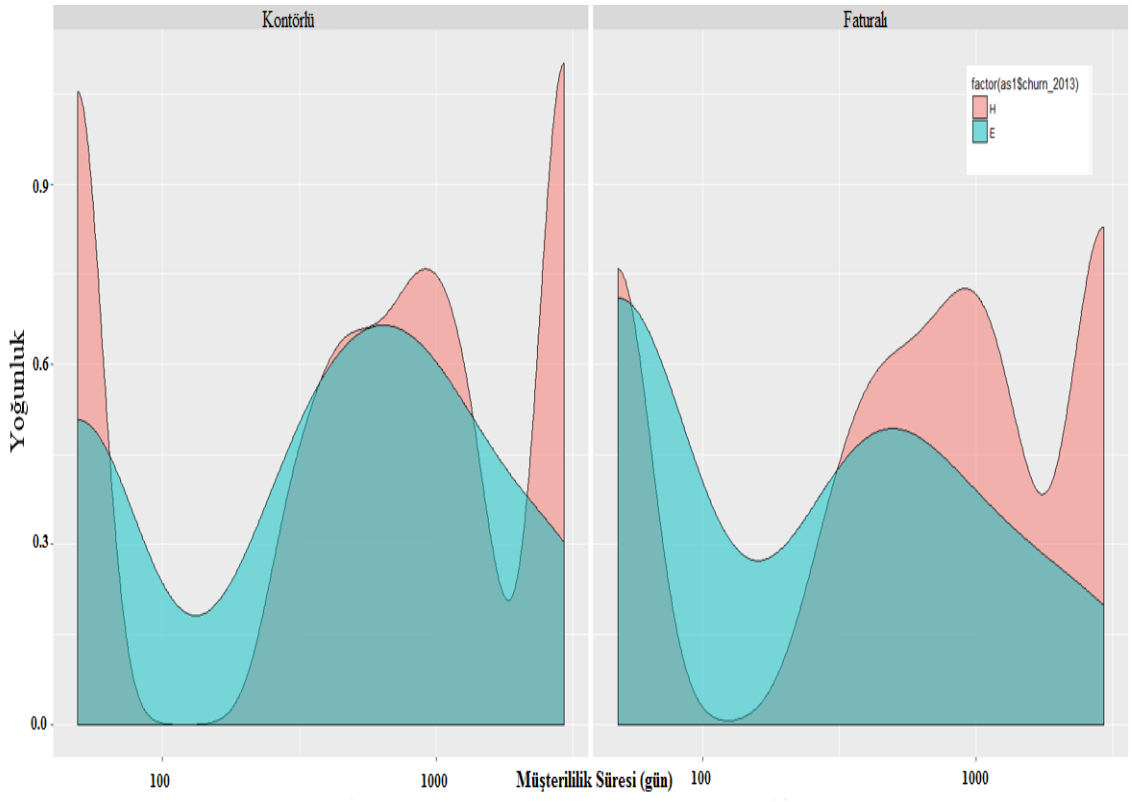
Şekil 7.2. Müşterilik süresi ile ayrılma (churn) değişkenlerinin yaşa göre gruplandırılması.

Şekil 7.2'deki grafiğe yapılabilecek yorum "65 yaşının üstündeki kullanıcılar aboneliklerinden ayrılmadıkları (H) görülmekte ve telefonlarını sadece konuşma maksatlı kullandıkları" şeklindedir. Ayrıca 16-24 ve 25-44 arası yaş gruplarında, "sürekli bir dalgalanmalı grafik görülmekte olup telefonlar sahipliği noktasında büyük ihtimalle akıllı cihazlar kullanmakta ve farklı kampanyalar arayışında oldukları" şeklinde yorum yapılabilir.



Şekil 7.3. Müşterilik süresi ile ayrılma (churn) değişkenlerinin kullanılan cihaza göre gruplandırılması.

Şekil 7.3'te telekomünikasyon veri setinde müşteri verilerinin bilinmeyen olarak girildiği bir çoğunluk olmasının yanı sıra tablet ve modül telefon kullananların ayrılmadığı (H), akıllı telefon, mobil telefon kullananlarda dalgalanma olduğu gözlenmektedir.

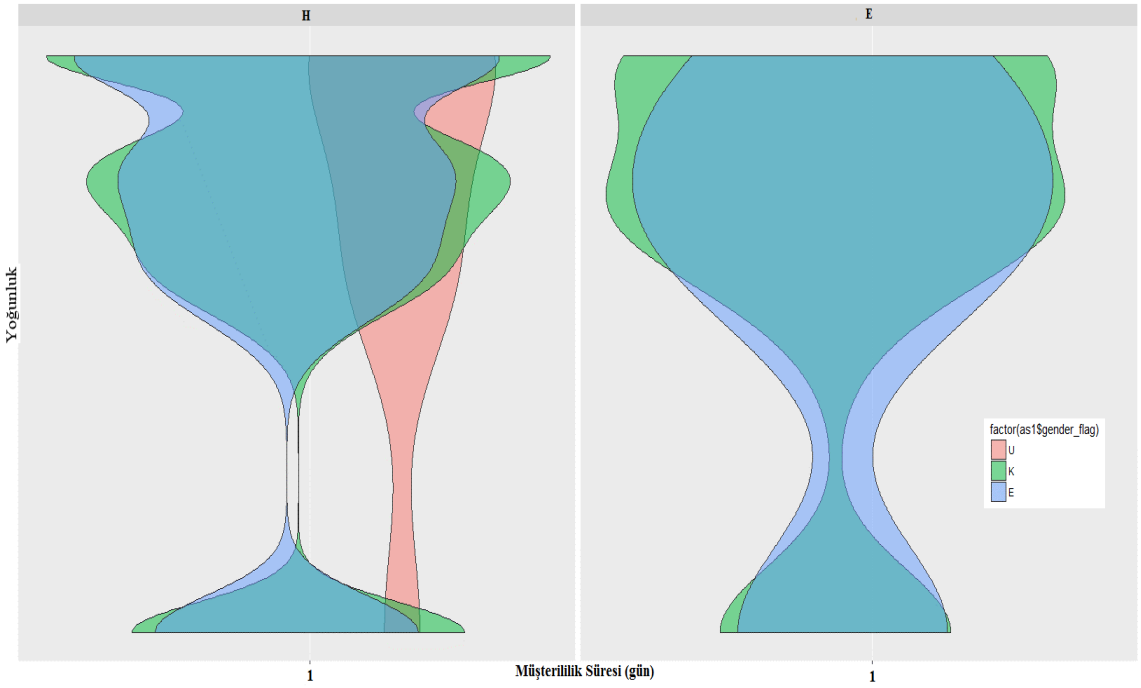


Şekil 7.4. Müşterilik süresi ile ayrılma (churn) değişkenlerinin kullanılan tarife tipine göre gruplandırılması.

Şekil 7.4'te kontörlü telefon kullananların faturalılara göre daha çok ayrıldığı görülmektedir. "Faturalı müşterilerin belki taahhütleri olduğundan dolayı ayrılmama (H) riski düşük" şeklinde yorum yapılabilir.

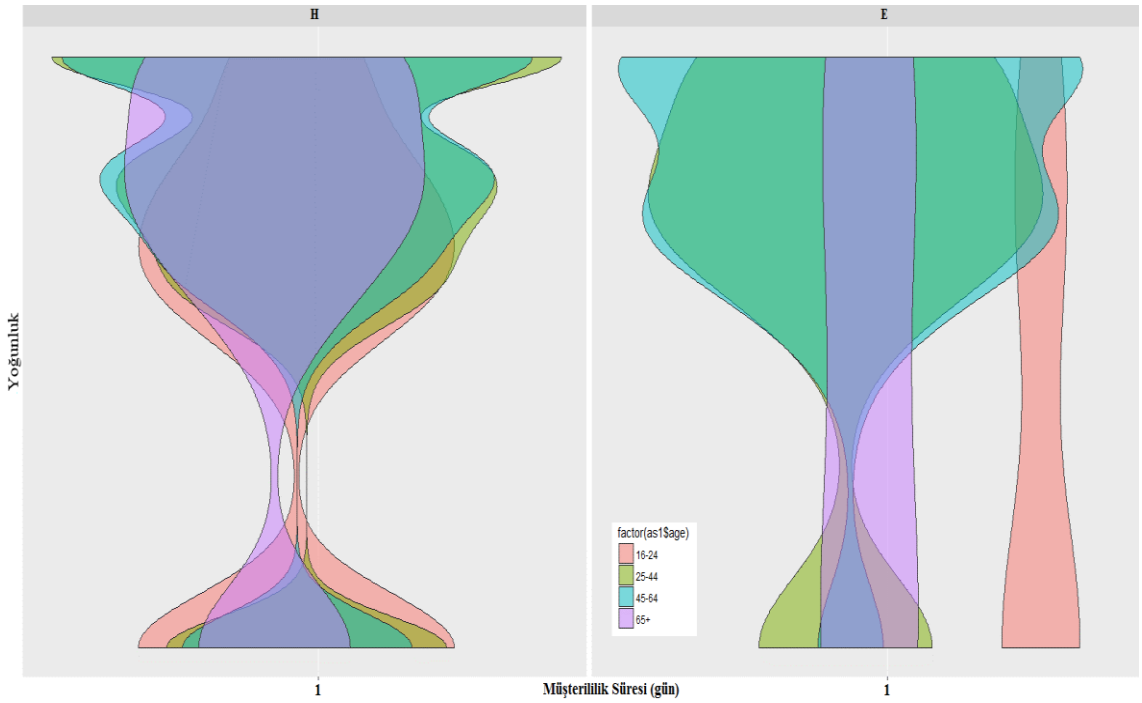
7.2. VIOLIN GRAFİĞİ İLE ELDE EDİLEN GRAFİKLER

Bu bölümde R paketlerinden olan ggplot2 paketi ile elde edilen violin grafiği üzerinde yapılan yorumlar anlatılacaktır. Violin grafiği yoğunluk grafiğinin iki boyutlu halidir.



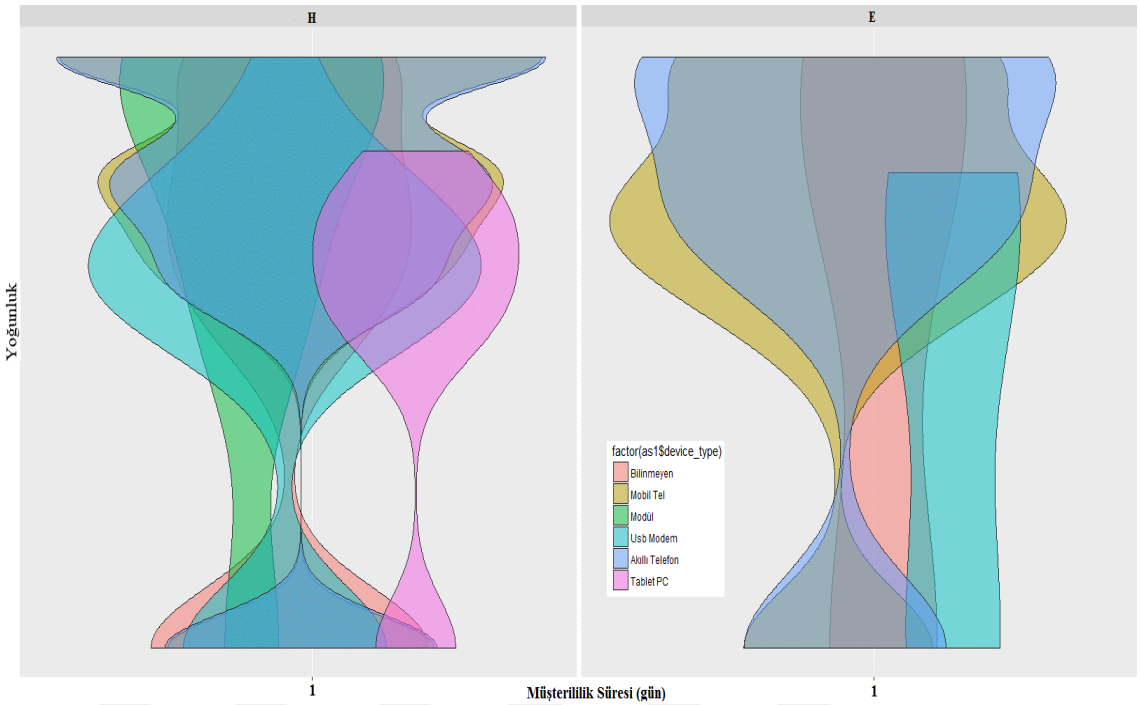
Şekil 7.5. Müşterilik süresi ile ayrılma (churn) değişkenlerinin cinsiyete göre gruplandırılması.

Şekil 7.5'teki grafiğin ayrılmayanlar (H) kısmında çeşitli nedenlerle cinsiyeti belli olmayarak veri setine cinsiyeti bilinmeyen girilenler (U) ile erkeklerin (E) ayrılması (E) kadınlara (K) göre daha kararsız olduğu görülmektedir. Density grafiğinde de benzer yorum yapılmaktadır.



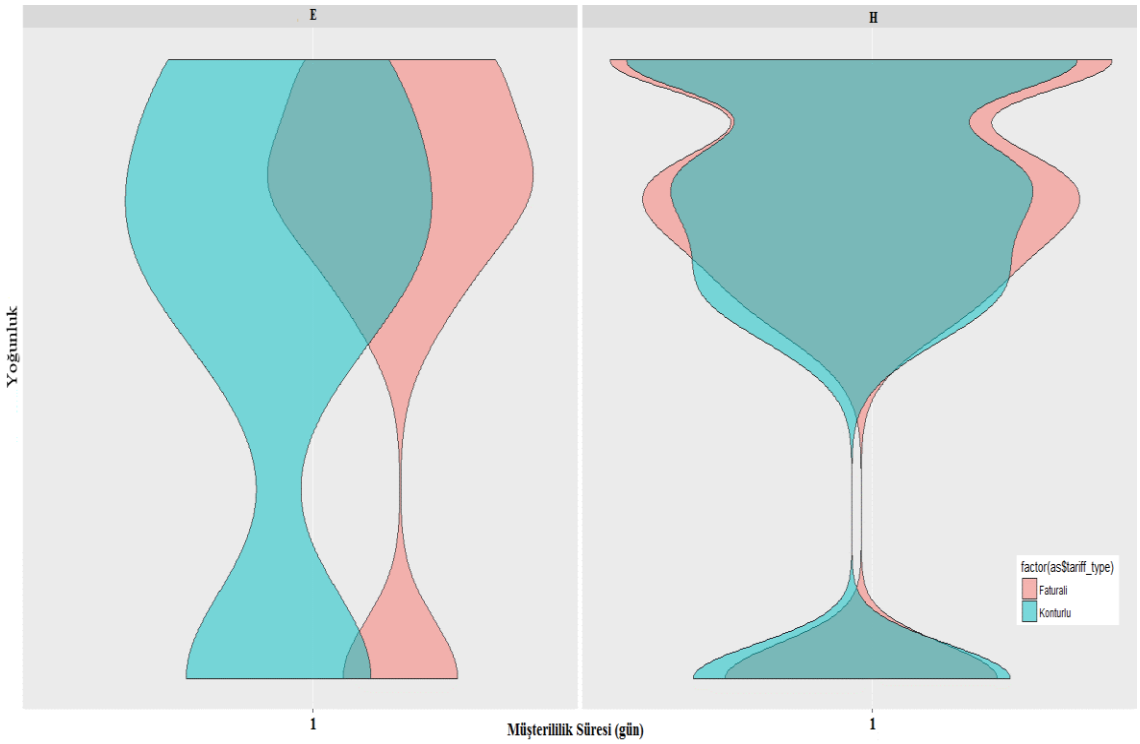
Şekil 7.6. Müşterilik süresi ile ayrılma (churn) değişkenlerinin yaşa göre gruplandırılması.

Şekil 7.6'daki grafiğe yapılabilecek yorum 65 yaşının üstündeki kullanıcılar aboneliklerinden daha az ayrıldıkları görülmekte ve density grafiğinde yapılan yorum tekrarlanarak telefonlarını sadece konuşma maksatlı kullandıkları söylenebilir. Diğer yaş gruplarında çeşitli dalgalanmalar mevcuttur. 25-44 ve 45-64 yaş grupları dengeli şekilde ayrılma (E) ve ayrılmama (H) eğilimi göstermektedir. Fakat 16-24 yaş grubunda ayrılma ve ayrılamama eğilimi dengesiz dağılmakta olması bu müşteri grubunun yeni müşteri olduğu veya yeni kampanya başlattığı yorumu yapılabilir. Bu yaş grubu, gençlerden oluştuğu için "karar verme süreçleri diğer yaş gruplarına göre daha düzensiz görünmesi normal karşılanabilir" şeklinde yorum yapılabilir.



Şekil 7.7. Müşterilik süresi ile ayrılma (churn) değişkenlerinin kullanılan cihaza göre gruplandırılması.

Şekil 7.7'deki tablet pc kullanların hiç ayrılmadığı görülmektedir. Hangi cihazı kullandıkları bilinmeyenlerin akıllı telefon kullananlara göre daha dengeli ayrıldıkları (E) görülmektedir. "Akıllı cihaz ve mobil telefon kullananların daha çok ayrılma eğiliminde olması konuşma, internet gibi yeni kampanyalar arasından en uygununun hangisi olduğu arayışında olan grup olduğu" şeklinde yorumu yapılabilir.



Şekil 7.8. Müşterilik süresi ile ayrılma (churn) değişkenlerinin kullanılan tarife tipine göre gruplandırılması.

Şekil 7.8'deki grafikte kontörlü hat kullananlar faturalılara göre daha çok ayrılma (E) eğiliminde oldukları görülebilmektedir. "Faturalı abonelerin ayrılmasının riskinin az olma sebebi taahhütleri bitmemiş" şeklinde yorum yapılabilir. Bu yorum density grafiğinde de aynı şekilde olduğu görülmektedir.

8. SHINY İLE R UYGULAMASI GELİŞTİRME VE MÜŞTERİ AYRILMA TAHMİNİ DEĞERLENDİRMEYE İLİŞKİN WEB TABANLI ÇALIŞMA

Shiny, RStudioda web tabanlı uygulama geliştirmek için kullanılan bir pakettir [103]. Shiny ile yapılan uygulamaları internet üzerinden de paylaşmanın yolu shiny server sayfasında (shinyapps.io) yayınlamaktır. Bu bölümde, çalışma boyunca kullanılan en iyi sınıflandırma algoritması seçilerek R paketlerinden Shiny ile yapılan web tabanlı uygulama anlatılacaktır. Uygulamalar geliştirilirken Shiny kütüphanesinden faydalanılmıştır [62].

Bu çalışmada, sınıflandırma algoritmaları ile kurulan modellerde kullanılan telekomünikasyon veri setinin tüm değişkenleri kullanılamamıştır. Bunun yerine müşterilerin kişisel verileri ve aylık telekomünikasyon verileri Bölüm 3.4.3'te anlatılan PCA analizindeki performanslarına göre azaltılarak alınmıştır. Müşteri kişisel verilerinden gender_flag (cinsiyet), age (yaş), lifestyle_segment (tarife türü), tariff_type (tarife tipi), device_type (cihaz tipi), crm_segment (crm), aylık telekomünikasyon verilerinden de hedef değişken olarak churn_2013_07_12 (abone churn durumu), payment_type_07_12 (fatura ödeme şekli), callcenter_count_07_12 (aylık cc şikayet arama sayısı), msmo_count_07_12 (aylık sms atma sayısı), mmo_count_07_12 (aylık kendi aboneleriyle konuşma sayısı (arama)), mmo_duration_07_12 (aylık kendi aboneleriyle konuşma süresi (arama), mmt_count_07_12 (aylık kendi aboneleriyle konuşma sayısı (aranma)), mmt_duration_07_12 (aylık kendi aboneleriyle konuşma süresi (aranma)) değişkenleri alınarak shiny uygulaması yapılmıştır.

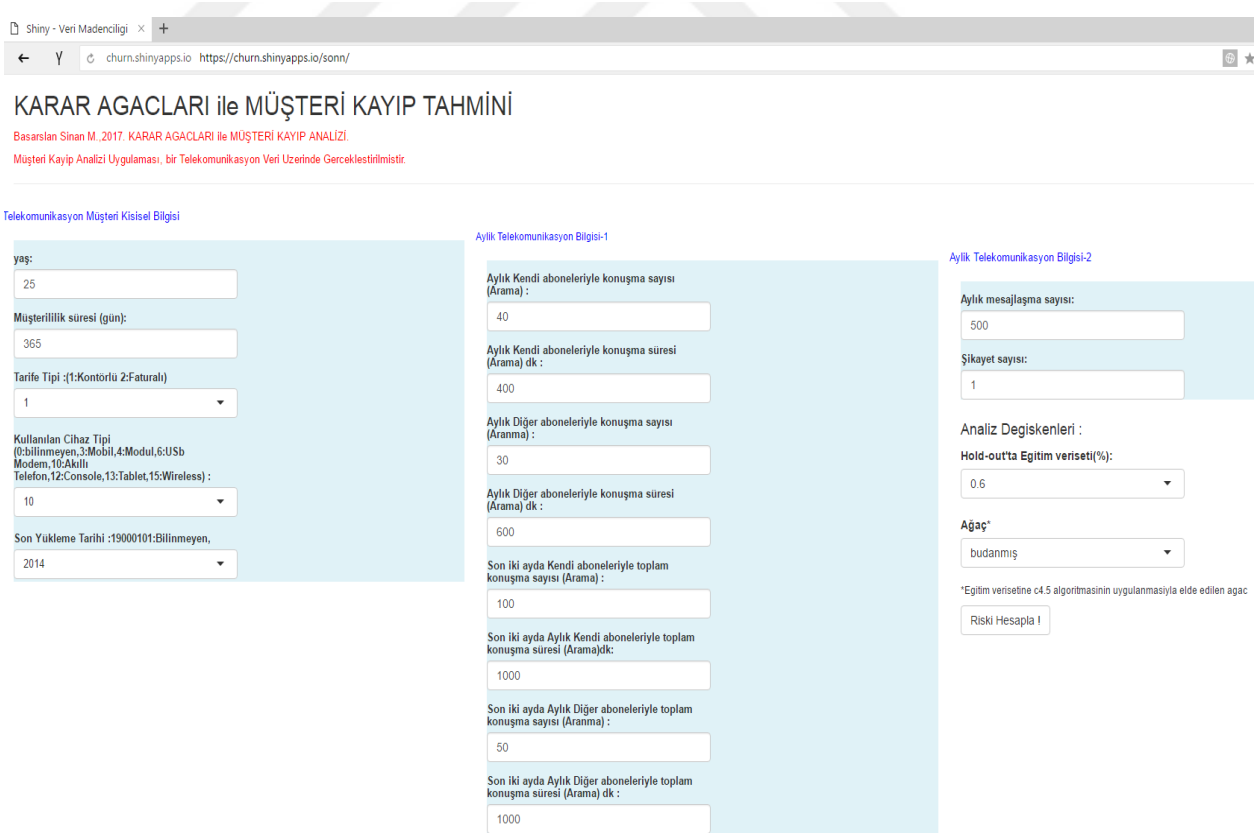
Bu çalışma boyunca RStudio [102], R Shiny paketi [104], performans değerlendirme ve model seçim yöntemlerinden hold-out için Caret [112] paketi kullanılmıştır. Shiny uygulamalarında kullanılan dosyalar da R uzantılı dosyalar olsa da bu dosyaların çalıştırılması ve bir araya getirilmesi normal R çalıştırma prensibinden farklıdır. Bu noktadaki fark, hem "kullanıcı" hemde "sunucu" taraflı olmasıdır. Kullanıcı arayüzünde kullanıcıya gösterilen metin, grafik ya da işlem başlatacak buton gibi bileşenleri içermesinin yanı sıra kullanıcıdan alınan girdilere göre çıktıları da gösterir.

Kullanıcı arayüzündeki formlar sunucu tarafı ile ilişkilendirilir. Yani kullanıcı arayüzünde kullanıcı forma veri girdiğinde bu veriler sunucu tarafına gönderilir. Sunucu tarafında R kodları ile yapılmak istenen amaca yönelik işlemler yapılarak kullanıcı arayüzünde gösterilmek üzere oluşturulan çıktılar kullanıcı arayüzüne gönderilir.

İlk olarak Rstudio'da uygulama klasörü açılır ve bu klasörün içine üç farklı dosya oluşturulur. Bunlar "ui.R", "server.R" ve "global.R" dosyalarıdır. "global.R" dosyası, "server.R" ve "ui.R" dosyaları tarafından erişilen bir dosyadır.

"global.R" dosyasında kullanılacak kütüphaneler, veri setinin okunacağı dosya dizini veya web url'nin bildirilmesi ve veri seti üzerinde yapılması gereken çeşitli işlemler bulunur. "ui.R" dosyasında kullanıcıların kişisel bilgi, aylık telekomünikasyon bilgisi gibi bilgileri girmeye ve bu bilgiler ile yapılan sınıflandırma sonucunu görmeye imkan sağlamak için tasarlanmış bir arayüz bulunur. "server.R" dosyasında da kullanıcı arayüzünden gelen bilgiler kullanılarak tanımlanan analizler yapılır.

Kullanıcı arayüzünde veri girme ve ekrana sonuçların yazdırılması ile ilgili görseller sırasıyla Şekil 8.1 ve Şekil 8.2'de görülmektedir.



Şekil 8.1. Kullanıcı arayüzünde veri girilen kısım.

ANALİZ PERFORMANS DEĞERLENDİRME:

```
Confusion Matrix and Statistics

          Reference
Prediction    E    H
         E    8   42
         H   49 3100

               Accuracy : 0.9716
                 95% CI : (0.9652, 0.977)
    No Information Rate : 0.9822
    P-Value [Acc > NIR] : 1.0000

                  Kappa : 0.1351
 Mcnemar's Test P-Value : 0.5294

            Sensitivity : 0.140351
            Specificity : 0.986633
         Pos Pred Value : 0.160000
         Neg Pred Value : 0.984440
             Prevalence : 0.017818
         Detection Rate : 0.002501
   Detection Prevalence : 0.015630
      Balanced Accuracy : 0.563492

       'Positive' Class : E
```

TAHMİN SONUCU :

```
Müşteri ayrılma riski yoktur.
```

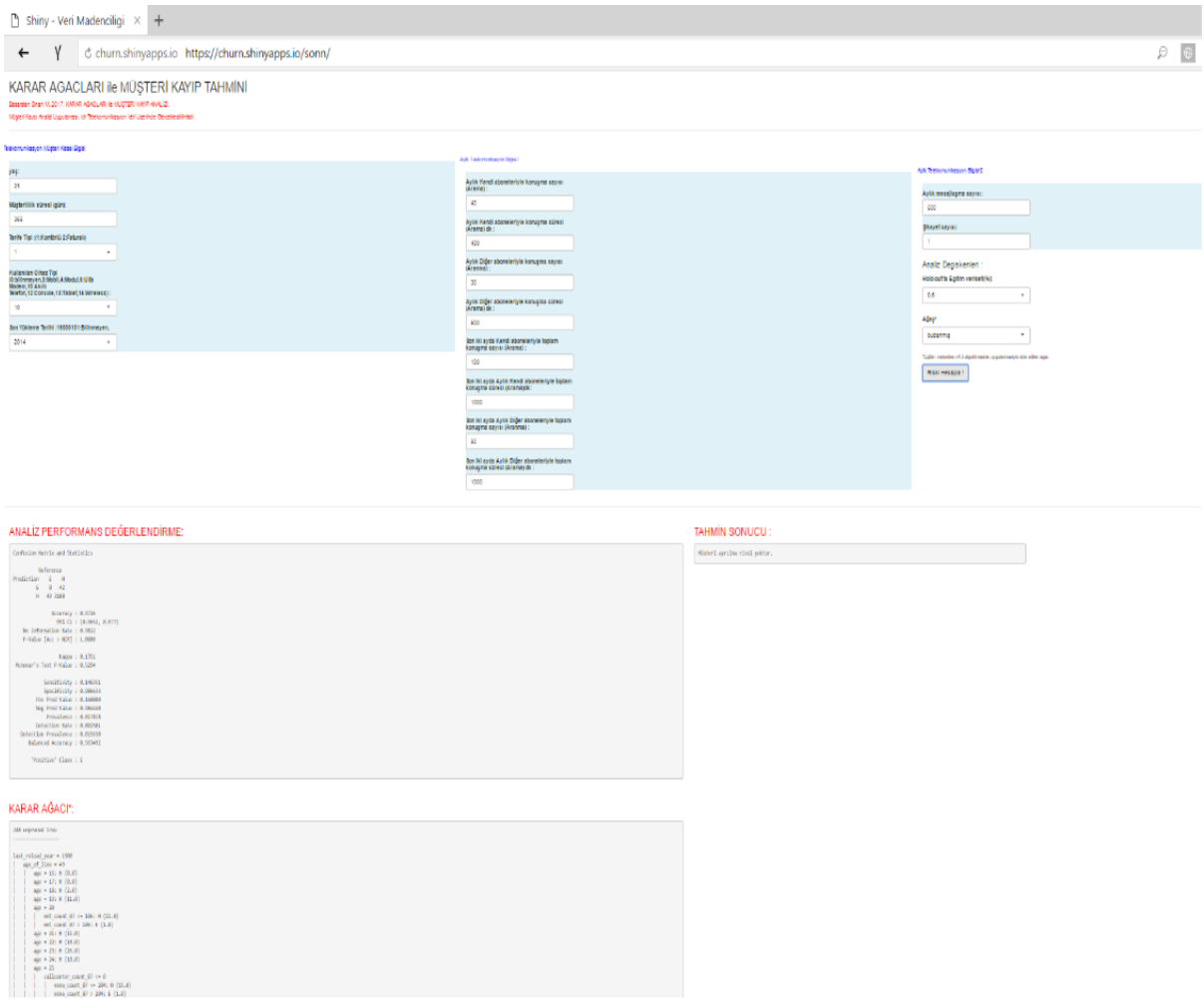
KARAR AĞACI*:

```
J48 unpruned tree
------------------

last_reload_year = 1900
|   age_of_line = 49
|   |   age = 16: H (0.0)
|   |   age = 17: H (0.0)
|   |   age = 18: H (2.0)
|   |   age = 19: H (11.0)
|   |   age = 20
|   |   |   mmt_count_07 <= 106: H (11.0)
|   |   |   mmt_count_07 > 106: E (1.0)
|   |   age = 21: H (11.0)
|   |   age = 22: H (18.0)
|   |   age = 23: H (26.0)
|   |   age = 24: H (18.0)
|   |   age = 25
|   |   |   callcenter_count_07 <= 0
|   |   |   |   msmo_count_07 <= 204: H (15.0)
|   |   |   |   msmo_count_07 > 204: E (1.0)
|   |   |   callcenter_count_07 > 0
|   |   |   |   tariff_type = 1: E (1.0)
|   |   |   |   tariff_type = 2: H (1.0)
|   |   age = 26
|   |   |   callcenter_count_07 <= 7: H (14.0)
|   |   |   callcenter_count_07 > 7: E (1.0)
|   |   age = 27
|   |   |   callcenter_count_07 <= 5: H (9.0)
|   |   |   callcenter_count_07 > 5: E (1.0)
|   |   age = 28: H (23.0)
|   |   age = 29: H (14.0)
|   |   age = 30
```

Şekil 8.2. Ekrana sonuçların yazdırılması.

Geliştirilen Shiny uygulamasının web adresinde yayınlanarak kullanıcıların web browser da karşılarına çıkan görsel Şekil 8.3'te görülmektedir. Ayrıca yapılan çalışma "https://churn.shinyapps.io/sonn/" web adresinde yayınlanarak kullanıcıların web üzerinden erişimine sunulmuştur. Bu bölüm boyunca yapılan tüm çalışmanın kodları EK-4'te yer almaktadır.



Şekil 8.3. Shiny uygulamasının görüntüsü.

9. TARTIŞMA VE SONUÇ

Bu çalışma kapsamında veri madenciliği, makine öğrenmesi, makine öğrenmesi süreci gibi çeşitli kavramlar ve çalışılan algoritmalar hakkında bilgi verilmiştir. Yapılan literatür taramasında tahmin için en çok çalışılan yöntemin sınıflandırma algoritmaları olduğu görülmüş ve bundan dolayı sınıflandırma algoritmaları ile modeller oluşturularak en iyi model belirlenmiştir. Bu çalışma ile beraber kullanılan programın da katkısıyla veri görselleştirme ve web tabanlı uygulama örnekleri de yapılmıştır.

Yapılan literatür çalışmasında veri seti bakımından kayıt ve değişken sayıları yakın denebilecek olanlar; Kamalraj [4], Yabaş [7], Hudaib [12], Branduşoiu [19], AlOmari [22], Gürsoy [23]'a ait çalışmalardır Yapılan tez çalışmasında elde edilen sonuçlar, literatürdeki çalışmaların sonuçları ile farklılıklar göstermektedir. Bunun sebebi her çalışmanın kendine özgü veri seti üzerinde yapılmış olmasıdır. Fakat telekomünikasyon sektörü üzerine çalışılmış olması ve genel olarak sınıflandırma algoritmaları ile çalışılmasından dolayı benzerlik kurulabilir. Bizim çalışmamızda karar ağaçları birbirine yakın sonuçlar vermiştir. Karar ağaçları ile en yakın komşu algoritması ile oluşturulan modeller benzer sonuçlar verirken; Bayes algoritması ile oluşturulan modellerin daha düşük sonuç verdiği görülmektedir. Bunun sebebi kategorik verilerimizin çok olmasıdır.

Bölüm 6.3.1'de yer alan model performans karşılaştırması başlığı altında 4-kat çapraz geçerleme, 5-kat çapraz geçerleme ve 10 kat çapraz geçerleme ile elde edilen sonuçlar ve hold-out ile elde edilen sonuçlar anlatılmaktadır. Sınıflandırma algoritmaları ile kurulan modellerin tablo olarak karşılaştırılması Bölüm 6.3.1'de model performans karşılaştırılma başlığı altında gösterilmiştir. Bu bölümde görüldüğü gibi tüm tabloların sonuçları birbirleri ile benzer sonuçlar vermiştir. Bu bölümde C4.5 karar ağacı ile oluşturulan model, performans değerlendirmelerinin hepsinde yaklaşık 0.98 değer ile doğruluk bakımından en iyi sonucu veren algoritma olmuştur. 4 kat, 5 kat ve 10 kat çapraz geçerleme ile performans değerlendirme de C4.5 algoritması ile kurulan modeli diğer karar ağaçlarından ID3 ve Gini takip etmiştir. Sırasıyla *k*-En yakın komşu ve Bayes de karar ağaçlarından sonra dördüncü ve beşinci iyi performansı gösteren algoritma olmuştur.

Hold-out performans değerlendirme ve model seçim yönteminde %60-%40, %75-%25, %80-%20 eğitim-test veri seti ayrımlarının hepsinde C4.5 karar ağacı ile oluşturulan model doğruluk, hata, tanısal üstünlük oranı ve F-ölçü değerlerine göre en iyi performansı gösteren algoritma olmuştur. C4.5 algoritması ile oluşturulan modeli tüm ayrımlarda *k*-en yakın komşu algoritması takip ederken sonrasındaki sıralamada sırasıyla ID3, Gini karar ağaçları ve Bayes algoritması gelmektedir. *k*-en yakın komşu algoritmasının ID3 ve Gini karar ağacının geçmesi hold-out ile rastgele ayrımda ID3 ve Gini karar ağaçlarına göre daha iyi performans göstermesinden dolayıdır.

Karar ağaçları birbiriyle yakın sonuç vermesi sürekli değer sayısının çok olmasından dolayı olabilir. Bu dezavantajı gidermek için veri dönüşümü yapılmıştır. Bu çalışma sonucunda C4.5 karar ağacı algoritması diğer modellere göre daha iyi bir performans göstermiş olmasına rağmen ileriki çalışmalarda karar ağaçlarının dezavantajları giderilerek yeni bir hibrit model elde edilerek yapılması planlanmaktadır.

Bölüm 5'te yer alan veri madenciliği yoluyla veri görselleştirme örnekleri R paketlerinin sağlamış olduğu grafikler ile yorumlanarak gösterilmiştir. Bu çalışmada amaç onlarca rakamın yer aldığı tablolarla uğraşmamak ve veri madenciliği tekniklerine girmeden grafikler ile elimizde bulunan veri seti hakkında bilgi sahibi olabilmeyi hatta çıkarım yapmayı sağlayan veri görselleştirme yapmaktır. Veri madenciliği yoluyla veri görselleştirme alt başlığında anlatılan bu bölümde asıl vurgulamak istediğimiz grafikleri yorumlayarak çıkarım yapmada üst düzey matematik yetenekleri gerektirmediğidir. Ayrıca grafiklerin çıkarıldığı sektör ve veri seti hakkında ilgili bilgisi olan herkesin elindeki veri setine ilişkin buna benzer grafikleri çıkarma ve elde ettiği sonuçlarla ilgili yorum yapabilme imkanına sahip olduğudur. Böylece herhangi bir veri setiyle ilgili uygun grafikleri üretmek ve yorumlamak oldukça kolay hal almış oluyor. Yapılan çalışmada density, violin grafikleri ile benzer senaryolarda aynı yorumsal sonuçların yapılabildiği görülmektedir. Bu da yapılan yorumun doğruluğunu artırmaktadır.

Bölüm 6'da R paketlerinden olan Shiny paketi ile sınıflandırma algoritmaları ile yapılan çalışmayı görsel ve web tabanlı hale taşımaya imkan sağlayan bir arayüz kazandırılması anlatılmıştır. Üstelik lokalde yapılan çalışma web server aracılığıyla internet ortamına da taşınabilmektedir. Bu çalışmada kullanıcıdan gelen veriler test verisi olarak alınıp eldeki veri seti eğitim verisi olarak kullanılarak C4.5 sınıflandırma algoritması aracılığı ile müşteri kaybı tahmin edilmeye çalışılmıştır. Shiny'de kullanılan model, Bölüm 6.3.1'de yer alan model performans karşılaştırma tablolarında en iyi performansı veren

C4.5 karar ağacı algoritması ile elde edilen modeldir. Ayrıca Shiny ile Telekomünikasyon veri seti üzerine yapılan bu çalışmaya benzer bir çalışma literatürde rastlanmamıştır.

Sonuç olarak, Veri madenciliği ve makine öğrenmesi ile telekomünikasyon sektörüyle alakalı özgün bir tez çalışması yapılarak ilgili telekomünikasyon veri seti üzerinde tahmine yönelik öğrenme modelleri oluşturulmuş ve en iyi performansı veren model belirlenmiştir. Bu model Shiny aracılığı ile dinamik hale getirilmiştir. Ayrıca çeşitli grafikler aracılığı ile veri seti hakkında yorum yapılmıştır.

Bu çalışmaların, makine öğrenmesi ve veri madenciliği algoritmaları ile çalışma yapacaklara yol göstermesi en büyük isteğimizdir.

10.KAYNAKLAR

[1] D. Yağan. (2015, 18 Ağustos). *Hanehalkı bilişim teknolojileri kullanım araştırması* [Online]. Erişim: http://www.tuik.gov.tr/PreHaberBultenleri. do?id=21779.

[2] S. Akyokuş, "Veri madenciliği yöntemlerine genel bakış", Türkiye Bilişim Derneği Veri Madenciliği Günü'nde sunuldu, İstanbul, Türkiye, 2006.

[3] W. Verbeke, K. Dejaeger, D. Martens, J. Hur. B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," *European Journal Of Operational Research,* vol. 218, no. 1, pp. 211-229, 2012.

[4] N. Kamalraj, A. Malathi. "Applying data mining techniques in telecom churn prediction," *International Journal of Advanced Research in Computer Science And Software Engineering,* vol. 3, no. 10, pp. 363-370, 2013.

[5] I. Brandusoiu, G. Toderean, "Churn prediction in the telecommunications sector using support vector machines," *Annals of the Oradea University, Fascicle of Management and Technological Engineering*, vol. 22, no. 1, 2013.

[6] G.D.O. Olle, S.Q. Cai, "A hybrid churn prediction model in Mobile telecommunication industry," *International Journal of e-Education, e-Business, e-Management and e-Learning*, vol. 4, no. 1, 2014.

[7] U.Yabaş, "Customer churn prediction for telecommunications industry," M.S. thesis, Computer Engineering and Computer Science and Control, University of Economics, Izmir, Turkey, 2014.

[8] N. Forhad, S. Hussain, and R. M. Rahman, "Churn analysis: predicting churners," presented at Digital Information Management (ICDIM), Nineth International conference on IEEE, Thailand, 2014.

[9] A. Amin, C. Khan, I. Ali, and A, Anwar, "Customer churn prediction in telecommunication industry: with and without counter-example," Mexican *International Conference On Artificial Intelligence*, Tuxtla Gutiérrez, Mexico, 2014, pp. 206-218.

[10] M. Kuyzu, E. Tufan, "Telekomünikasyon sektöründe müşterilerin ürün grupları ve tarifeler arası geçiş analizi," *İktisat, İşletme ve Finans,* c. 29, s. 345, ss. 41-82, 2014.

[11] M. Kaur, P. Mahajan, P, "Churn prediction in telecom industry using R," *International Journal of Engineering and Technical Research (IJETR),* vol. 3, no. 5, pp. 46-53, 2015.

[12] A. Hudaib, R. Dannoun, O. Harfoushi, R. Obiedat, and H. Faris, "Hybrid data mining models for predicting customer churn," *International Journal of Communications,* vol. 4, no. 1, 2014.

[13] M. Yıldız, S. Albayrak, "Customer churn prediction in telecommunication," *Signal Processing and Communications Applications Conference (SIU)*, Malatya, Turkey, 2015, pp. 252-255.

[14] A. Backiel, Y. Verbinnen, B. Baesens, and G. Claeskens, "Combining local and social network classifiers to improve churn prediction," *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ACM, Paris, France, 2015, pp. 651-658.

[15] K. Dahiya, S. Bhatia, "Customer churn analysis in telecom industry," *Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions), 2015 4th International Conference on IEEE*, Noida, India, 2015, pp. 1-6.

[16] P.K. Dalvi, S.K. Khandge, and A. Deomore, "Analysis of customer churn prediction in telecom industry using decision trees and logistic regression," *Colossal Data Analysis and Networking (CDAN), Symposium on IEEE*, Madhya Pradesh, India, 2016, pp. 1-4.

[17] N. Gordini, V. Veglio, "Customers churn prediction and marketing retention strategies. An application of support vector machines based on the auc parameter-selection technique in b2b e-commerce industry," *Industrial Marketing Management*, vol. 62, pp. 100-107, 2017.

[18] Q. Yihui, Z. Chiyu, "Research of indicator system in customer churn prediction for telecom industry," *Computer Science & Education (ICCSE), 2016 11th International Conference on IEEE*, Kyoto, Japan, 2016, pp. 123-130.

[19] I. Branduşoiu, G. Toderean, and H. Beleiu, "Methods for churn prediction in the pre-paid mobile telecommunications industry," *Communications (COMM), 2016 International Conference on IEEE*, Bucharest, Romania, 2016, pp. 97-100.

[20] M. Oskarsdottir, C. Bravo, W. Verbeke, C. Sarraute, B. Baesens, J. Vanthienen, "A comparative study of social network classifiers for predicting churn in the telecommunication industry," *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on IEEE*, San Francisco, USA, 2016, pp. 1151-1158.

[21] R. Yu, X. An, B. Jin, J. Shi, O. Move, A. Y. Liu, "Particle classification optimization-based BP network for telecommunication customer churn prediction," *Neural Computing and Applications*, vol. 28, no. 150, pp. 1-14, 2016.

[22] D. AlOmari, M.M. Hassan,"Predicting telecommunication customer churn," *9th International Conference on Internet and Distributed Computing Systems*, New Jersey, USA, 2016, pp. 167-178.

[23] U.T.Ş. Gürsoy, "Customer churn analysis in telecommunication sector," *Istanbul University Journal of the School of Business*, vol. 39, no. 1, pp. 35-49, 2010.

[24] C. Shearer, "The CRISP-DM model: the new blueprint for data mining," *Journal of data warehousing*, vol. 5. no. 4, pp. 13-22, 2000.

[25] S. E. Seker, "Müşteri kayıp analizi (customer churn analysis)*," YBS Ansiklopedi*, c. 3, s. 1, 2016.

[26] P. Kotler, *Marketing insights from A to Z: 80 concepts every manager needs to*

know, 1st ed., New Jersey, USA: John Wiley & Sons, 2003.

[27] F. Rechinhheld, W. Sasser, "Zero defections: Quality comes to service," *Harvard Business review*, vol. 68, no. 5, pp. 105-111, 1990.

[28] Y. Özkan, *Veri madenciliği yöntemleri*, 2. basım, İstanbul, Türkiye: Papatya Yayıncılık, 2013.

[29] J. Han, M. Kamber, *Data mining: concepts and techniques*, 1st ed., San Francisco, USA: Morgan Kaufmann Publisher, 2000.

[30] L.S. Gordon M.J.A. Berry, *Mastering data mining: For marketing, sales, and customer relationship management,* 2nd ed., New York, USA: Willey Puslishing, 2004.

[31] L.B. Ayre, *Data mining for ınformation professionals*, San Diego, California: USA, 2006.

[32] E. Alpaydın, "Ham veriden altın bilgiye ulaşma yöntemleri," Bilişim 2000 Eğitim Semineri'nde sunuldu, İstanbul, Türkiye, 2000.

[33] P. Yıldırım, M. Uludağ ve A. Görür, "Hastane bilgi sistemlerinde veri madenciliği," Çanakkale On Sekiz Mart Üniversitesi Akademik Bilişim'08 Sempozyumu'nda sunuldu, Çanakkale, Türkiye, 2008.

[34] B. Işıklı. (2009, 16 Şubat). *Veri madenciliği nerelerde kullanılır-3* [Online]. Erişim: https://burakisikli.wordpress.com/tag/veri-madenciligi.

[35] Ç. Nergiz, *İş zekası ve veri ambarı*, 1. basım, Ankara, Türkiye: ODTÜ Yayıncılık, 2010.

[36] H. Tatlıdil, *Uygulamalı çok değişkenli istatistiksel analiz*, 1. basım Ankara, Türkiye: Ziraat Matbaacılık, 2002.

[37] G.M. Jacquez, R.Grimson, and L.A. Waller, "The analysis of disease clusters, part II: introduction to techniques," *Infection Control & Hospital Epidemiology*, vol. 17, no. 6, pp. 385-397, 1996.

[38] M. Dener, M. dörterler, and A. Orman, "Açık kaynak kodlu veri madenciliği programları: Weka'da örnek uygulama," Akademik Bilişim'09 Konferansı'nda sunuldu, Şanlıurfa, Türkiye, 2009.

[39] KNIME, *Bilgisayar Programı*, Konstanz Üniversitesi, Zürih Teknopark, 2004.

[40] T.T. Bilgin, "Veri akışı diyagramları tabanlı veri madenciliği araçları ve yazılım geliştirme ortamları," Akademik Bilişim'09 Konferansı'nda sunuldu, Şanlıurfa, Türkiye, 2009.

[41] YALE, *Bilgisayar Programı*, Yale Üniversitesi, 2001.

[42] RAPIDMINER, *Bilgisayar Programı*, Dormunt Teknoloji Üniversitesi Yapay Zeka Birimi, 2006.

[43] B. Sanlı. (2014, 24 Ağustos). *R ile Enerji Verilerinin Analizi ve Modellemesi* [Online]. Erişim: http://www.barissanli.com/calismalar/dersler/r/giris.php.

[44] SPSS, *Bilgisayar Programı*, SPSS Inc, 1968.

[45] A. Demirci. (2015, 28 Eylül). *Data Driven Kavramı* [Online]. Erişim: http://devveri.com/kategori/haberler.

[46] M. Kaya, S.A. Özel, "Açık kaynak kodlu veri madenciliği yazılımlarının

karşılaştırılması," Akademik Bilişim'14 Konferansı'nda sunuldu, Mersin, Türkiye, 2014.

[47] G. Piatetsky. (2016, 6 Haziran). *R, Python Duel As Top Analytics, Data Science Software-KDnuggests 2016* [Online]. Erişim: http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html.

[48] S. Buuren, K. Groothuis-Oudshoorn, A. Robitzsch, G. Vink, L. Doove, S. Jolani, R. Schouten, P. Gaffert, and F. Meinfelder. (2015, 5 Şubat). *Package mice* [Online]. Erişim: http://cran. r-project. org/packages=mice.

[49] T.M. Mitchell, *Machine Learning*, 1st ed., New York, USA: McGraw-Hill, 1997.

[50] J. Brownlee. (2014, 12 Şubat). *Solve Machine Learning Problems Step-by-Step* [Online]. Erişim: https://machinelearningmastery.com/applied-machine-learning-process.

[51] J. Hu. (2013, 12 Mart). *About Data Mining: Basic steps of applying machine learning methods* [Online]. Erişim: http://www.aboutdm.com/2013/03/basic-steps-of-applying-machine.html.

[52] J.J. Rossi, *MicroRNA Methods*, 1st ed., Amsterdam, Netherlands: Elsevier textbooks, 2007.

[53] F.Ö.K. Bakioğlu, E. Kartal, Z. Özen, Ç. Erol, S. Gülseçen, "Aspects of students about ınformation technology courses in social science," *Procedia-Social and Behavioral Sciences*, vol. 176, pp. 148-154, 2015.

[54] S. Ahsan, S. Abad. "Data, information, knowledge, wisdom: A doubly linked chain," *the proceedings of the 2006 International Conference on Information Knowledge Engineering*, Nevada, USA, 2006.

[55] J. Carpenter, J. Bartlett, and M. Kenward. (2015, 2 Şubat). *Introduction to missing data* [Online]. Erişim: http://missingdata.lshtm.ac.uk/index.php?option=com_content&view=section&id=7&Itemid=96.

[56] M.M. Rahman, D.N. Davis, "Machine learning-based missing value imputation method for clinical datasets," *IAENG Transactions on Engineering Technologies*. Springer Netherlands, vol. 229, pp. 245-257, 2013.

[57] R.S. Somasundaram, R Nedunchezhian, "Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values," *International Journal of Computer Applications*, vol. 21, no. 10, 2011.

[58] E.W. Steyerberg, *Clinical prediction models: a practical approach to development, validation, and updating*, 1st. ed., Berlin, Germany: Springer Science & Business Media, 2008.

[59] A. Gelman, J. Hill, *Data analysis using regression and multilevelhierarchical models*, 1st ed., Cambridge, England: Cambridge University Press, 2007.

[60] G.B. Durrant. (2005, 10 Şubat). *Imputation methods for handling item-nonresponse in the social sciences: a methodological review* [Online]. Erişim: http://missingdata.lshtm.ac.uk/preprints/durrantOct05.pdf.

[61] G.E. Batista, M.C. Monard, "An analysis of four missing data treatment methods for supervised learning," *Applied artificial intelligence*, vol. 17, no. 5-6, pp. 519-533, 2003.

[62] W. Chang, J. Cheng, J.J. Allaire, Y. Xie, and J. McPherson. (2005, 11 Kasım). *Shiny: Web Application Framework for R* [Online]. Erişim: http://CRAN.R-project.org/package=shiny.

[63] F.E. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, vol. 11, no. 1, pp. 1-21, 1969.

[64] J.F. Hair, R.E. Anderson, R.L. Tatham, and C. William, *Multivariate data analysis*, 7th ed., New Jersey, USA: Prentice Hall, 1998.

[65] P. Flach, Machine Learning: The art and science of algorithms that make sense of data, *Lecture Notes,* University of Bristol, England, 2012.

[66] P. Rai, Model selection and feature, *Ders Notları*, School of Computing of The University of Utah, USA, 2011.

[67] H.Arslan, "Sakarya üniversitesi web sitesi erişim kayıtlarının web madenciliği ile analizi," Yüksek lisans tezi, Elektronik-Bilgisayar Eğitimi, Sakarya Üniversitesi, Sakarya, Türkiye, 2008.

[68] S. Akbulut, "Veri madenciliği teknikleri ile bir kozmetik markanın ayrılan müşteri analizi ve müşteri segmentasyonu," Yüksek lisans tezi, Endüstri Mühendisliği, Gazi Üniversitesi, Ankara, Türkiye, 2006.

[69] P. Harrington, *Machine learning in action,* New York, USA: Manning, 2012.

[70] Anonim, (15 Nisan 2016). [Online]. Erişim: http://www.matlabyar.com/wp-content/uploads/edd/2016/03/knnng.png.

[71] A. Özdemir, U. Yavuz, Uz. ve Ayık, "Lise türü ve lise mezuniyet başarısının, kazanılan fakülte ile ilişkisinin veri madenciliği tekniği ile analizi," *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi,* c. 10, s. 2, ss. 441-454, 2007.

[72] E. Kartal, M.E. Balaban, *Sınıflandırmaya dayalı makine öğrenmesi teknikleri ve kardiyolojik risk değerlendirmesine ilişkin bir uygulama*, 1. baskı, İstanbul, Türkiye: Çağlayan Kitapevi, 2016.

[73] L. Jiang, C. Li, "An empirical study on attribute selection measures in decision tree learning," *Journal of Computational Information Systems* vol. 6, no. 1, pp. 105-112, 2010.

[74] R.L. De Mantaras, J. Cerquides, and P. Garcia, "Comparing information-theoretic attribute selection measures: a statistical approach," *AI Communications*, vol. 11, no. 2, pp. 91-106, 1998.

[75] J.R. Quinlan, "Improved use of continuous attributes in C4.5," *Journal of artificial intelligence research*, vol. 4, pp. 77-90, 1996.

[76] R. Kohavi, J.R. Quinlan, "Data mining tasks and methods: Classification: decision-tree discovery," *Handbook of data mining and knowledge discovery*, 1st ed., New York, USA: Oxford Universtiy Press, 2002.

[77] A. Demiriz, Karar ağaçları ile sınıflandırma, *Ders Notları*, Sakarya Üniversitesi, 2016.

[78] P. Refaeilzadeh, L. Tang, L. and H. Liu, "Cross-validation," *Encyclopedia of*

database systems, 1st ed., Tempe, USA: Springer US, 2009, pp. 532-538.

[79] R. Gutierrez, Lecture 13: Validation (Intelligent Sensor Systems), Lecture *Notes*, Computer Science & Engineering of Texas A&M Univresity, 2013.

[80] G.K.D. Saharidis, I.P. Androulakis, and M.G. Ierapetritou, "Model building using bi-level optimization," *Journal of Global Optimization*, vol. 49, no.1, pp. 49-67, 2011.

[81] R. Remesan, J. Mathew, *Hydrological data driven modelling*, 1st ed., Basel, Switzerland: Springer International Publishing, 2016.

[82] F.R. Scott. (2012, 10 Temmuz). *Understanding the Bias-Variance Tradeoff* [Online]. Erişim: http://scott.fortmann-roe.com/docs/BiasVariance.html.

[83] Z. Martinasek, J. Hajny, and L. Malina, "Optimization of power analysis using neural network," presented at International Conference on Smart Card Research and Advanced Applications, Berlin, Germany, 2013.

[84] S. Rajeev, *Research Developments in Computer Vision and Image Processing: Methodologies and Applications: Methodologies and Applications*, 1st ed., Pennsylvania, USA: IGI Global, 2013.

[85] F. Guil-Reyes, M.T. Daza-Gonzalez, "Summarizing frequent itemsets via pignistic transformation," *Portuguese Conference on Artificial Intelligence*, Lisbon, Portugal, 2011, pp. 297-310.

[86] D.L. Olson, D. Delen, *Advanced data mining techniques*, 1st. ed., Berlin, Germany: Springer-Verlag Berlin Heidelberg, 2008.

[87] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *Ijcai*, vol. 14, no. 2, 1995.

[88] S. Rogers, M. Girolami, *A First Course in Machine Learning*, 2nd ed., Florida, USA: CRC Press, 2016.

[89] N. Japkowicz, "Performance evaluation for learning algorithms," presented at International conference on machine learning, Edinburg, Scotland, 2012.

[90] M. Clark, An Introduction to machine learning with Applications in R, *Lecture Notes*, University of Notre Dame, 2015.

[91] H. Nizam, S.S. Akın, "Sosyal medyada makine öğrenmesi ile duygu analizinde dengeli ve dengesiz veri setlerinin performanslarının karşılaştırılması," XIX. Türkiye'de İnternet Konferansı'nda sunuldu, İstanbul, Türkiye, 2014.

[92] A.D. Ertorsun, B. Bağ, G. Uzar, and M.A. Turanoğlu, ROC (Receiver Operating Characteristic) Eğrisi Yöntemi ile Tanı Testlerinin Performanslarının Değerlendirilmesi, XIII. Öğrenci Sempozyumu'nda sunuldu, Ankara, Türkiye, 2009.

[93] L. Hamel, *Encyclopedia of Data Warehousing and Mining*, 2nd ed., IGI Global, 2009, pp.1316-1323.

[94] G. Memiş, *Yarı otomatik ders program sistemi*, Yüksek lisans, tezi, Bilgisayar Mühendisliği, Başkent Üniversitesi, Ankara, Türkiye, 2008.

[95] R. Parikh, S. Parikh, E. Arun, and R. Thomas, "Likelihood ratios: clinical application in day-to-day practice," *Indian Journal of Ophthalmology*, vol. 57, no. 3, pp. 217, 2009.

[96] A.S. Glas, J.G. Lijmer, M.H. Prins, G.J. Bonsel, and P.M.M. Bossuyt, "The diagnostic odds ratio: a single indicator of test performance," *Journal of clinical epidemiology*, vol. 56, no. 11, pp. 1129-1135, 2003.

[97] P. Flach, The many faces of ROC analysis in machine learning, *Lecture Notes*, University of Bristol, 2004.

[98] M.E. Balaban, E. Kartal, Veri madenciliği ve makine öğrenmesi temel algoritmaları ve R dili ile uygulamaları, 1. basım, İstanbul, Türkiye: Çağlayan Kitabevi, 2015.

[99] M. Sokolova, G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45 no. 4, pp.427-437, 2009.

[100] K. Merih. (2016, 27 Haziran), *Principal Components Analizi (PCA) ve R veri Görselleştirme* [Online]. Erişim: https://tr.linkedin.com/pulse/principal-components-analizi-pca-ve-r-veri-ile-uygylamasi-datalab-tr.

[101] r-project. (2016, 10 Şubat). *The R Project for Statistical Computing* [Online]. Erişim: http://www.r-project.org/.

[102] RStudio 2015a. (2016, 10 Şubat). *Home – RStudio* [Online]. Erişim: http://www.rstudio.com/.

[103] RStudio 2015b. (2016, 10 Şubat). *Shiny* [Online]. Erişim: http://shiny.rstudio.com/.

[104] RStudio 2015c. (2016, 10 Şubat). *Shinyapps.io* [Online]. Erişim: https://www.shinyapps.io/.

[105] C. Müssel, L. Lausser, M. Maucher, and H.A. Kestler, "Multi-objective parameter selection for classifiers," *Journal of Statistical Software*, vol. 46, no. 5, 2012.

[106] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. (2014, 14 Ocak). *E1071: Misc Functions of the Department of Statistics (e1071)* [Online]. Erişim: http://CRAN.R-project.org/package=e1071.

[107] B.Wei, F. Yang, X. Wang, and Y. Ge. (2012, 12 Temmuz). *KnnGarden: Multi-distance based k-Nearest Neighbors* [Online]. Erişim: http://CRAN.R-project.org/package=knnGarden.

[108] K. Hornik, C. Buchta, C. and A. Zeileis, "Open-source machine learning: R meets Weka," *Computational Statistics*, vol. 24, no. 2, pp. 225-232, 2009.

[109] I.H. Witten, E. Frank, *Data Mining: Practical machine learning tools and techniques*, 4th ed., Cambridge, Massachusetts, USA: Morgan Kaufmann, 2016.

[110] T. Hothorn, T. ve A. Zeileis. (2014, 10 Mart). *Partykit: A Modular Toolkit for Recursive Partytioning in R* [Online]. Erişim: http://EconPapers.RePEc.org/RePEc:inn:wpaper:2014-10.

[111] W. Hadley, C. Winston. (2016, 30 Aralık). *KMggplot2* [Online]. Erişim: https://cran.r-project.org/ packages/RcmdrPlugin.KMggplot2.

[112] M. Kuhn, J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca,

Y. Tang, C. Candan, and T. Hunt. (2015, 23 Kasım). *Package caret* [Online]. Erişim: https://cran.r-project.org/web/packages/caret.

[113] T. Therneau, B. Atkinson, B. Ripley. (2015, 24 Şubat). *Package rpart* [Online]. Erişim: https://cran.r-project.org/web/packages/rpart/rpart.

[114] S.Milborrow. (2015, 24 Şubat). *Package rpart.plot* [Online]. Erişim: https://cran.r-project.org/web/packages/rpart.plot/rpart.plot.

11.EKLER

11.1. EK 1: R TEMEL İŞLEMLER

Veri Okuma ve Diğer Temel İşlemler

```
#Veri Okuma
dataset <- read.csv("C:/Users/sinan/Desktop/deneme.csv")#veri okuma
View(dataset)#veriyi tablosal olarak Shiny de görme
##Veriyi Yazdırma
write.table(data, file = "dataset.csv", row.names=FALSE, na="", sep=",")
###sütun ve satır okuma adları okuma
colnames(dataset) #sütun okuma
rownames(dataset)#satır adı okuma
#Sütun Adı Değiştirme
colnames(dataset)[12]<-"mmo_count"
##tarihleri gün,ay, yıl olarak ayırma
library(lubridate)
dataset3$last_reload_date <- ymd(as.character(dataset3$last_reload_date))
tarihh <- as.Date(as.character(dataset3$last_reload_date), format='%Y%m%d')
sene <- format(tarihh, '%Y')
ay <- format(tarihh, '%m')
gun <- format(tarihh, '%d')
#Kütüphane İşlemleri
require(plyr) #kütüphane gereklilikleri sorgulama
install.packages("plyr")#kütüphane yükleme
library(plyr)#kütüphane çağırma
```

Eksik Veriler – Aykırı Veriler

```
Mice Eksik Veri Tamamlama
library(mice)
#library(VIM)
eks_plot2 <- aggr(a, col=c(white,black),
numbers=TRUE, sortVars=TRUE, labels=names(a),
cex.axis=.10, gap=2, ylab=c("eksik veri histogramı","Pattern"))
#############Nümerik Değişkenler
#Eksik olan yerler bu şekilde mice ile doldurulmuştur.
dataset14 <- mice(dataset[14:20],m=5, maxit=50, meth='pmm',seed=700)
dataset21 <- mice(dataset[21:24],m=5, maxit=50, meth='pmm',seed=700)
mmocount3 <- complete(dataset14,1)
mmocount2 <- complete(dataset21,1)
## Eksik veri tamamlama-2 Klasik (nümeriklere ortalam, kategoriklere en çok tekrar eden) tamamlama
############## Nominal(Kategorik) Değişkenler
dataset[which(is.na(dataset[,86:90]))]<-(names(which.max(table(dataset[,86:90]))))
############## Nümerik Değişkenler
dataset[is.na(dataset[,62:67])]<-(mean(table(dataset[,62:67])))
# ######Eksik Veri Tamamlama-3 EM(Expectation-Maximization) Algoritması İle
## EM(Expectation-Maximization) -1
library(e1071)
data <- dataset
var(na.omit(data) )
var(data.imputed)
EM <- function(x, tol=.001){
tam <- is.na(x)
```

```
new.impute<-x
old.impute <- x
count.iter <- 1
reach.tol <- 0
iz <- as.matrix(var(na.exclude(x)))
mean.vec <- as.matrix(apply(na.exclude(x),2,mean))
 while(reach.tol != 1) {
for(i in 1:nrow(x)) {
pick.miss <-( c( tam[i,]) )
if ( sum(pick.miss) != 0 ) {
inv.S <- solve(iz[!pick.miss,!pick.miss]) # Kovaryansın tersi alınır
# Em çalışmaya başlar
new.impute[i,pick.miss] <- mean.vec[pick.miss] +
iz[pick.miss,!pick.miss]
inv.S
(t(new.impute[i,!pick.miss])- t(t(mean.vec[!pick.miss]))) }}
iz <- var((new.impute))
mean.vec <- as.matrix(apply(new.impute,2,mean))
if(count.iter > 1){ #  ilk yinelemede çalışmasını istenilmiyor
for(l in 1:nrow(new.impute)){
for(m in 1:ncol(new.impute)){
if( abs((old.impute[l,m]-new.impute[l,m])) > tol ) {
reach.tol < - 0} else {
reach.tol <- 1}}}}
count.iter <- count.iter+1 # Hata ayıklama amacıyla düzgün yineleme işlemini sağlamak için kullanılır
old.impute <- new.impute} return(new.impute)}
data.imputed <- EM(data, tol=.0001)
plot(data.imputed[,1], data.imputed[,2], pch=16, main="Eksik değer dağılımı",
sub="Kayıp değer", xlab="X",ylab="Y")
# Değer eşleştirme
plot.imputed <- data.imputed[
row.names(
subset(data, is.na( data[,2] ) | is.na( data[,3]) )  ),]
points(plot.imputed[,2],plot.imputed[,3], pch=16, col='red')
## EM(Expectation-Maximization) -3 Amelia paketi ile
library(Amelia)#
data(dataset)
head(dataset,12)
data.out <- amelia(dataset, m =5, boot.type = "none")
data.out <- amelia(dataset, m =5, idvars=2)
head(data.out$imputations[[1]],12)
data.out <- amelia(dataset, m =5, idvars=2,ords="non_percent",noms="mmo_duration")# Eksik Veri
########## Eksik Veri Tamamlama-4 NA Değerlerin Atılması İle
datasetNAextract<-na.omit(dataset)
```

Outlier Tespit Ve Düzenleme

```
#Outlier Tespiti
outlier(dataset$age)#Üst değer
outlier(dataset$age,opposite = TRUE)#alt değer
boxplot(dataset$mmo_count_07)
###############################################
#Outlier Düzeltme
###############################################
dt <- dataset$age
ceyrek <- quantile(t, probs=c(.25, .75), na.rm = T)
kes <- quantile(dt, probs=c(.05, .95), na.rm = T)
sin <- 1.5 * IQR(dt, na.rm = T)
dt[dt < (ceyrek[1] - sin)] <- kes[1]
dt[dt > (ceyrek[2] + sin)] <- kes[2]
summary(dt)
```

```
###################
length(data1)#42 değer
summary(data1)
bench<-15.50+1.5*IQR(data1)#Q3+1.5IQR
bench
data1[data1<35]
data1[data1>35]
data1<-data1[data1<bench]
summary(data1)
boxplot(data1)
length(data1)
```

Veri Dönüştürme

```
#Sürekli Olan Yaş Değerlerini Kesikli Hale Getirme##############
agecat<-cut(dataset$age,breaks = c(15, 25, 45, 65, 75))
agelabs <- c( "16-24", "25-44",  "45-64", "65+")
dataset$age<-cut(dataset$age,breaks = c(15, 25, 45, 65, 75), right=FALSE, labels=agelabs)
#Sürekli Olan Aylk Konuşma Sürelerini Kesikli Hale Getirme##############
mmodur07<-cut(dataset $mmo_duration_07,breaks = c(0, 120, 240, 360, 480,600,720,840,960,1080))
#mmodur7ad <- c( "0-2 saat", "2-4 saat",  "4-6 saat", "6-8 saat",  "8-12 saat", "12-14 saat" "14-16
saat", "16-18 saat","18-20 saat","20-22 saat","22-24 saat")
dataset$mmo_duration_07<-cut(dataset $mmo_duration_07,breaks = c(0, 120, 240, 360,
480,600,720,840,960,1080), right=FALSE, labels=mmodur7ad)
```

11.2. EK 2: SINIFLANDIRMA MODELLERİ KURMA

Hold-out Ayrım ile Oluşturulan Modeller

```
#################################################################
######C4.5 KARAR AĞACI ALGORİTMASI ile YAPILAN MODEL
#################################################################
#Kütüphane yükleme
library(rpart)
library(RWeka)
library(partykit)
library(caret)
library(ROCR)
library(rpart)
library(e1071)
#########################################################################
#Rastgele Eğitim ve Test Ayrımı
set.seed(2016)
randomayir60 <- createDataPartition(y = dataset$churn_2013_07, p = .60, list = FALSE) #%60 ayrım
#%75,%80 ayrımda bu sayı değiştirilmiştir.
trainSET60 <- dataset[randomayir60,]
testSET60 <- dataset[-randomayir60,]
#Eğitim kümesi ile modelin oluşturulması
DT_model60 <-J48(churn_2013_07~., data=trainSET60)
print(DT_model60)
summary(DT_model60)
plot(DT_model60)
#str(dataset)
#summary(dataset)
#dim(dataset)
#Test kümesi üzerinde tahminleme işlemlerinin yapılması
pred_testSET60<- predict(DT_model60, newdata = testSET60[,-113])
#Test kümesi için confusion matrisinin oluşturulması
DT_table60 <- table(pred_testSET60, testSET60$churn_2013_07, dnn = c("Tahminler", "Gerçekler"))
DT_table60
#performans değerlendirme
(tp<-table60[1])
(fp<-table60[3])
(fn<-table60[2])
(tn<-table60[4])
paste0("Dogruluk = ",(dogruluk <- (tp+tn)/sum(table60)))
paste0("Hata = ",(hata <- 1-dogruluk))
paste0("TPR = ",(TPR <- tp/(tp+fn)))
paste0("SPC = ",(SPC <- tn/(fp+tn)))
paste0("PPV = ",(PPV <- tp/(tp+fp)))
paste0("NPV = ",(NPV <- tn/(tn+fn)))
paste0("FPR = ",(FPR <- fp/(fp+tn)))
paste0("FNR = ",(FNR <- fn/(fn+tp)))
paste0("LR_p = ",(LR_p <- TPR/FPR))
paste0("LR_n = ",(LR_n <- FNR/SPC))
paste0("DOR = ",(DOR <- LR_p/LR_n))
paste0("F_measure = ",(F_measure <- (2*PPV*TPR)/(PPV+TPR)))
#################################################################
####Gini ALGORİTMASI ile YAPILAN MODEL
#################################################################
set.seed(2015)
tahmin<-createDataPartition(y=telekom$churn_2013_07, p=.60, list=FALSE)
train <- telekom[tahmin,]
test <- telekom[-tahmin,]
```

```
tree_model<-rpart(churn_2013_07~.,train,method="class",minsplit=20,parms=list(split="gini"))
show(tree_model)
prp(tree_model)
tree_predict<-predict(tree_model,test,type="class")
TGini<-table(tree_predict, test$churn_2013_07, dnn=c("Tahmini", "Gerçek"))
TGini
(tp <- TGini[1])
(fp <- TGini[3])
(fn <- TGini[2])
(tn <- TGini[4])
paste0("Dogruluk = ",(dogruluk <- (tp+tn)/sum(TGini)))
paste0("Hata = ",(hata <- 1-dogruluk))
paste0("TPR = ",(TPR <- tp/(tp+fn)))
paste0("SPC = ",(SPC <- tn/(fp+tn)))
paste0("PPV = ",(PPV <- tp/(tp+fp)))
paste0("NPV = ",(NPV <- tn/(tn+fn)))
paste0("FPR = ",(FPR <- fp/(fp+tn)))
paste0("FNR = ",(FNR <- fn/(fn+tp)))
paste0("LR_p = ",(LR_p <- TPR/FPR))
paste0("LR_n = ",(LR_n <- FNR/SPC))
paste0("DOR = ",(DOR <- LR_p/LR_n))
paste0("F_measure = ",(F_measure <- (2*PPV*TPR)/(PPV+TPR)))
#################################################################
####ID3 ALGORİTMASI ile YAPILAN MODEL
#################################################################
ID3 <- function(node, dataset) {
node$osSay <- nrow(dataset)
#if the dataset-set is pure (e.g. all toxic), then
if (IsPure(dataset)) {
chd <- node$AddChid(unique(dataset[,ncol(dataset)]))
node$degisken <- tail(names(dataset), 1)
chd$osSay <- nrow(dataset)
child$degisken <- "
} else {
# information gain hesabı;
ggg <- sapply(colnames(dataset)[-ncol(dataset)]
function(x) InformationGain(
table(dataset[,x], dataset[,ncol(dataset)])))
degisken <- names(which.max(ggg))
node$ degisken <- degisken
chdOs <- split(dataset[ ,names(dataset) != degisken, drop = FALSE],
dataset[ , degisken],
drop = TRUE)
for(i in 1:length(chdOs)) {
chd <- node$AddChd(names(chdOs)[i])
TrainID3(chd, chdOs[[i]])}}}
library(data.tree)
data(dataset)
dataset
tree <- Node$new("dataset ")
TrainID3(tree, age)
print(tree, " degisken ", "osSay")
tahmin <- function(tree, degisken) {
if (tree$ chid [[1]]$isLeaf) return (tree$ chid [[1]]$name)
child <- tree$chid[[degisken [[tree$ degisken]]]]
return (tahmin (chid, degisken))}
#################################################################
####Bayes ALGORİTMASI ile YAPILAN MODEL
#################################################################
set.seed(2017)
```

```
egitimIndisleriBayes60<-createDataPartition(y=dataset$churn_2013_07,p=.60,list=FALSE)
egitimBayes60<-dataset[egitimIndisleriBayes60,]
testBayes60<-dataset[-egitimIndisleriBayes60,]
testDeğişkenleriBayes60<-testBayes60[,-113]
testHedefDeğişkenleriBayes60<-testBayes60[[113]]
egitimDeğişkenleriBayes60<-egitimBayes60[,-113]
egitimHedefDeğişkenleriBayes60<-egitimBayes60[[113]]
naiveBayes_modeli60<-
naiveBayes(egitimDeğişkenleriBayes60,egitimDeğişkenleriDeğişkenleriBayes60)
naiveBayes_modeli60
(tahminSiniflarBayes60<-predict(naiveBayes_modeli60,testdeğişkenleriBayes60))
(tablomBayes60<-table(tahminSiniflarBayes60,testHedefDeğişkenleriBayes60,dnn=c("Tahmini
Siniflar","Gerçek Siniflar")))
(tp<-tablomBayes60[1])
(fp<-tablomBayes60[3])
(fn<-tablomBayes60[2])
(tn<-tablomBayes60[4])
paste0("Dogruluk = ",(dogruluk <- (tp+tn)/sum(tablomBayes60)))
paste0("Hata = ",(hata <- 1-dogruluk))
paste0("TPR = ",(TPR <- tp/(tp+fn)))
paste0("SPC = ",(SPC <- tn/(fp+tn)))
paste0("PPV = ",(PPV <- tp/(tp+fp)))
paste0("NPV = ",(NPV <- TN/(tn+fn)))
paste0("FPR = ",(FPR <- fp/(fp+tn)))
paste0("FNR = ",(FNR <- fn/(fn+tp)))
paste0("LR_p = ",(LR_p <- TPR/FPR))
paste0("LR_n = ",(LR_n <- FNR/SPC))
paste0("DOR = ",(DOR <- LR_p/LR_n))
paste0("F_measure = ",(F_measure <- (2*PPV*TPR)/(PPV+TPR)))
#################################################################
###K-EN YAKIN KOMŞU ALGORİTMASI ile YAPILAN MODEL
#################################################################
set.seed(2017)
egitimIndisleri60<-createDataPartition(y=dataset$churn_2013_07,p=.60,list = FALSE)
egitim60 <- dataset[egitimIndisleri60,]
test60 <- dataset[-egitimIndisleri60,]
testDeğişkenleriKomsu60<-test60[-113]
testHedefDeğişkenleriKomsu60<-test60[[113]]
egitimDeğişkenleriKomsu60<-egitim60[-113]
egitimHedefDeğişkenleriKomsu60<-egitim60[[113]]
#k için keyfi değer seçildi.
k_degeri=3
set.seed(2016)
(tahminSiniflar60=knn(egitimDeğişkenleriKomsu60,testDeğişkenleriKomsu60,egitimHedefDeğişkenleri
Komsu60,k=k_degeri))
(tablom60<-table(tahminSiniflar60,testHedefDeğişkenleriKomsu60,dnn=c("Tahmini Siniflar","Gerçek
Siniflar")))
(tp<-tablom60[1])
(fp<-tablom60[3])
(fn<-tablom60[2])
(tn<-tablom60[4])
paste0("Dogruluk = ",(dogruluk <- (tp+tn)/sum(tablom60)))
paste0("Hata = ",(hata <- 1-dogruluk))
paste0("TPR = ",(TPR <- tp/(tp+fn)))
paste0("SPC = ",(SPC <- tn/(fp+tn)))
paste0("PPV = ",(PPV <- tp/(tp+fp)))
paste0("NPV = ",(NPV <- tn/(tn+fn)))
paste0("FPR = ",(FPR <- fp/(fp+tn)))
paste0("FNR = ",(FNR <- fn/(fn+tp)))
paste0("LR_p = ",(LR_p <- TPR/FPR))
```

```
paste0("LR_n = ",(LR_n <- FNR/SPC))
paste0("DOR = ",(DOR <- LR_p/LR_n))
paste0("F_measure = ",(F_measure <- (2*PPV*TPR)/(PPV+TPR)))
```

K-kat Çapraz Geçerleme

```
#################################################################
###K-EN YAKIN KOMŞU ALGORİTMASI ile YAPILAN MODEL
#################################################################
tekrar<-1
katSayisi<-5
set.seed(2017)
kat<-generateCVRuns(dataset$churn_2013_07,ntimes = tekrar,nfold =katSayisi,leaveOneOut =
FALSE,stratified =TRUE)
katdogruluk<-rep(0,times=katSayisi)
katdogruluk
churn_gercek<-NULL
model_churn<-NULL
uzunluklar<-sapply(kat[[tekrar]],length)
names(uzunluklar)<-NULL
for(j in 1: katSayisi){
churn_gercek[[tekrar]]<-append(churn_gercek[[tekrar]],values=list(rep(NA,times=uzunluklar[j])))
model_churn[[tekrar]]<-append(model_churn[[tekrar]],values=list(rep(NA,times=uzunluklar[j])))}
dataset$churn_2013_07<-as.numeric(dataset$churn_2013_07)
k_degeri=3
set.seed(123)
for(i in 1:katSayisi){
testIndisleri<-kat[[tekrar]][[i]]
test<-dataset[testIndisleri,]
hurn_gercek[[tekrar]][[i]]<-test$churn_2013_07
egitim<-dataset[-testIndisleri,]
testDeğişkenleri<-test[-113]
testHedefDeğişkenleri<-test[[113]]
egitimDeğişkenleri<-egitim[-113]
egitimHedefDeğişkenleri<-egitim[[113]]
print("Eğitim verisi Sınıf değerleri ")
print(tab(egitim[,113]))
print("Test seti Sınıf değerleri")
print(tab(test[,113]))
model_churn[[tekrar]][[i]]<-knn(egitimDeğişkenleri, testDeğişkenleri, egitimHedefDeğişkenleri,
k=k_degeri)
print("Gerçek churn değerleri =")
print(churn_gercek[[tekrar]][[i]])
print("Tahimini churn değerleri =")
print(model_churn[[tekrar]][[i]])
#performans değerlendirmesi
(tablom<-table(model_churn[[tekrar]][[i]],model_churn[[tekrar]][[i]],
dnn=c("Tahimini Siniflar","Gerçek Siniflar")))
katdogruluk[i]<-mean(model_churn[[tekrar]][[i]]==churn_gercek[[tekrar]][[i]])}
round(katdogruluk,4)
print(paste0(katSayisi, "-Kat çapraz geçerleme sonucu elde edilen ortalama doğruluk ",
round(mean(katdogruluk),4),"'dir."))
#en iyi 3 çıktı
 i<-3
 testIndisleri<-kat[[tekrar]][[i]]
 test<-dataset[testIndisleri,]
churn_gercek[[tekrar]][[i]]<-test$churn_2013_07
egitim<-dataset[-testIndisleri,]
print("Eğitim seti sınıf değerleri dağılımı")
```

```
print(tab(egitim[,34]))
print("Test veri seti Sinif değerleri dağılımı")
print(tab(test[,34]))
model_churn[[tekrar]][[i]]<-knn(egitimDeğişkenleri, testDeğişkenleri, egitimHedefDeğişkenleri,
k=k_degeri)
model_churn[[tekrar]][[i]]<-knn(egitimDeğişkenleri, testDeğişkenleri, egitimHedefDeğişkenleri,
k=k_degeri)
print("Gerçek churn değerleri =")
print(churn_gercek[[tekrar]][[i]])
print("Tahimini churn değerleri =")
print(model_churn[[tekrar]][[i]])
#################################################################
###ID3 ALGORİTMASI ile YAPILAN MODEL
#################################################################
data(dataset)
rpart.model <- rpart(age~., data= dataset, method="class")
print(rpart.model)
rcart.prediction <- predict(rpart.model, newdata= dataset, type="class")
confusion.matrix <- table(dataset $age, rcart.prediction)
print(confusion.matrix)
accuracy.percent <- 100*sum(diag(confusion.matrix))/sum(confusion.matrix)
print(paste("acc:",accuracy.percent,"%"))
library(plyr)
set.seed(1)
form <- "dataset"
fld <- split(dataset, cut(sample(1:nrow(dataset)),10))
err <- rep(NA, length(fld))
for (i in 1:length(folds)) {
 test <- ldply(fld[i], data.frame)
train <- ldply(fld[-i], data.frame)
dt.model <- rpart(form , train, method = "class")
dt.predict <- predict(tmp.model, newdata = test, type = "class")
conf.mat <- table(test$churn, tmp.predict)
err[i] <- 1-sum(diag(conf.mat))/sum(conf.mat)}
print(("Ortalama hata k-fold cross-validation: %.3 ", 100*mean(err))
#################################################################
###BAYES ALGORİTMASI ile YAPILAN MODEL
#################################################################
tekrar<-1
katSayisi<-5
set.seed(2017)
kat<-generateCVRuns(dataset$churn_2013_07,ntimes = tekrar,
nfold =katSayisi,leaveOneOut = FALSE,stratified =TRUE)
katdogruluk<-rep(0,times=katSayisi)
katdogruluk
churn_gercek<-NULL
model_churn<-NULL
uzunluklar<-sapply(kat[[tekrar]],length)
names(uzunluklar)<-NULL
for(j in 1: katSayisi){
churn_gercek[[tekrar]]<-append(churn_gercek[[tekrar]],values=list(rep(NA,times=uzunluklar[j])))
model_churn[[tekrar]]<-append(model_churn[[tekrar]],values=list(rep(NA,times=uzunluklar[j])))}
for(i in 1:katSayisi){
testIndisleri<-kat[[tekrar]][[i]]
test<-dataset[testIndisleri,]
egitim<-dataset[-testIndisleri,]
testDeğişkenleri<-test[-113]
testHedefDeğişkenleri<-test[[113]]
egitimDeğişkenleri<-egitim[-113]
egitimHedefDeğişkenleri<-egitim[[113]]
```

```
churn_gercek[[tekrar]][[i]]<-test$churn_2013_07
print("Eğitim veri seti Sınıf Değerleri dağılımı")
print(tab(egitim[,34]))
print("Test seti Sınıf Değerleri dağılımı")
print(tab(test[,34]))
naiveBayes_modeli60<-naiveBayes(egitimDeğişkenleri, egitimHedefDeğişkenleri)
model_churn[[tekrar]][[i]]<-predict(naiveBayes_modeli60,newdata=test[-34])
print("Gerçek churn değerleri =")
print(churn_gercek[[tekrar]][[i]])
print("Tahimini churn değerleri =")
print(model_churn[[tekrar]][[i]])
#performans değerlendirmesi
(tablom<-table(model_churn[[tekrar]][[i]],model_churn[[tekrar]][[i]],dnn=c("Tahimini
Siniflar","Gerçek Siniflar")))
katdogruluk[i]<-mean(model_churn[[tekrar]][[i]]==churn_gercek[[tekrar]][[i]])}
round(katdogruluk,4)
print(paste0(katSayisi, "-Kat çapraz geçerleme sonucu elde edilen ortalama doğruluk ",
round(mean(katdogruluk),4),"'dir."))
#en iyi 4 çıktı
i<-4
testIndisleri<-kat[[tekrar]][[i]]
test<-dataset[testIndisleri,]
churn_gercek[[tekrar]][[i]]<-test$churn_2013_07
egitim<-dataset[-testIndisleri,]
print("Eğitim seti Sınıf değerleri dağılımı")
print(tab(egitim[,34]))
print("Test veri seti Sınıf değerleri dağılımı")
print(tab(test[,34]))
naiveBayes_modeli60<-naiveBayes(egitimDeğişkenleri, egitimHedefDeğişkenleri)
summary(naiveBayes_modeli60)
naiveBayes_modeli60
model_churn[[tekrar]][[i]]<-predict(naiveBayes_modeli60,newdata=test[-113])
print("Gerçek churn değerleri =")
print(churn_gercek[[tekrar]][[i]])
print("Tahimini churn değerleri =")
print(model_churn[[tekrar]][[i]])
################################################################
###C4.5 ALGORİTMASI ile YAPILAN MODEL
################################################################
install.packages("TunePareto")
library(TunePareto)
tekrar<-1
katSayisi<-5
set.seed(2017)
kat<-generateCVRuns(telekom$churn_2013_07,ntimes = tekrar,nfold =katSayisi,leaveOneOut =
FALSE,stratified =TRUE)
katdogruluk<-rep(0,times=katSayisi)
katdogruluk
churn_gercek<-NULL
model_churn<-NULL
uzunluklar<-sapply(kat[[tekrar]],length)
names(uzunluklar)<-NULL
for(j in 1: katSayisi){
churn_gercek[[tekrar]]<-append(churn_gercek[[tekrar]],values=list(rep(NA,times=uzunluklar[j])))
model_churn[[tekrar]]<-append(model_churn[[tekrar]],values=list(rep(NA,times=uzunluklar[j])))}
library(RWeka)
for(i in 1:katSayisi){
testIndisleri<-kat[[tekrar]][[i]]
test<-dataset[testIndisleri,]
churn_gercek[[tekrar]][[i]]<-test$churn_2013_07
```

```
egitim<-dataset[-testIndisleri,]
print("Eğitim veriseti Sınıf Değerleri Dağılımları"
print(table(egitim[,34]))
print("Test veriseti Sınıf Değerleri Dağılımları")
print(table(test[,34]))
modelC4_5<-J48(egitim$churn_2013_07~.,data=egitim)
model_churn[[tekrar]][[i]]<-predict(modelC4_5,newdata=test[-113])
print("Gerçek churn değerleri =")
print(churn_gercek[[tekrar]][[i]])
print("Tahimini churn değerleri =")
print(model_churn[[tekrar]][[i]])
#performans değerlendirmesi
(tablom<-table(model_churn[[tekrar]][[i]],model_churn[[tekrar]][[i]],dnn=c("Tahimini
Siniflar","Gerçek Siniflar")))
katdogruluk[i]<-mean(model_churn[[tekrar]][[i]]==churn_gercek[[tekrar]][[i]])}
round(katdogruluk,4)
print(paste0(katSayisi, "-Kat çapraz geçerleme sonucu elde edilen ortalama doğruluk ",
round(mean(katdogruluk),4),"'dir."))
#en iyi 3 kat çıktı
i<-3
testIndisleri<-kat[[tekrar]][[i]]
test<-dataset[testIndisleri,]
churn_gercek[[tekrar]][[i]]<-test$churn_2013_07
egitim<-dataset[-testIndisleri,]
print("Eğitim veriseti Sınıf Değerleri Dağılımları")
print(table(egitim[,34]))
print("Test veriseti Sınıf Değerleri Dağılımları")
print(table(test[,34]))
modelC4_5<-J48(egitim$churn_2013_07~.,data=egitim)
summary(modelC4_5)
modelC4_5
library(partykit)
plot(modelC4_5)
model_churn[[tekrar]][[i]]<-predict(modelC4_5,newdata=test[-113])
print("Gerçek churn değerleri =")
print(churn_gercek[[tekrar]][[i]])
print("Tahimini churn değerleri =")
print(model_churn[[tekrar]][[i]])
#################################################################
###Gini ALGORİTMASI ile YAPILAN MODEL
#################################################################
library(rpart)
tekrar<-1
katSayisi<-5
set.seed(2017)
kat<-generateCVRuns(telekom$churn_2013_07,ntimes = tekrar,nfold =katSayisi,leaveOneOut =
FALSE,stratified =TRUE)
katdogruluk<-rep(0,times=katSayisi)
katdogruluk
churn_gercek<-NULL
model_churn<-NULL
uzunluklar<-sapply(kat[[tekrar]],length)
names(uzunluklar)<-NULL
for(j in 1: katSayisi){
churn_gercek[[tekrar]]<-append(churn_gercek[[tekrar]],values=list(rep(NA,times=uzunluklar[j])))
model_churn[[tekrar]]<-append(model_churn[[tekrar]],values=list(rep(NA,times=uzunluklar[j])))}
for(i in 1:katSayisi){
testIndisleri<-kat[[tekrar]][[i]]
test<-telekom[testIndisleri,]
churn_gercek[[tekrar]][[i]]<-test$churn_2013_07
```

```
egitim<-telekom[-testIndisleri,]
print("Eğitim veriseti Sınıf değerleri Dağılımları")
print(table(egitim[,34]))
print("Test veriseti Sınıf değerleri Dağılımları")
print(table(test[,34]))
tree_model<-rpart(churn_2013_07~.,egitim,method="class")
model_churn[[tekrar]][[i]]<-predict(tree_model,newdata=test[-113],type = "class")
print("Gerçek churn değerleri =")
print(churn_gercek[[tekrar]][[i]])
print("Tahimini churn değerleri =")
print(model_churn[[tekrar]][[i]])
#performans değerlendirmesi
(tablom<-table(model_churn[[tekrar]][[i]],model_churn[[tekrar]][[i]],
dnn=c("Tahimini Siniflar","Gerçek Siniflar")))
katdogruluk[i]<-mean(model_churn[[tekrar]][[i]]==churn_gercek[[tekrar]][[i]])}
round(katdogruluk,4)
print(paste0(katSayisi, "-Kat çapraz geçerleme sonucu elde edilen ortalama doğruluk ",
round(mean(katdogruluk),4),"'dir."))
#en iyi 1 kat
i<-1
testIndisleri<-kat[[tekrar]][[i]]
test<-telekom[testIndisleri,]
churn_gercek[[tekrar]][[i]]<-test$churn_2013_07
egitim<-telekom[-testIndisleri,]
print("Eğitim veriseti Sınıf değerleri Dağılımları")
print(table(egitim[,34]))
print("Test veriseti Sınıf değerleri Dağılımları")
print(table(test[,34]))
tree_model<-rpart(churn_2013_07~.,egitim,method="class")
summary(tree_model)
show(tree_model)
prp(tree_model)
plot(tree_model)
model_churn[[tekrar]][[i]]<-predict(tree_model,newdata=test[-113])
print("Gerçek churn değerleri =")
print(churn_gercek[[tekrar]][[i]])
print("Tahimini churn değerleri =")
print(model_churn[[tekrar]][[i]])
```

11.3. EK 3: VERİ GÖRSELLEŞTİRME

Veri Görselleştirme Senaryo (Müşterililik Süresi (Gün) ve Kendi Operatör Aboneleri Arası Konuşma Sürelerinin Çeşitli Gruplamalanması)

```
##############
#Kütüphane Yükleme
Library(ggplot2)
#################################################
##############################VIOLIN GRAFİKLERİ
#################################################
gplot<-ggplot(dataset)
deg<-aes(log10(dataset$mmo_count_07), log10(dataset$age_of_line), fill=factor(dataset$churn_2013))
baslik<-"Kendi aboneleri arasındaki konuşma süreleri ile müşterililik süresine göre churn durumu
(Ayrıldı: E, Ayrılmadı: H) Violin Grafiği "
violin_grafik<-geom_violin (alpha=0.5)
lnx<-scale_x_log10()
lny<-scale_y_log10()
gplot+deg+violin_grafik+ggbaslikle(baslik)+lnx+xlab("Aynı Operatör Aboneleri Arası Konuşma Süreleri
(dk)")+lny+ylab(" Müşterililik süresi (gün)")
#################################################################Son Yükleme Tarihi
gplot<-ggplot(dataset)
deg<-aes(log10(dataset$mmo_count_07), log10(dataset$age_of_line),
fill=factor(dataset$last_reload_year))
baslik<-"Kendi aboneleri arasında konuşma süreleri ile müşterililik süresi arasında son yükleme yılına
göre churn durumu (Ayrıldı: E, Ayrılmadı: H)  Violin Grafiği "
violin_grafik<-geom_violin(alpha=0.5)
lnx<-scale_x_log10()
lny<-scale_y_log10()
facetn<- facet_wrap(~ dataset$last_reload_year)
facet<-facet_wrap (~dataset$churn_2013)
gplot+deg+violin_grafik+ggbaslikle(baslik)+lnx+xlab("Kendi Abone Arası Konuşma Süreleri
(dk)")+lny+ylab("Müşterililik süresi (gün)")+facet
#############################################################Cinsiyet
gplot<-ggplot(dataset)
deg<-aes(log10(dataset$mmo_count_07), log10(dataset$age_of_line), fill=factor(dataset$gender_flag))
baslik<-"Kendi aboneleri arasında konuşma süreleri ile müşterililik süresi arasında cinsiyetine göre churn
durumu (Ayrıldı: E, Ayrılmadı: H) Violin Grafiği "
violin_grafik<-geom_violin(alpha=0.5)
lnx<-scale_x_log10()
lny<-scale_y_log10()
facetn<- facet_wrap(~ dataset$gender_flag)
facet<-facet_wrap (~dataset$churn_2013)
gplot+deg+violin_grafik+ggbaslikle(baslik)+lnx+xlab("Kendi Abone Arası Konuşma
Süreleri(dk)")+lny+ylab("Müşterililik süresi (gün)")+facet
######## ########################################Yaş
gplot<-ggplot(dataset)
deg<-aes(log10(dataset$mmo_count_07), log10(dataset$age_of_line), fill=factor(dataset$age))
baslik<-"Kendi aboneleri arasında konuşma süreleri ile müşterililik süresi arasında yaşa göre churn
durumu (Ayrıldı: E, Ayrılmadı: H) Violin Grafiği "
violin_grafik<-geom_violin(alpha=0.5)
lnx<-scale_x_log10()
lny<-scale_y_log10()
facetn<- facet_wrap(~ dataset$age)
facet<-facet_wrap (~dataset$churn_2013)
gplot+deg+violin_grafik+ggbaslikle(baslik)+lnx+xlab("Kendi Abone Arası Konuşma
Süreleri(dk)")+lny+ylab("Müşterililik süresi (gün)")+facet
#############################################Aygıt Tipi
gplot<-ggplot(dataset)
deg<-aes(log10(dataset$mmo_count_07), log10(dataset$age_of_line), fill=factor(dataset$device_type))
```

```
baslik<-"Kendi aboneleri arasında konuşma süreleri ile müşterililik süresi arasında kullandığı cihaza göre
churn durumu (Ayrıldı: E, Ayrılmadı: H)  violin grafiği "
violin_grafik<-geom_violin(alpha=0.5)
lnx<-scale_x_log10()
lny<-scale_y_log10()
facetn<- facet_wrap(~ dataset$device_type)
facet<-facet_wrap (~dataset$churn_2013)
gplot+deg+violin_grafik+ggbaslikle(baslik)+lnx+xlab("Kendi Abone Arası Konuşma Süreleri
(dk)")+lny+ylab("Müşterililik süresi (gün)")+facet
################################################Tarife Tipi
gplot<-ggplot(dataset)
deg<-aes(log10(dataset$mmo_count_07), log10(dataset$age_of_line), fill=factor(dataset$tariff_type))
baslik<-"Kendi aboneleri arasında konuşma süreleri ile müşterililik süresi arasında kullandığı tarife tipine
göre churn durumu (Ayrıldı: E, Ayrılmadı: H)  violin grafiği "
violin_grafik<-geom_violin(alpha=0.5)
lnx<-scale_x_log10()
lny<-scale_y_log10()
facetn<- facet_wrap(~ dataset$tariff_type)
facet<-facet_wrap (~dataset$churn_2013)
gplot+deg+violin_grafik+ggbaslikle(baslik)+lnx+xlab("Kendi Abone Arası Konuşma Süreleri
(dk)")+lny+ylab("Müşterililik süresi (gün)")+facet
#payment_type
gplot<-ggplot(dataset)
deg<-aes(log10(dataset$mmo_count_07), log10(dataset$age_of_line),
fill=factor(dataset$payment_type_07))
baslik<-"Kendi aboneleri arasında konuşma süreleri ile müşterililik süresi arasında ödeme türüne göre
churn durumu (Ayrıldı: E, Ayrılmadı: H)  violin grafiği "
violin_grafik<-geom_violin(alpha=0.5)
lnx<-scale_x_log10()
lny<-scale_y_log10()
facetn<- facet_wrap(~ dataset$payment_type_07)
facet<-facet_wrap (~dataset$churn_2013)
gplot+deg+violin_grafik+ggbaslikle(baslik)+lnx+xlab("Kendi Abone Arası Konuşma Süreleri
(dk)")+lny+ylab("Müşterililik süresi (gün)")+facet
################################################################################
##############################DENSİTY GRAFİĞİ###################################
################################################################################
#######konuşma Süreleri
gplot<-ggplot(dataset)
ad<-aes(dataset$age_of_line, fill=factor(dataset$churn_2013))
density<-geom_density(alpha=0.5)
baslik<-"Müşterililık süresi ve Churn durumu Arası Density Grafigi"
gplot+ad+density+ggbaslikle(baslik)+lnx+xlab("Kendi Aboneleriyle Konuşma Süreleri ve Churn ")
###    tarife tipi
gplot<-ggplot(dataset)
ad<-aes(dataset$age_of_line, fill=factor(dataset$churn_2013))
facet<-facet_wrap (~dataset$tariff_type)
density<-geom_density(alpha=0.5)
baslik<-"Müşterililık süresi ile Churn durumunu Tarife tipine göre guruplama yaparak Density Grafigi"
gplot+ad+density+ggbaslikle(baslik)+lnx+xlab("Müşterililık süresi (gün) ve Churn durumunu (Ayrıldı: E,
Ayrılmadı: H) Tarif tipine göre gruplanmasının yoğunluk grafiği ")+facet
#################################################################Cinsiyet
gplot<-ggplot(dataset)
ad<-aes(dataset$age_of_line, fill=factor(dataset$churn_2013))
facet<-facet_wrap (~dataset$gender_flag)
density<-geom_density(alpha=0.5)
baslik<-"Müşterililık süresi ile Churn durumunu cinsiyete göre guruplama yaparak Density Grafigi"
gplot+ad+density+ggbaslikle(baslik)+lnx+xlab("Müşterililık süresi (gün) ve Churn durumunu(Ayrıldı: E,
Ayrılmadı: H)  cinsiyete göre (E:Erkek, K:Kadın, U:Belirsiz) gruplanmasının yoğunluk grafiği ")+facet
#################################################################Yaş
```

```
gplot<-ggplot(dataset)
ad<-aes(dataset$age_of_line, fill=factor(dataset$churn_2013))
facet<-facet_wrap (~dataset$age)
density<-geom_density(alpha=0.5)
baslik<-"Müşterililık süresi ile Churn durumunu yaşa göre guruplama yaparak Density Grafigi"
gplot+ad+density+ggbaslikle(baslik)+lnx+xlab("Müşterililık süresi (gün) ve Churn durumunu (Ayrıldı: E, Ayrılmadı: H)  yaşa göre gruplanmasının yoğunluk grafiği ")+facet
#################################################################Aygıt Tipi
gplot<-ggplot(dataset)
ad<-aes(as$age_of_line, fill=factor(dataset$churn_2013))
facet<-facet_wrap (~dataset$device_type)
density<-geom_density(alpha=0.5)
baslik<-"Müşterililık süresi ile Churn durumunu Kullanılan cihaz tipine göre guruplama yaparak Density Grafigi"
gplot+ad+density+ggbaslikle(baslik)+lnx+xlab("Müşterililık süresi (gün) ve Churn durumunu (Ayrıldı: E, Ayrılmadı: H)  Kullanılan cihaz tipine göre gruplanmasının yoğunluk grafiği ")+facet
############################Ödeme Tipi
gplot<-ggplot(dataset)
ad<-aes(as$age_of_line, fill=factor(dataset$churn_2013))
facet<-facet_wrap (~dataset$payment_type_07)
density<-geom_density(alpha=0.5)
baslik<-"Müşterililık süresi ile Churn durumunu Kullanılan cihaz tipine göre guruplama yaparak Density Grafigi"
gplot+ad+density+ggbaslikle(baslik)+lnx+xlab("Müşterililık süresi (gün) ve Churn durumunu (Ayrıldı: E, Ayrılmadı: H)  Ödeme türüne göre gruplanmasının yoğunluk grafiği ")+facet
```

11.4. EK 4: SHINY

Global.R

```
library(shiny)
library(caret)
library(e1071)
library(RWeka)
musteri<-read.table('C:/Users/sinan/Desktop/telekom/sonn/data/dataset.csv',header = FALSE,
sep=",",dec=".")
colnames(musteri)<-
c("age","age_of_line","tariff_type","device_type","last_reload_year","mmo_count_07","mmo_duration_0
7", "mmt_count_07","mmt_duration_07","mmo_total_count_07","mmo_total_duration_07",
"mmt_total_count_07","mmt_total_duration_07","msmo_count_07","callcenter_count_07","churn_2013_
07")
for(i in 1:16){
if(i == 6 || i == 7 || i == 8  || i == 9 || i == 10 || i == 11 || i == 12|| i == 13 || i == 14 || i == 15 )
musteri[[i]]<-as.numeric(x=musteri[[i]])
else
musteri[[i]]<-as.factor(x=musteri[[i]])}
```

Ui.R

```
shinyUI(
fluidPage(
fluidRow(column(12,titlePanel(title="KARAR AGACLARI ile MÜŞTERİ KAYIP TAHMİNİ",
windowTitle = "Shiny - Veri Madenciligi"),
div("Basarslan Sinan M.,2017. KARAR AGACLARI ile MÜŞTERİ KAYIP ANALİZİ.",
br(),
style="color:red;font-size:small"),
h6("Müşteri Kayip Analizi Uygulaması, bir Telekomunikasyon Veri Uzerinde Gerceklestirilmistir.
",style="color:red;font-size:small"),
hr())),
fluidRow(
h6(" Telekomunikasyon Müşteri Kisisel Bilgisi ",style="color:blue;font-size:small"),
column(4,div(tags$h4(style="font-size:small;background-color:#E0F2F7;",
br(),
numericInput(inputId="age",label ="yaş:",value=0),
numericInput(inputId="age_of_line",label ="Müşterililik süresi (gün):",value=0),
selectInput(inputId="tariff_type",label="Tarife Tipi :(1:Kontörlü
2:Faturalı)",choices=levels(musteri$tariff_type)),
selectInput(inputId="device_type",label="Kullanılan Cihaz Tipi
(0:bilinmeyen,3:Mobil,4:Modul,6:USb Modem,10:Akıllı
Telefon,12:Console,13:Tablet,15:Wireless):",choices=levels(musteri$device_type)),
selectInput(inputId="last_reload_year",label="Son Yükleme Tarihi
:19000101:Bilinmeyen,",choices=levels(musteri$last_reload_year)))),
h6("Aylik  Telekomunikasyon Bilgisi-1",style="color:blue;font-size:bold"),
column(4,
div(tags$h4(style="font-size:small;background-color:#E0F2F7;",
br(),
numericInput(inputId="mmo_count_07",label ="Aylık Kendi aboneleriyle konuşma sayısı (Arama)
:",value=0),
numericInput(inputId="mmo_duration_07",label ="Aylık Kendi aboneleriyle konuşma süresi (Arama)
dk  :",value=0),
numericInput(inputId="mmt_count_07",label ="Aylık Diğer aboneleriyle konuşma sayısı (Aranma)
:",value=0),
numericInput(inputId="mmt_duration_07",label ="Aylık Diğer aboneleriyle konuşma süresi (Arama)
dk :",value=0),
numericInput(inputId="mmo_total_count_07",label ="Son iki ayda Kendi aboneleriyle toplam
```

```
konuşma sayısı (Arama) :",value=0),
numericInput(inputId="mmo_total_duration_07",label ="Son iki ayda Aylık Kendi aboneleriyle
toplam konuşma süresi (Arama)dk:",value=0),
numericInput(inputId="mmt_total_count_07",label ="Son iki ayda Aylık Diğer aboneleriyle toplam
konuşma sayısı (Aranma)  :",value=0),
numericInput(inputId="mmt_total_duration_07",label ="Son iki ayda Aylık Diğer aboneleriyle toplam
konuşma süresi (Arama) dk :",value=0)))),
h6("Aylik  Telekomunikasyon Bilgisi-2",style="color:blue;font-size:small"),
column(4,
div(tags$h4(style="font-size:small;background-color:#E0F2F7;",
br(),
numericInput(inputId="msmo_count_07",label ="Aylık mesajlaşma sayısı:",value=0),
numericInput(inputId="callcenter_count_07",label ="Şikayet sayısı:",value=0)))),
column(4, div(tags$h4("Analiz Degiskenleri : "),  style="font-size:small,background-
color:#F6CECE;",
selectInput(inputId="holdout",label = "Hold-out'ta Egitim veriseti(%):",choices = c(.60,.75,.80)),
selectInput(inputId="agac",label="Ağaç*",choices = c("budanmış","budanmamış")),
h6("*Egitim verisetine c4.5 algoritmasinin uygulanmasiyla elde edilen agac"),
actionButton(inputId="hesapla",label="Riski Hesapla !")))),
fluidRow(
hr(),
column(6,   div(tags$h3("ANALİZ PERFORMANS DEĞERLENDİRME:"), style="font-
size:small;color:red;"),
verbatimTextOutput("performansDegerlendirme"),
br(),
div(tags$h3("KARAR AĞACI*:"),style="color:red;font-size:small",
verbatimTextOutput("kararAgaci"))),
column(3,div(tags$h3("TAHMİN SONUCU :"), style="font-size:small;color:red;"),
verbatimTextOutput("karar"),
plotOutput("grafik")))))
```

Server.R

```
shinyServer(function(input,output,session){
egtmIndisleri<-reactive({set.seed(1)
egitimIndisleri<-createDataPartition(y=musteri$churn_2013_07,p=as.numeric(input$holdout),list =
FALSE)  egitimIndisleri } )
model<-reactive({
egitim<-musteri[egtmIndisleri(),]
if(input$agac=="budanmamış")
C45_modeli<-J48(churn_2013_07~.,data=egitim)
else{C45_modeli<-J48(churn_2013_07~.,data=egitim,control=Weka_control(M=1,U=TRUE))}
C45_modeli})
tahminSonucu<-eventReactive(input$hesapla,{
                        yeniOrnek<-musteri[0, ]
                        yeniOrnek[1,1]<-as.numeric(input$age)
                        yeniOrnek[1,2]<-as.numeric(input$age_of_line)
                        yeniOrnek[1,3]<-as.numeric(input$tariff_type)
                        yeniOrnek[1,4]<-as.numeric(input$device_type)
                        yeniOrnek[1,5]<-as.numeric(input$last_reload_year)
                        yeniOrnek[1,6]<-as.numeric(input$mmo_count_07)
                        yeniOrnek[1,7]<-as.numeric(input$mmo_duration_07)
                        yeniOrnek[1,8]<-as.numeric(input$mmt_count_07)
                        yeniOrnek[1,9]<-as.numeric(input$mmt_duration_07)
                        yeniOrnek[1,10]<-as.numeric(input$mmo_total_count_07)
                        yeniOrnek[1,11]<-as.numeric(input$mmo_total_duration_07)
                        yeniOrnek[1,12]<-as.numeric(input$mmt_total_count_07)
                        yeniOrnek[1,13]<-as.numeric(input$mmt_total_duration_07)
                        yeniOrnek[1,14]<-as.numeric(input$msmo_count_07)
                        yeniOrnek[1,15]<-as.numeric(input$callcenter_count_07)
                        yeniOrnek[1,16]<-NA
```

```
                        tahminiSinif<-predict(model(),newdata=yeniOrnek[1,-16])
                        if(tahminiSinif=="H")
                         "Müşteri ayrılma riski yoktur."
                        else
                         "Müşteri Ayrıla riski vardır."})
kararAgaci<-eventReactive(
input$hesapla,{
model()})
performansHesabi<-eventReactive(input$hesapla,{
tahminiSiniflar<-predict(model(),newdata=musteri[-egtmIndisleri(),-16])
cm<-caret::confusionMatrix(data=tahminiSiniflar,reference=musteri[-egtmIndisleri(),16]) cm})
performanslar<-eventReactive(
input$hesapla,{
fourfoldplot(performansHesabi()$table,color = rainbow(2))})
output$karar<-renderText({
tahminSonucu()})
output$kararAgaci<-renderPrint({
kararAgaci()})
output$performansDegerlendirme<-renderPrint({performansHesabi()})
output$grafik<-renderPlot({performanslar()})})
```

11.5. EK 5: TEMEL BİLEŞEN ANALİZİ (PRINCIPAL COMPONENT ANALYSIS- PCA)

PCA kodları

```
#Kütüphane Çağırma
library(FactoMineR)
library(factoextra)
library(corrplot)
# Veri setini belirlemek
dset <-  dataset
print(dim(dset))
print(head(dset,3))
datas<-na.omit(dset)
datas<-as.data.frame(scale(dset))
print(dim(datas))
print(head(datas))
# Scatterplot Matrisler
par(ask=TRUE)
if(ncol(dset)<28) {p<-scatterplotMatrix(dset, diagonal="histogram"); print(p)}
# corrplot ile korelasyon matrisleri
library(corrplot)
par(ask=TRUE)
print(res.pca)
corel<-cor(datas)
corrplot(corel, method = "number")
corrplot(corel)
corrplot(corel, order= "AOE")
corrplot(corel, order= "hclust", addrect=2)
# FactoMineR   ile PCA
res.pca<-PCA(datas, scale=FALSE, graph=FALSE)
print(res.pca)
par(ask=TRUE)
par(mfrow=c(1,2))
plot(res.pca)
plot(res.pca, choix="var")
indiv<-TRUE
scores <- res.pca$x
loadings <- res.pca$rotation
# facto_summarize() ile pca() sonuçları dökümü
# 1,2,3 değişkenlerinin özeti
fsv<- facto_summarize(res.pca, "var", axes = 1:3)[,-1]
print(fsv)
fsvmax<- max(fsv$cos2)
print(fsvmax)
# fviz_eig() özdeğerler
p <- fviz_eig(res.pca, addlabels = TRUE, hjust = -0.4,
linecolor = "#FC4E07", barfill="red", barcolor ="white" )+ ylim(0, 85)
tit<-paste(tit,"Bileşenlerin Varyans Yüzdeleri  y-Range %85")
print(p+tma+ggtitle(tit))
p<- fviz_eig(res.pca, choice = "eigenvalue", addlabels=TRUE)
print(p+tma+ggtitle("Eigenvalue Barplot"))
# fviz_pca_var() ile Değişken Grafikleri
# Defaultplot: Korelasyon Çemberi
p<- fviz_pca_var(res.pca)+ tma
print(p+ggtitle("PCA Değişkenler Korelasyon Çemberi"))
```

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Muhammet Sinan BAŞARSLAN

Doğum Tarihi ve Yeri : 22.12.1991

Yabancı Dili : İngilizce

E-posta : muhammetsinanbasarslan@hotmail.com

ÖĞRENİM DURUMU

Derece	Alan	Okul/Üniversite	Mezuniyet Yılı
Y. Lisans	Bilgisayar Müh.	Düzce Üniversitesi	2017
Lisans	Bilgisayar Müh.	Düzce Üniversitesi	2015
Lise	Fen Bilimleri	Yalova Çiftliköy Anadolu Lisesi	2009