ANKARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

TELEKOMÜNİKASYON SEKTÖRÜNDE MÜŞTERİ KAYBI ANALİZİ

Mehmet Sabri KUNT

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

ANKARA
2019

Her hakkı saklıdır

TEZ ONAYI

Mehmet Sabri KUNT tarafından hazırlanan **"TELEKOMÜNİKASYON SEKTÖRÜNDE MÜŞTERİ KAYBI ANALİZİ"** adlı tez çalışması 26/06/2019 tarihinde aşağıdaki jüri tarafından oy birliği ile Ankara Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Danışman : Dr. Öğr. Üyesi Bülent TUĞRUL
Ankara Üniversitesi / Bilgisayar Mühendisliği Anabilim Dalı

Jüri Üyeleri:

Başkan: Doç. Dr. İhsan Tolga MEDENİ
Yıldırım Beyazıt Üniversitesi / Yönetim Bilişim Sistemleri Anabilim Dalı

Üye : Dr. Öğr. Üyesi Bülent TUĞRUL
Ankara Üniversitesi / Bilgisayar Mühendisliği Anabilim Dalı

Üye : Dr. Öğr. Üyesi Ömer Özgür TANRIÖVER
Ankara Üniversitesi / Bilgisayar Mühendisliği Anabilim Dalı

Yukarıdaki sonucu onaylarım.

Prof. Dr. Özlem YILDIRIM
Enstitü Müdür Vekili

ETİK

Ankara Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırladığım bu tez içindeki bütün bilgilerin doğru ve tam olduğunu, bilgilerin üretilmesi aşamasında bilimsel etiğe uygun davrandığımı, yararlandığım bütün kaynakları atıf yaparak belirttiğimi beyan ederim.

Tarih

26/06/2019

İmza

Öğrencinin Adı Soyadı

Mehmet Sabri KUNT

ÖZET

Yüksek Lisans Tezi

TELEKOMÜNİKASYON SEKTÖRÜNDE MÜŞTERİ KAYBI ANALİZİ

Mehmet Sabri KUNT

Ankara Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Bülent TUĞRUL

Müşteri kaybı analizi, mevcut kullanıcıların analiz edilerek, hizmet veya ürünü kullanmayı bırakma ihtimali yüksek olan müşterilerin tespit edilmesi işlemidir. Potansiyel kitle tespit edildikten sonra, pazarlama ve müşteri ilişkileri departmanları ile ortak çalışma yapılarak, müşteriyi memnun edecek kampanya veya promosyon çalışmalarının yapılmasına zemin hazırlanır. Müşteri kaybı analizi, telekomünikasyon, bankacılık, online ticaret gibi müşteri sayısı ile gelir miktarının doğru orantılı olduğu sektörlerinde hayati öneme sahiptir. Bu çalışmada genel olarak ürün veya hizmeti kullanmayı bırakma ihtimali yüksek olan müşterilerin, veri madenciliği altında çalışan karar ağaçları ve onun gelişmiş bir versiyonu olan, random forest yöntemleri ve xgboosting yöntemi ve ayrıca yaygın olarak kullanılan sınıflandırma yöntemlerinden olan naif bayes ve lojistik regresyon yöntemleri ile nasıl tespit edileceği incelenmiştir.

Haziran 2019, 67 sayfa

Anahtar Kelimeler: Müşteri kaybı analizi, churn analizi, karar ağaçları, random forest, xgboosting, naif bayes, lojistik regresyon, veri madenciliği, sınıflandırma, aykırı veri temizleme

ABSTRACT

Master Thesis

CHURN ANALYSIS IN TELECOMMUNICATION SECTOR

Mehmet Sabri KUNT

Ankara University
Graduate School of Natural and Applied Sciences
Departmant of Computer Engineering

Supervisor: Dr. Öğr. Üyesi Bülent TUĞRUL

Customer loss analysis is the process of identifying customers who are likely to quit using the service or product by analyzing existing users. Once the potential customers are identified, a joint study is carried out with marketing and customer relations departments to prepare the campaign or promotion activities that will satisfy the customer. Customer loss analysis is so important for the sectors that number of customers and income is directly proportional such as telecommunications, banking, online trade. In this study, generally we analyze how to determine the customers who are likely to stop using the product or service by using of decision trees method that is a part of data mining area and Random Forest method that is a advanced version of decision trees and xgboosting as well as the methods of naive bayes and logistic regresyon, one of the commonly used classification methods.

June 2019, 67 pages

Key Words: Customer loss analysis, churn analysis, decision tree, random forest, xgboosting, naive bayes, logistic regression, data mining, classification, handling outliers data

ÖNSÖZ ve TEŞEKKÜR

Çalışmalarımı yönlendiren, araştırmalarımın her aşamasında bilgi, öneri ve yardımlarını esirgemeyen danışman hocam Dr. Öğr. Üyesi Bülent TUĞRUL'a, çalışmalarım sırasında desteğini esirgemeyen yöneticim Osman GÜYÜM, çalışma arkadaşlarım Cansu ÜÇKUŞ, Vehbi ULU ve Yasin SARI'ya ve tüm çalışma süreci boyunca fedakarlık gösteren ve beni destekleyen eşim Seher ve kızım Esma'ya teşekkür ederim.

Mehmet Sabri KUNT
Ankara, Haziran 2019

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

KISALTMALAR DİZİNİ

ETL	Extract Transform and Load
KNIME	Konstanz Information Miner
WEKA	Waikato Environment for Knowledge Analysis

1. GİRİŞ

Yapılan araştırmalara göre günümüz rekabet koşullarında çoğu sektör için yeni müşteri kazanmak, mevcut müşteriyi mutlu edip elde tutmaktan çok daha maliyetli olmaktadır (Kotler ve Keller 2015). Bu yüzden yeni müşteri kazanmak yerine var olan müşterilerimizi elde tutmak daha kazançlı hale gelmektedir (Euler 2005). Yeni müşteri kazanabilmek için ayrılan pazarlama ve alt yapı maliyetlerinin toplamı ile aynı dönemlerde kazanılan yeni müşterilerin sayısı oranlanırsa, müşteri başına oldukça yüksek bütçeler çıktığı görülür.

Telekomünikasyon, finans, online ticaret gibi sektörlerde faaliyet gösteren firmaların değerleri, kasalarında bulunan paranın yanı sıra, sahip oldukları müşteri sayısı ile de ölçülmektedir. Bu firmalar için daha karlı satış yapmanın yolu, daha az maliyet ile daha çok müşteriye sahip olmaktan geçer. Bu da sahip oldukları müşterilerin daha uzun süre hizmet veya ürünü kullanmaya devam etmeleri ile olur. Sahip olunan müşterileri elde tutabilmek için ilk olarak hizmet veya ürünü kullanmayı bırakma ihtimali yüksek olan müşterileri tespit etmek ve bu kitleye yönelik pazarlama ve promosyon çalışmaları yapmak gerekir. Yani bu kitle tespit edip, memnun edebilirsek, müşteri kaybı en aza indirilmiş olur.

Şirketler genel olarak, kazançlarının %80'ini, müşterilerinin %20'sinden elde ederler. Bu yüzden aboneliğini iptal etme olasılığı yüksek olan müşterileri önceden tahmin etmek oldukça önemlidir (Anonymous 2018). Eğer kaybettiğimiz müşteriler gelirin %80'ini elde ettiğimiz müşterilerden ise bu önem daha da artmaktadır.

Telekomünikasyon sektöründe şirketlerin çok yüksek sayılarda hizmet veya ürünlerini kullanan müşterileri vardır. Sayıları milyonlara ulaşan müşterilere ait geçmişe dönük verilerde düşünüldüğünde önümüze çok büyük bir veri kitlesi çıkmaktadır. Bu çalışmamızda son zamanlarda iyice gelişen veri tabanı sistemleri ve veri madenciliği teknikleri ile bu verilerin işlenmesi ve modellenmesi sağlanmıştır. Model, iptale gidecek olan müşterileri tespit edecek şekilde oluşturulmuş ve test verileri üzerinde uygulanarak, iptal etme potansiyeli olan müşterilerin doğru tespit edilip edilmediği ölçülmüştür.

Bu çalışmada veri madenciliği tahminleme ve sınıflandırma yöntemlerinden olan ve birçok müşteri kayıp analizi çalışmasında kullanılan karar ağaçları, random forest ve xgboosting modelleri ayrıca yaygın sınıflandırma metotlarından olan naif bayes ve lojistik regresyon yöntemleri kullanılmıştır. Beş yöntem ile de aynı veri üzerinde model oluşturma ve sonrasında doğrulama ve test işlemleri yapılarak sonuçlar karşılaştırılmıştır. Çalışmamız internet abonelerine yönelik yapılmış ve internet hizmetini iptal etme ihtimali yüksek olan abonelerin tespiti sağlanmıştır.

2. LİTERATÜR TARAMASI

2009 yılında yapılan bir çalışmada bir telekom şirketinin müşterilerine ait yapısal veriler kullanılarak müşteri kayıp analizi yapılmış. Ayrıca bu verilere ek olarak çağrı merkezi ve web sitesi loglarından oluşan yapısal olmayan veriler metin ve web madenciliği teknikleri ile yapısal hale getirilerek ana veri setine eklenerek model tekrardan oluşturulmuştur. Yapısal olmayan bu veriler içerisinden çıkarılan yapısal verinin kullanımı ile model başarı oranları artmıştır. Bu çalışmada metin ve web madenciliği gibi teknikler ile şirketlere ait anlamsız yığınlar halinde tutulan yapısal olmayan verilerin aslında ne kadar değerli olduğu ifade edilmiştir (Dolgun vd. 2009).

2009 yılında Burez ve Van den Poe tarafından yapılan çalışmada müşteri kayıp analizi konusu üzerinden çalışılmış ve bu çalışmada makine öğrenmesi tekniklerinde kullanılan düzensiz(imbalanced) veri 6 farklı kategoride ele alınmıştır. Çalışmada düzensiz verinin nasıl işleneceği ve makine öğrenmesi yöntemlerinde kullanılacağı üzerine çalışılmıştır ve düzensiz verinin ele alınması sonrası model başarısına ne kadar etkisi olduğu araştırılmıştır. Bu çalışmada sınıflandırma yöntemi olarak random forest ve lojistik regresyon kullanılmıştır (Burez vd. 2009).

Coşkun ve Baykal 2011 yılında yaptıkları çalışmada veri madenciliği tekniğinin temel amacının analiz yöntemi ile bilgi çıkarımı zor olan veri yığınlarını inceleyerek bu yığınlar içerisinde kalmış gizli, faydalı ve anlamlı verileri açığa çıkarmak olarak tanımlamışlardır. Bu açığa çıkarılan bilgileri içerisinde barındıran bir modelin oluşturulması ile daha sonra gelecek bir veri nesnesi hakkında da bilgi sahibi olmamızı ve yorum yapmamızı sağlamaktadır (Coşkun ve Baykal 2011).

2014 yılında yapılan bir çalışmada telekomünikasyon sektörüne ait müşterilerin aboneliklerini iptal etmeleri durumunda bağlantıda oldukları diğer abonelerin de iptale gitme olasılıkları değerlendirilmiş. Bu değerlendirmeyi yapabilmek için abonelerin birbirleri ile olan telefon konuşmalarının bulunması için şirkete ait arama kayıtları (CDR) işlenerek bir network oluşturulmuş. Bu sayede iptale giden müşterinin etkileyeceği diğer müşteriler tespit edilerek modele dahil edilmiştir (Kim vd 2014). Bizim çalışmamızda kullanacağımız müşteri setinde sadece internet aboneleri olduğu

için arama kayıtları verisinden söz edilememektedir. Bu yüzden bizim çalışmamızda abonelerin birbiri ile oluşturdukları network kullanılmamıştır.

Telekom sektörü dışında bankacılık sektörü içinde müşteri kaybı analizi önemli olduğu için, 2015 yılında karaağaç tarafından bu sektöre özel bir çalışma yapılmış ve yapılan çalışmada karar ağaçları ve lojistik regresyon yöntemleri kullanılmıştır. Karaağaç da çalışmasında diğer çalışmalarda ve bizim çalışmamızda olduğu gibi veri setini ön işleme aşamalarından geçirmiş ve oluşan veri ile modelini oluşturmuştur. Oluşturduğu model sonrası ortaya çıkardığı karışıklık matrisine göre doğruluk oranı %89 olmuştur (Karaağaç 2015).

2015 yılında yapılan bir diğer çalışmada yine telekom sektöründe müşteri kayıp analizi incelenmiş ve yöntem olarak yapay sinir ağları, destek vektör makineleri, karar ağaçları, naif bayes ve regresyon analizi algoritmaları uygulanmış ve sonuçları birbiri ile karşılaştırılmıştır. Bu çalışmada farklı olarak her yöntem tek tek uygulandıktan sonra bir de performans yükseltici (boosting algorithm) bir başka yöntem var olan algoritmalar üzerine uygulanmıştır. Performans yükseltici yöntem kullanılmadan önce en iyi sonucu yapay sinir ağları verirken, performans yükseltici uygulandıktan sonra en iyi sonucu destekçi vektör makinaları vermiştir (Vafeiadis vd. 2015).

Bizim çalışmamızda olduğu gibi müşteri kayıp analizi ile ilgili çalışmalardan biri 2017 yılında yapılmış ve "Makine Öğrenmesi Yöntemleriyle Müşteri Kaybı Analizi" başlığı ile yayınlanmıştır. Bu çalışmada Destek Vektör Makineleri, Naif Bayes ve Çok Katmanlı Yapay Sinir Ağları kullanılmıştır. Bizim çalışmamızda olduğu gibi bu çalışmada da firma ürününü kullanmaya devam eden ve kullanmayı bırakan olmak üzere veriyi 2 sınıfa ayırmayı hedeflemiştir. Çalışmada kullanılan müşteri verisinde 21 adet öz nitelik kullanılmıştır. Bu çalışmada bizim çalışmamızdakine göre çok küçük bir veri seti ile çalışılmıştır. Bu çalışmadaki veri setinde 4667 adet müşteriye ait bilgi bulunmaktadır. Bu çalışma sonucunda en başarılı model %91,35 doğruluk oranı ile Yapay sinir ağları olmuştur. Çalışmada Naif Bayes %87,15 ve Destek Vektör Makinaları ise %77,89 doğruluk değerine ulaşmıştır. Destek vektör makinalarının düşük

performans değerler üretmesi veri setindeki niteliklerden kaynaklandığı sonucu çıkarılmıştır (Kaynar vd. 2017).

Coussement tarafından 2017 yılında yapılan çalışmada müşteri kayıp analizi öncesi veri hazırlama aşamasının ne kadar önemli olduğuna değinilmiştir. Veri setinde yer alan niteliklerin sürekli (continues) tipinde olanların ayrık (discrete) tipler olarak çevrilmesinin sınıflandırma algoritmalarının çalışması açısından ne kadar önemli olduğuna değinilmiştir. Bu işlem sonrası aykırı ve eksik verilerinde belirli bir kategori altına toplanarak temizlenmiş olması da ayrı bir avantaj olarak anlatılmıştır. Çalışmada Avrupa'da hizmet veren bir mobil telefon operatörüne ait müşterilerin 156 adet kategorik 800 adet sürekli niteliği kullanılmıştır. Çalışmada %4,52 oranında iptal oranı olan 30.104 adet müşteriye ait veri bulunmaktadır (Coussement 2017).

Diğer çalışmalardan farklı olarak P.Spanoudes ve T.Nguyen müşteri kayıp analizi çalışmasını sadece derin öğrenme tekniğini kullanarak ele almış ve 3 farklı telekom operatörüne ait verileri kullanmıştır. Bu çalışmada veri seti boyutlarının küçültülmesinin sonuca pozitif etki ettiğinden bahsedilmiş ve kullanılan derin öğrenme yönteminin müşteri kayıp analizi konusunda diğer tekniklere yakın sonuçlar çıkardığı gözlemlenmiştir (Spanoudes ve Nguyen 2017).

Bir diğer çalışmada 3333 adet telekom müşterisine ait 20 farklı nitelik bulunduran veri seti ile müşteri kayıp analizi çalışması yapılmış ve çalışmada lojistik regresyon, naif bayes, karar ağaçları, destek vektör makineleri ve yapay sinir ağları yöntemlerinin yanı sıra hibrit bir modelde kullanılmıştır. Bu çalışmada WEKA uygulaması kullanılmış ve bizim çalışmamızda da kullandığımız KNIME üzerinden yaptığımız korelasyon öznitelik çalışması WEKA üzerinde yapılmıştır. Korelasyon çalışması sonrası 20 olan öznitelik sayısı 10'a düşürülmüş ve tüm çalışma bu 10 öznitelik ile yapılmıştır. Çalışma sonrası en iyi sonucu yapay sinir ağları %91 ile vermiş ve %89 ile onu karar ağaçları takip etmiştir. Bu çalışmada da yapay sinir ağlarının müşteri kayıp analizinde ne kadar başarılı olduğu görülmüştür (Günay 2018).

Yapılan başka bir çalışmada telekomünikasyon sektöründe müşteri kayıp analizi için bizim çalışmamızda da kullandığımız ilişkisiz sınıflandırıcıların yanı sıra ilişkili sınıflandırıcılarda kullanılarak iki farklı tipteki model yapısının karşılaştırılması yapılmış. Çalışmada ilişkili sınıflandırıcı modeli oluşturmak için müşterilerin birbiri ile olan ilişkileri tespit edilerek bir çizge(graph) yapısı oluşturulmuş. Çalışmanın sonucunda ilişkili modelin ilişkisiz modele göre çok daha zayıf kaldığı fakat ilişkisiz modelin yanlış sınıflandırdığı bazı müşterileri doğru sınıflandırdığı tespit edilmiştir. İlişkili ve ilişkisiz modellerin beraber kullanılması sonrası sonuçların pozitif olarak etkilendiği tespit edilmiş (Verbeke vd 2018).

3. MATERYAL

Kullanılan tahminleme ve sınıflandırma modellerinde başarı oranı kullanılan verinin kalitesi ve miktarı ile doğru orantılıdır. Bu yüzden modeli eğitmek için kullanılacak olan verinin hazırlanması en önemli aşamalardandır. Bu çalışmamızda modeli oluşturma, eğitme ve kullanma aşamalarına geçmeden önce modeli eğitirken ve tahmin yaparken kullanacağımız veri seti özellikleri çıkarılmıştır. Belirlenen veri seti özelliklerine göre de veri seti oluşturulmuş ve uygun formata getirilmiştir.

Çalışmamız bir telekomünikasyon şirketine ait anonim veriler ile yapılmıştır. İlgili telekomünikasyon şirketinin internet abonelerine ait veriler oluşturulurken her abonenin veri işleme izni verip vermediği kontrol edilmiş ve sadece veri işleme izni olan abonelere ait veriler anonimleştirilmiş ve kişisel verilerden ayrıştırılarak hazırlanmıştır. Bu kapsamda kullandığımız veriler KVKK kapsamında kullanılmaya uygun hale gelmiştir.

Müşteri kaybı analizi için 2 farklı abone verisine ihtiyaç vardır. Bunlardan birincisi aboneliğini iptal etmiş olan abonelere ait veriler, bir diğeri ise aktif olarak hizmeti kullanmaya devam eden abonelere ait verilerdir. Biz veri setimizde abonelik durumu ifade eden bu veriyi “churn oldumu” niteliği ile isimlendirecek ve bu niteliği eğitim, doğrulama ve test aşamalarında sınıf etiketi olarak kullanacağız.

Telekomünikasyon sektöründe aboneliğini iptal eden aboneler iki farklı şekilde incelenmiştir. Bunlardan birincisi aboneliğini kendi isteği ile iptal eden abonelerdir ve çalışmamızdan gönüllü kayıp olarak adlandırılmıştır. İkincisi ise aboneliğini kendi isteği dışında bir sebepten dolayı iptal eden abonelerdir ve çalışmamızda gönülsüz kayıp olarak adlandırılmıştır.

Gönüllü kayıp, müşterinin kendi isteği ile ürün veya hizmeti kullanmayı bırakarak aynı ürün veya hizmeti başka bir firma üzerinden kullanması veya hiç kullanmaması durumudur. Günümüz rekabet ortamında genel olarak müşteriler kendileri için en iyi teklifi veren bir başka firmaya geçme eğilimindedir (Odabaş 2017).

Gönülsüz Kayıp, müşterinin kendi tercihi dışında oluşan olaylardan dolayı yaşanan kayıplardır. Genel olarak müşterinin mevcut firmanın hizmet vermediği bir bölgeye taşınması veya ülke değiştirmesi veya ölüm gibi sebeplerden dolayı yaşanan iptallerdir (Odabaş 2017).

Veri madenciliği ile tahmin çalışmaların da gönülsüz kayıplar göz ardı edilir ve eğitim setine dâhil edilmezler. Bu çalışmada da veri setini hazırlarken gönülsüz kayıplar dışarıda bırakılmış ve iptal olan aboneler listesine sadece gönüllü olarak kendi isteği ile aboneliğini iptal eden aboneler eklenmiştir.

3.1 Veri Setinin Oluşturulması

Modeli eğitmek için kullanılacak eğitim modeli içerisinde hem iptal hem de aktif olan aboneler yer almaktadır. İptal aboneler seçilirken son 2 yıl içinde internet aboneliğini iptal etmiş abonelere ait veriler kullanılmıştır. Aktif abone verisi için ise bu çalışmanın yapıldığı tarihte aktif olarak hizmet alan aboneler seçilmiş ve bu abonelerin önceki 6 ay boyunca aktif olması şartı aranmıştır.

Bir aboneyi iptale götüren veya aktif olarak hizmeti kullanmaya devam etmesini sağlayan sebepler genel olarak bir anda ortaya çıkmazlar ve bir süre abonenin memnuniyetsizliği devam eder. Çalışmamızda bu devam eden süreci yakalayabilmek ve doğru analiz yapabilmek için 6 aylık süreç değerlendirilmiştir.

İptal olan abonelerin verileri alınırken iptal tarihi ve 6 ay öncesine kadar olan veriler bulunmuş ve o tarih aralığındaki veri oluşturulmuştur. Ayrıca aktif abonelerin de günümüzden 6 ay önceki tarihe göre verileri oluşturulmuştur. Bu durum bize bir abonenin iptale gitmeden önceki 6 aylık süreçteki durumunu ve hala aktif olarak hizmeti kullanan abonenin 6 aylık durumunu değerlendirmemizi sağlamıştır.

İptale giden abonelerin 6 aylık verisini aldığımız için, geçmiş 2 yıldaki iptal eden abonelerin hepsine ait, iptal tarihinden itibaren 6 ay önceki verilerin bulunması ve işlenmesi gerekmiştir. Bu durumda veri setini oluştururken çok büyük verilerin işlenmesini gerektirmiştir.

Örnek veri setini kullandığımız telekomünikasyon şirketi bünyesinde internet, televizyon ve telefon hizmeti verilmektedir. Bu kapsamda bireysel ve tüzel internet abonelerine ait aşağıdaki bilgilerin kullanılmasına karar verilmiş, veriler veri tabanından SQL scriptler ile oluşturularak özet haline getirilmiş ve CSV formatında dosyalara yazılmıştır. Oluşturulan sql cümlecikleri bu çalışmanın kapsamı dışında olduğu için detaylı olarak verilmemiştir.

Abonelere ait aşağıdaki maddelerde yer alan bilgiler çıkarılmış ve veri seti oluşturulmuştur.

- Televizyon aboneliği var mı?
- Televizyon Sinema paketi aboneliği var mı?
- Telefon aboneliği var mı?
- Abonelik yaşı
- Taahhüt vermiş mi?
- Kalan taahhüt
- Son 6 ayda kullandığı tüm hizmetlere ödediği fatura ortalaması
- Son 6 ayda internet hizmeti için ödediği fatura ortalaması
- Son 6 ayda aboneye çıkan fatura sayısı
- Son 6 ayda geç ödediği fatura sayısı
- Son 6 ayda yaptığı kota aşım sayısı
- Son 6 ayda yaptığı ortalama download miktarı
- Son 6 ayda yaptığı ortalama upload miktarı
- Son 6 ayda abonenin bıraktığı toplam çağrı sayısı
- Abonenin durumu (churn oldumu)

Yukarıdaki bilgiler daha öncede ifade edildiği gibi veri işleme izni alınan aboneler için oluşturulmuş ve içerisine müşteri ile ilişki kurulacak bir bilgi eklenmemiştir. Bu işlem sonrasında veriler tamamen anonim hale gelmiştir. Ayrıca verilerin oluşturulması ve işlenmesinde kurum iç networkünde yer alan sunucu sistemleri kullanılmış ve kurum

dışına ham veri çıkarılmamıştır. Bu hali ile veri seti KVKK kapsamında kullanılabilir bir veri haline getirilmiştir.

Veri setini oluşturan nitelikler kurum içerisinde çalışan veri analizi uzmanı, müşteri ilişkileri uzmanı, çağrı merkezi uzmanı, veri tabanı uzmanı ve iş zekâsı uzmanı, veri bilimi uzmanı kişiler ile çalışmalar sonucunda belirlenmiştir. Bu çalışmalar sırasında wide band delphi yöntemi kullanılmıştır.

Verileri oluştururken her satıra yeni oluşturulan sıralı bir ID alanı eklenmiştir. Çalışmamızın ileriki aşamalarında modelin tahminlerinin başarısını gözlemlerken bu ID alanı üzerinden ilişki kurulabilecektir. Bu ID alanı üzerinden verinin ait olduğu aboneye ulaşmak mümkün değildir. ID alanı sadece bu ilişki için kullanılmış ve sonuç özet verilerinde yer almamıştır.

Yukarıda maddeler olarak verdiğimiz veri setimize ait aktif ve pasif abonelerin olduğu örnek veri seti **çizelge 3.1** de verilmiştir.

Çizelge 3.1 Örnek veri seti

Row id	TV Abonesi Mi?	TV Sinema Paket Abonesi Mi?	Telefon Abonesi Mi?	Abonelik Yaşı Ay	Taahütü Var Mı?	Kalan Taahhüt Ay	Toplam Fatura Ortalaması	İnternet Fatura Ortalaması	Fatura Sayısı	Geç Ödenen Fatura Sayısı	Kota Aşım Sayısı	Download GB	Upload GB	Toplam Çağrı Sayısı	Churn Oldu Mu?
1	1	0	0	143	0		45	24	6	0	0	2,9	0,3	1	0
2	1	1	0	119	1	3	122	21	6	6	0	15,8	1,3	1	0
3	1	1	0	29	0		60	0	7	0	7	0	0	0	1
4	0	0	0	25	0		20	0	8	8	0	0	0	0	1
5	1	1	1	92	1	15	40	10	6	3	0	9,7	0,5	0	0
6	1	0	0	36	0		45	23	7	7	0	15,8	0,8	0	1
7	1	0	0	24	0		37	21	7	0	0	9,4	1	0	1
8	1	1	1	74	1	16	47	15	6	1	0	41,8	1,4	0	0
9	0	0	0	12	0		45	25	6	5	0	36,7	5,2	0	1
10	1	1	0	43	1	6	102	59	6	3	0	32,4	1,5	1	0
11	1	0	0	39	1	8	42	22	6	0	0	60,6	2,6	0	0
12	1	1	1	34	1	17	59	15	6	4	0	119,9	5,8	0	0
13	1	1	1	33	1	14	40	10	6	6	0	1,3	1,6	2	0
14	1	1	0	35	1	20	92	16	7	3	0	61,3	3,8	0	1

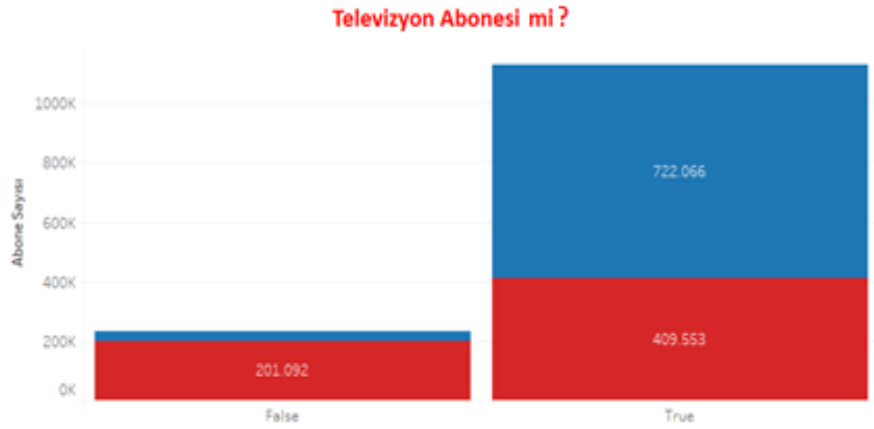
Çizelge 3.1 Örnek veri seti (Devam)

15	1	0	1	27	1	14	45	20	6	5	0	8,1	0,6	0	0
16	1	0	0	24	0		22	6	7	0	0	10,7	1	0	1
17	0	0	1	14	0		39	25	7	3	0	10,8	1,2	0	1
18	1	1	0	25	0		33	7	7	1	0	1,6	0,1	0	1
19	1	0	0	18	1	6	36	13	6	0	0	86,6	4,1	0	0
20	1	0	0	15	1	7	42	22	6	0	0	100,4	5,7	3	0
21	1	0	1	12	1	12	32	14	6	6	0	76,2	5,9	0	0
22	1	1	1	7	1	16	47	15	6	3	0	36,2	3,7	0	0
23	1	0	0	11	1	15	743	0	7	0	0	0,7	0,2	0	1

Veri tabanında bulunan verilerden üretilen ve yukarıda **çizelge 3.1**'de verilen veri seti üzerinden model oluşturmaya geçmeden önce, birçok ön hazırlık çalışması yapmak gerekir. Bu ön hazırlık çalışmalarından bazıları, eksik verilerin incelenmesi ve değerlendirilmesi, aykırı verilerin (outliers) tespiti ve düzenlenmesi ve sınıf etiketi ile diğer alanlar arasındaki ilişkinin(korelasyon) tespiti olarak verilebilir. Bu çalışmada da veri seti üzerinde bu veri ön hazırlık çalışmaları uygulanmıştır.

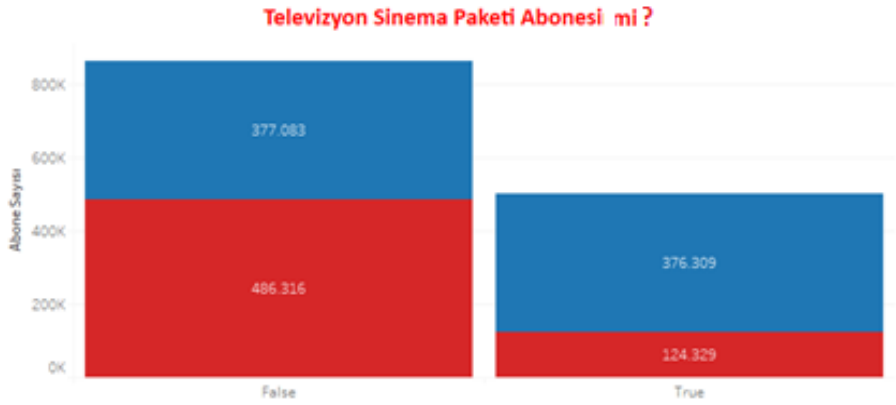
Veri oluşturma ve sonrasındaki ön hazırlık aşamaları, veri madenciliği ve makine öğrenmesi yöntemleri için hayati öneme sahiptir ve genel olarak en çok zaman bu aşamalarda harcanır. Bu çalışmada yapılan veri ön hazırlık çalışmaları yöntem başlığı altında detaylandırılmış ve uygulama başlığı altında nasıl uygulandığı anlatılmıştır.

Veriler hazırlandıktan sonra yapılan temizleme, eksik verilerin tamamlanması, aykırı(outlier) verilerin bulunması ve ilişkisiz olan alanların veri setinden çıkarılması işlemleri yapılmadan önce veri setindeki değerlerin istatistiksel dağılımı oluşturulmuş ve grafiksel olarak aşağıda verilmiştir **(Şekil 3.1-3.14)**. Bu dağılımlar oluşturulurken, kırmızı kısımların iptal abonelere ait, mavi kısımların ise hizmet almaya devam eden aboneliklere ait olacak şekilde oluşturulmuştur. Diyagramlar incelendiğinde bazı alanlarda olmaması gereken değerlerin geldiği görülmüştür. Örneğin son altı ayın fatura ortalamaları alınmasına rağmen fatura sayısı bilgisi 6'dan büyük gelmektedir veya abonelik yaşı kolonu incelendiğinde negatif olmaması gereken bir bilginin negatif olarak da geldiği görülmüştür. Bu durum veride temizlik yaparken göz önünde bulundurulacak ve gereksiz olanlar temizlenecektir.



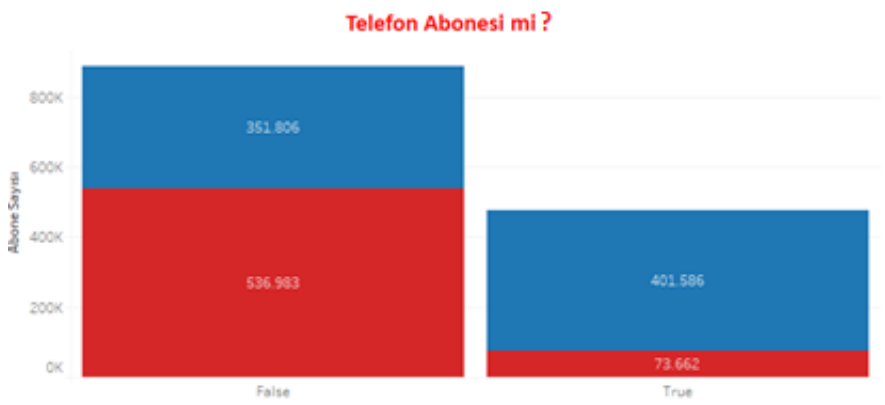
Şekil 3.1 Televizyon aboneliği durumuna göre abone iptal sayıları

Yukarıda **Şekil 3.1'**deki dağılım grafiği incelendiğinde televizyon abonesi olmayan abonelerde çok yüksek miktarda iptal oranı olduğu görülmektedir. Televizyon abonesi olan kısımda ise abonelerin yarıdan daha fazla oranda aboneliğe devam ettiği görülmektedir.



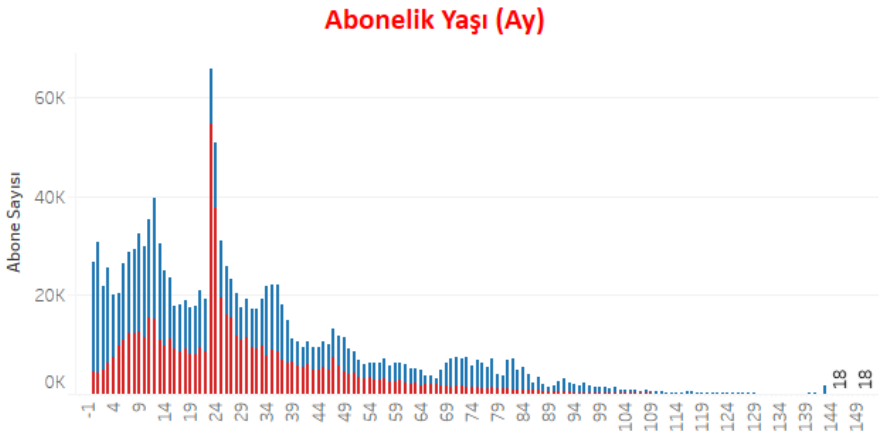
Şekil 3.2 Televizyon sinema aboneliği durumuna göre abone iptal sayıları

Yukarıda **Şekil 3.2'**deki dağılım grafiği incelendiğinde televizyon sinema paketi abonesi olmayan abonelerin yarıdan fazlası aboneliğini iptal ettirmişken paket aboneliği olanların büyük bir oranda aboneliğine devam ettiği görülebilmektedir. Bu durum paket aboneliği durumunun sınıf etiketimiz ile ilişkili olduğunuda göstermektedir.



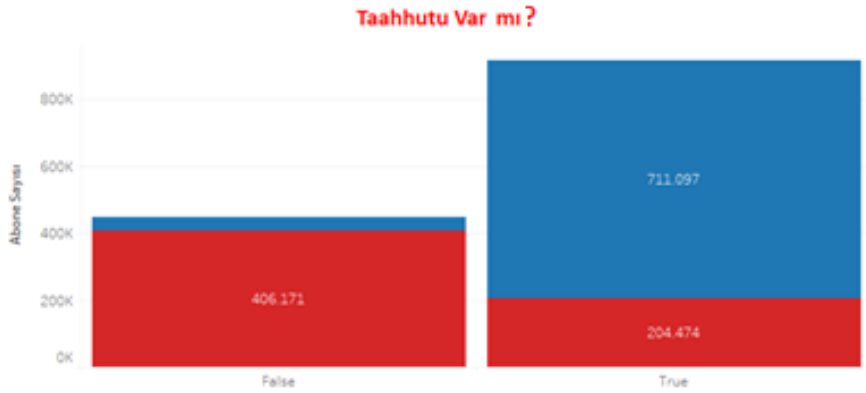
Şekil 3.3 Telefon aboneliği durumuna göre abone iptal sayıları

Yukarıda **Şekil 3.3'**deki dağılım grafiği incelendiğinde telefon abonesi olmayan abonelerin yarıdan fazlası aboneliğini iptal ettirmişken telefon aboneliği olanların büyük bir oranda aboneliğine devam ettiği görülebilmektedir.



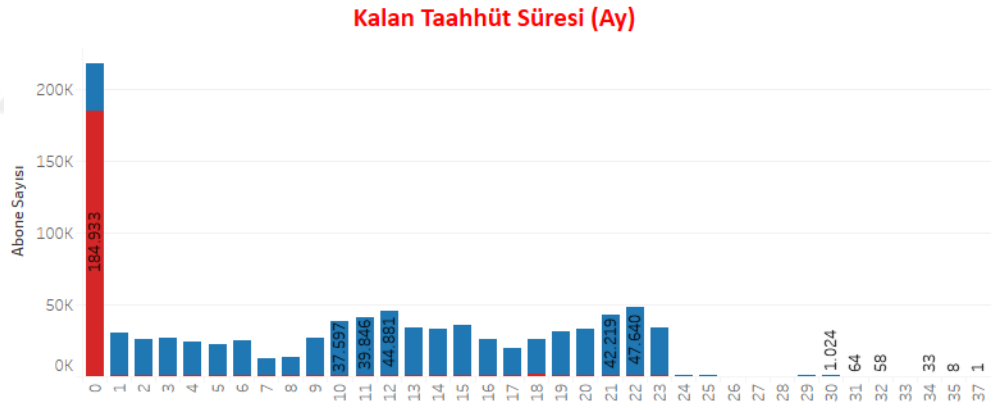
Şekil 3.4 Abonelik yaşı bilgisine göre abone iptal sayılar

Yukarıda **Şekil 3.4**'deki dağılım grafiği incelendiğinde aboneye ait hizmetin yaşını ifade eden abonelik yaşı 2 yılı bulan abonelerde yüksek iptal oranları gözlemlenebilmektedir. Ayrıca 100 üzeri abonelik yaşı olan az miktarda abone verisinin var olduğu görülebilmektedir. Bu kısım aykırı veri olarak değerlendirilebilir.



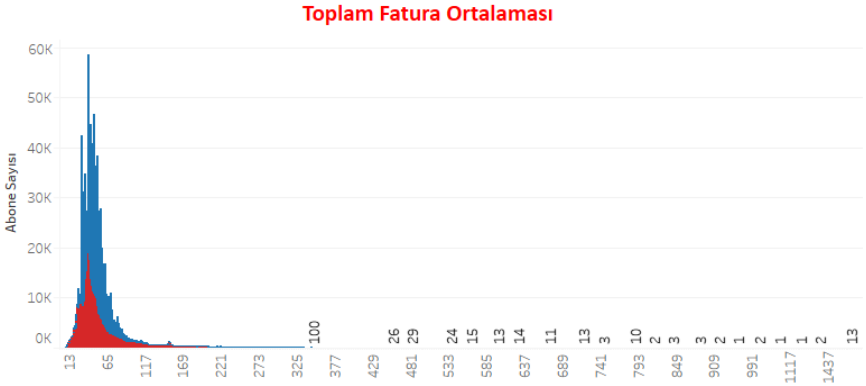
Şekil 3.5 Aboneye ait taahhüt var mı bilgisine göre abone iptal sayıları

Yukarıda **Şekil 3.5**'deki dağılım grafiği incelendiğinde taahhütü olan abonelerde iptal oranının çok düşük olduğu görülürken taahütü olmayan abonlerde bu durumun tam tersi olduğu görülebilmektedir. Sınıf etiketinin belirlenmesinde bu bilginin ne kadar belirleyici olduğu bu dağılımda açıkça görülebilmektedir.



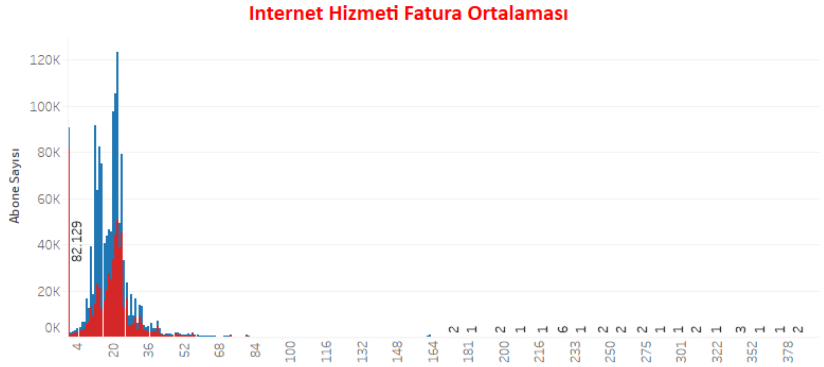
Şekil 3.6 Taahhüt veren abonelerin kalan taahhüt bilgisine göre abone iptal sayıları

Yukarıda **Şekil 3.6**'deki dağılım grafiği incelendiğinde geriye kalan taahhütü yok ise yüksek oranda iptal ile karşılaşırken bir ay ve üstü taahhütü kalan abonelerde neredeyse hiç iptal olmadığı görülebilir. 24 aydan daha yüksek taahhüt alınmadığı için bu grafikte görülen 24 üstü veriler bozuk verilere işaret etmektedir. Bu bozuk veriler aykırı veri olarak ileriki aşamalarda ele alınmıştır.



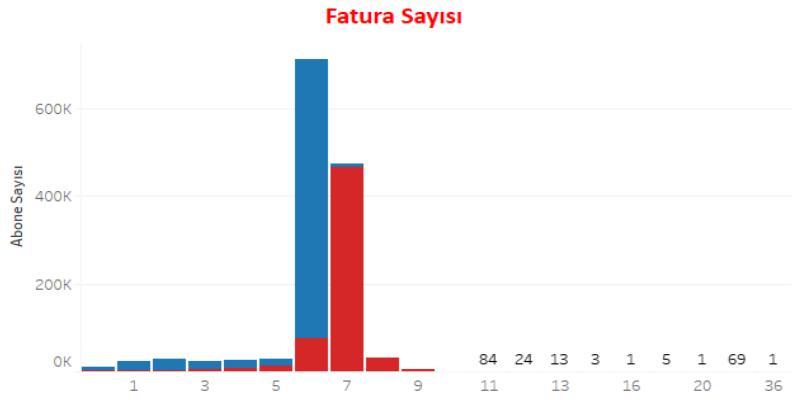
Şekil 3.7 Abonenin tüm hizmetlerine ait son 6 ay fatura ortalama bilgisine göre abone iptal sayıları

Yukarıda **Şekil 3.7**'deki dağılım grafiği incelendiğinde abonenin tüm hizmetlerine ait fatura ortalamasının sınıf etiketine göre dağılımı aykırı verilerden dolayı net olarak anlaşılamamaktadır. Aykırı veriler temizlendiği zaman bu dağılımın daha düzenli olması beklenmektedir.



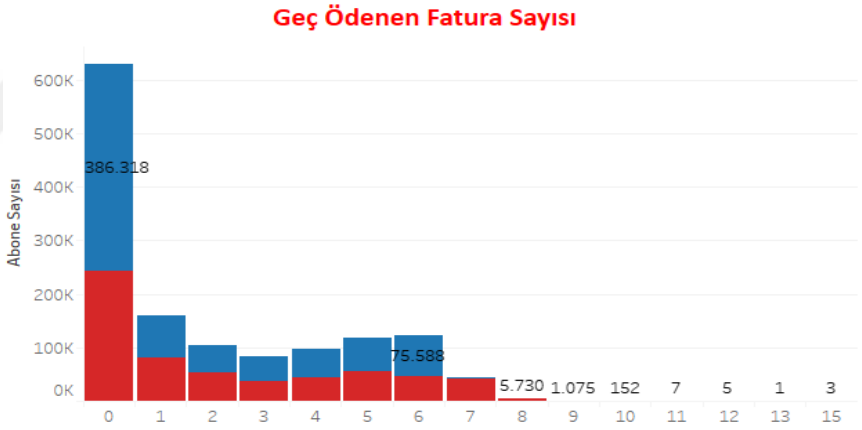
Şekil 3.8 Abonenin internet hizmetine ait son 6 ay fatura ortalama bilgisine göre abone iptal sayıları

Yukarıda **Şekil 3.8**'deki dağılım grafiği incelendiğinde abonenin internet hizmetine ait fatura ortalaması yükseldikçe iptal oranının arttığı görülebilmektedir. 50TL üzeri fatura ortalaması olan abone sayısı çok az olduğu için bu kısımlar aykırı veri olarak değerlendirilebilir. Aykırı veri temizliği sonrası bu dağılımın daha düzenli hale gelmesi beklenmektedir.



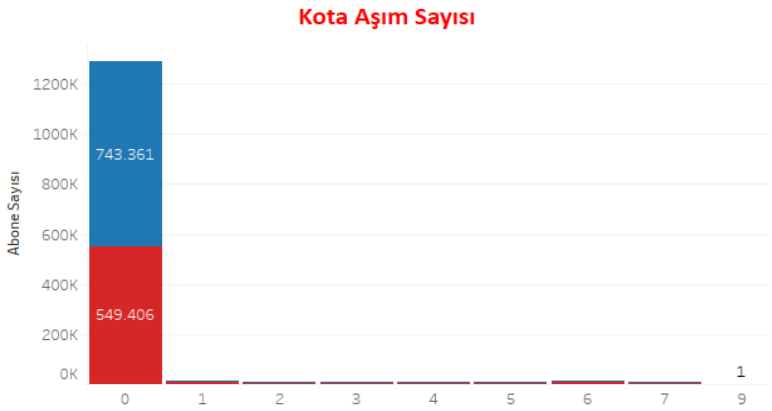
Şekil 3.9 Aboneye son 6 ayda çıkan fatura sayısı bilgisine göre abone iptal sayıları

Yukarıda **Şekil 3.9**'deki dağılım grafiği aboneye ait son 6 ayda çıkan fatura sayısını göstermektedir. Grafik incelendiğin ilk olarak 7'den daha büyük faturası olan aboneler olduğu görülmektedir ki bu aykırı veri olduğu anlamına gelmektedir.



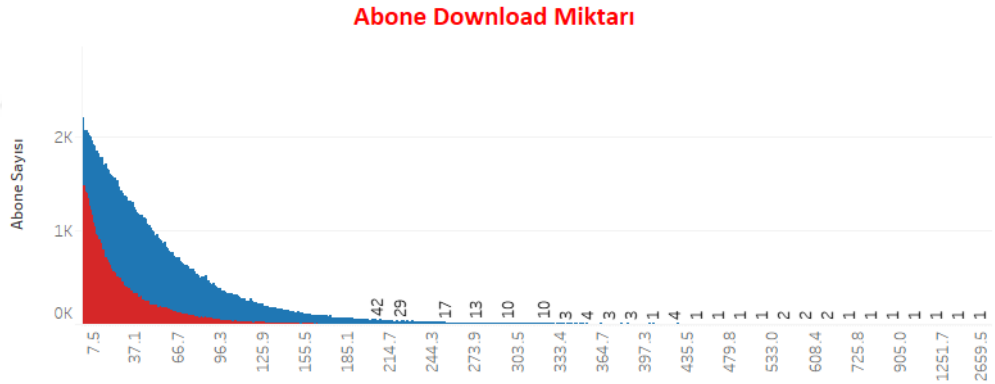
Şekil 3.10 Abonenin son 6 ayda geç ödediği fatura sayısı bilgisine göre abone iptal sayıları

Yukarıda **Şekil 3.10**'deki dağılım grafiği incelendiğinde **Şekil 3.9'da** olduğu gibi 7 ve üzeri fatura sayısı olduğu görülmektedir. Bu durumun aykırı veri temizleme işlemi sonrası düzelmesi beklenmektedir.



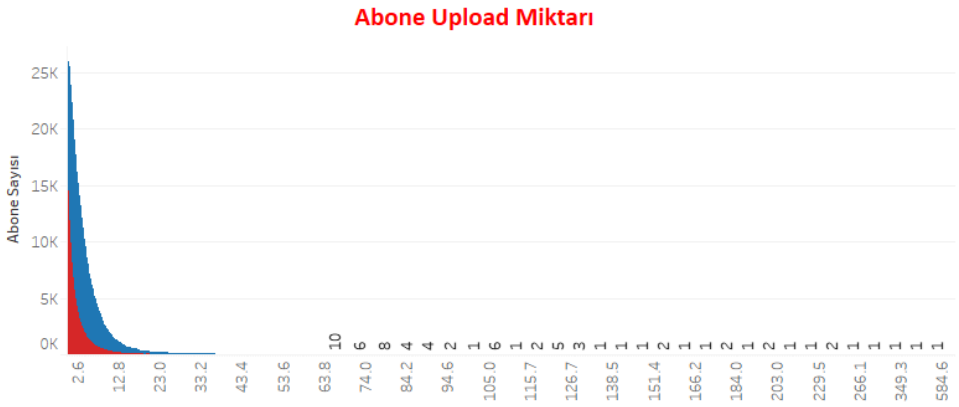
Şekil 3.11 Abonenin son 6 ayda yaptığı kota aşım sayısına bilgisine göre abone iptal sayıları

Yukarıda **Şekil 3.11**'deki kota aşım sayısına ait dağılım grafiği incelendiğinde verinin büyük kısmının 0 değerinde toplandığı görülmektedir. Bu durumda aslında bu bilginin sınıf etiketi ile ilişkisininde az olduğu anlamına gelmektedir.



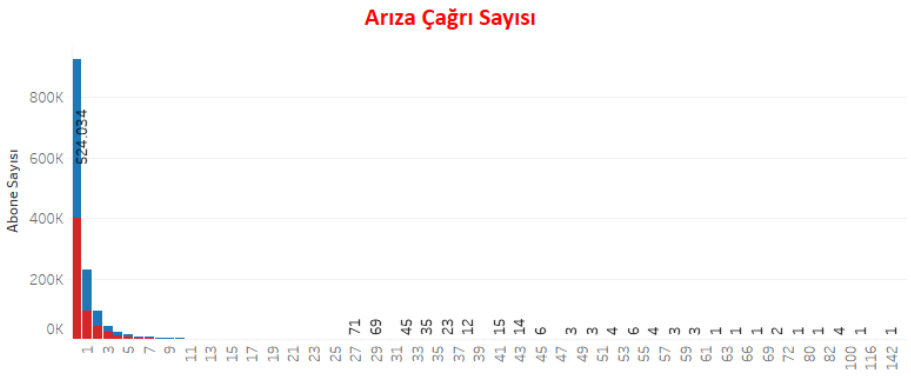
Şekil 3.12 Abonenin son 6 ayda yaptığı ortalama download miktarı bilgisine göre abone iptal sayıları

Yukarıda **Şekil 3.12**'deki aboneye ait download miktarına ait dağılım grafiği incelendiğinde abone download miktarı arttıkça iptal sayısının düştüğü gözlemlenebilmektedir. Çok yüksek miktarlarda downloadu olan abonelere ait veriler dağılım düzenini bozmaktadır. Bu dağılımın aykırı veri temizleme aşamasından sonra düzelmesi beklenmektedir.



Şekil 3.13 Abonenin son 6 ayda yaptığı ortalama upload miktarı bilgisine göre abone iptal sayıları

Yukarıda **Şekil 3.13**'deki aboneye ait upload miktarına ait dağılım grafiği incelendiğinde abone upload miktarı arttıkça iptal sayısının düştüğü gözlemlenebilmektedir. Çok yüksek miktarlarda uploadu olan abonelere ait veriler dağılım düzenini bozmaktadır. Bu dağılımın aykırı veri temizleme aşamasından sonra düzelmesi beklenmektedir.



Şekil 3.14 Abonenin son 6 ayda açtığı çağrı sayısı bilgisine göre abone iptal sayıları

Yukarıda **Şekil 3.14**'deki abonenin çağrı merkezine bıraktığı çağrı sayısı ile abonenin iptal olma durumlarının dağılım grafiği incelendiğinde çok yüksek miktarda aykırı veri olduğu görülebilmektedir.

4. YÖNTEM

Çalışmamızda hizmetini iptal etme olasılığı yüksek olan aboneleri bulmak için makine öğrenmesi tekniklerinden olan karar ağaçları, random forest, xgboosting, naif bayes ve lojistik regresyon yöntemleri kullanılmıştır.

Kullandığımız bütün veriler açık kaynak bir veritabanı sistemi olan PostgreSQL veritabanından SQL script dili ile hazırlanmış ve nihai veri setleri oluşturularak csv formatında export edilmiştir. Kullanılan PostgreSQL veritabanı sürümü 11 olarak belirlenmiştir.

Çalışmamızda karar ağaçları ve random forest algoritmalarını uygulamak ve gerekli olan veri taşıma, oluşturma ve temizleme gibi işlemleri yapmak için kullanımı ücretsiz olan KNIME analytics uygulaması kullanılmıştır.

4.1 Makine Öğrenmesi

Makina öğrenmesi yapay zekanın branşlarındandır. Yapısal olarak öğrenebilen ve büyük veriler üzerinde anlamlı çıkarımlar yapabilen bilgisayar algoritmalarına verilen isimdir. Günümüzde verilerin bilgisayarlar olmadan işlenemeyecek boyutlara gelmiş olması, klasik algoritmalar ile de verilerin sınıflandırılması ve ileriye dönük tahminlerin yapılmasının mümkün olmaması sebebiyle makina öğrenmesi konusu önemli hale gelmiştir ve bu konuda birçok araştırma ve geliştirme çalışmaları yapılmaktadır. Makina öğrenme teknikleri sınıflandırma problemlerinde başarılı bir şekilde kullanılmaktadır. Bu çalışmada da sınıflandırma amaçlı olarak makine öğrenmesi algoritmaları kullanılmıştır.

Makina öğrenmesi temel olarak gözetimli öğrenme ve gözetimsiz öğrenme olarak iki farklı kategoride değerlendirilmektedir. Gözetimli öğrenmede veri kümesinin sınıf niteliği vardır ve algoritmanın görevi bu sınıf niteliği tahmin etmektir. Gözetimsiz öğrenmede ise veri setinde sınıf niteliği yoktur ve algoritma veri setindeki benzer örnekleri bulmak ve bunları gruplamaktadır (Başarslan 2017).

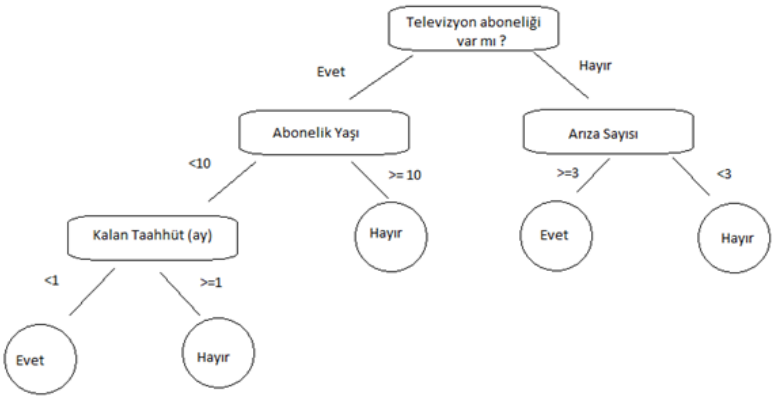
Bu çalışmada kullandığımız veri setinde sınıf etiketi olarak abonenin durumunu gösteren “churn oldumu” etiketi kullanılmış ve gözetimli öğrenme algoritmalarından olan karar ağaçları ile çalışılmıştır ve karar ağaçlarının birleşmesi ile oluşturulan random forest algoritması da çalışmamızda kullanılmıştır. Bu iki yönteme ek olarak xgboosting, naif bayes ve lojistik regresyon yöntemleri de kullanılarak sonuçların karşılaştırılması sağlanmıştır.

4.2 Karar Ağaçları

Karar ağaçları, tümevarım metodunu kullanarak verilerin sınıflandırılmasını veya sonuç tahmini yapılmasını sağlayan veri madenciliği ve makine öğrenmesi yöntemidir. Temel olarak yaprak ve dal olmak üzere iki özellik barındırır. Yapraklar; verimizde bulunan ve sınıf etiketi olarak adlandırılan her bir kolonu barındırır. Dallar ise; her bir sınıf etiketi arasındaki geçişi sağlayan ve özellik olarak adlandırılan değerlere göre şekillenirler.

Ağaç, bir kök düğümden, bir iç düğüm setinden (bölünmeler) ve bir dizi terminal düğümden (yapraklar) oluşur. Bir karar ağacındaki her düğümde bir ana düğüm ve iki veya daha fazla soydan düğüm bulunur. Bu çerçevede, bir veri seti, ağaç tarafından tanımlanan karar çerçevesine göre sırayla bölünerek sınıflandırılır ve gözlemin düştüğü yaprak düğümüne göre her gözlem için bir sınıf etiketi atanır.

Bu çalışmada kullanılan verimiz içinde alınan bir alt küme için örnek karar ağacı **şekil 4.1** de verilmiştir.



Şekil 4.1 Abone iptal durumu örnek karar ağacı

Yukarıdaki **şekil 4.1**'deki örnekte de görebileceğiniz gibi, karar ağaçları, temelinde karar vermemize yardımcı olan birer akış diyagramıdır. Gözle bakıldığında dahi insanın anlaması ve yorumlaması çok kolaydır. Veri setindeki nitelik sayısı çok fazla olduğu durumlarda oluşan ağaç çok büyük olacağı için anlaması ve yorumlaması zorlaşabilmektedir.

Bir karar ağacı, karar sürecinin kronolojik bir temsilidir. Ağacın kökünü oluşturan hücre, günümüze karşılık gelen bir düğümdür. Ağaç, bu düğümden geleceğe, kararların alınması gereken zamanları veya doğa durumlarını temsil eden bir düğüm ağı ve doğanın olası kararlarını veya durumlarını temsil eden dalları kullanılarak geleceğe doğru oluşturulur (Lawrence ve Pasternack 2002)

Karar ağacı kök ve dallardan ve düğümlerden oluşur. Kök ağacın başlangıç noktasıdır ve vereceğimiz karar veya bulacağımız sınıfa karşılık gelir. Ağaç oluşurken her düğümden çıkan dallar gidilebilecek alternatiflere denk gelir. Her bir dalın kendine ait bir maliyeti veya getirisi vardır. (Ulucan 2007)

Karar ağaçları, maksimum olasılık sınıflandırması gibi uzaktan algılamada kullanılan geleneksel denetimli sınıflandırma prosedürlerine göre birkaç avantaja sahiptir. Özellikle, karar ağaçları kesinlikle parametrik değildir ve girilen verilerin dağıtımına ilişkin varsayımlar gerektirmez. Ayrıca, özellikler ve sınıflar arasındaki doğrusal

olmayan ilişkileri ele alırlar. Eksik değerlere izin verirler ve hem sayısal hem de kategorik girdileri ele alabilirler. Karar ağaçları belirgin bir sezgiselliğe sahiptirler çünkü sınıflandırma yapısı açıktır ve bu nedenle kolayca yorumlanabilirler.

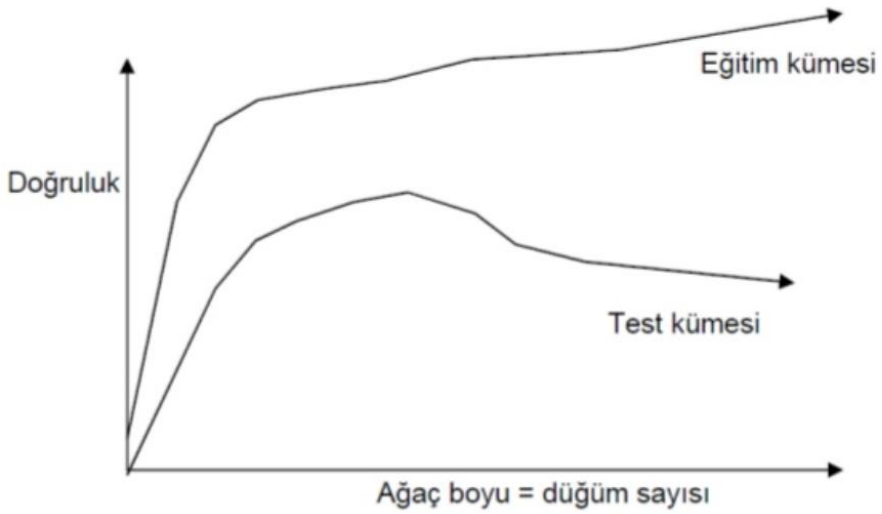
Karar ağaçları gözetimli (supervised) öğrenme algoritmalarındandır. Karar ağacı oluşturmada en popüler algoritma C4.5 algoritmasıdır.

4.2.1 Avantajları

- Karar ağacı oluşturmak diğer modellere göre daha kolaydır.
- İnsan tarafından anlaşılması ve yorumlanması mümkündür.
- Hem sayısal hem de kategorik verileri işleyebilir.
- Çok çıktılı problemleri ele alabilir.
- Veri kümesindeki nitelik sayısı ile karmaşıklık sayısı arasında logaritmik ilişki vardır.

4.2.2 Dezavantajları

- Sürekli nitelik değerleri tahmin etmekte çok başarılı değildir.
- Sınıf sayısı fazla ve öğrenme kümesi örnekleri sayısı az olduğunda model oluşturma başarısız olur.
- Büyük öğrenme kümeleri için ağaç oluşturma karmaşıklığı fazladır.
- Veri setinden kaynaklı aşırı öğrenme (over fitting) durumu sıkça yaşanabilir. Buna engel olmak için budama yapılması gerekir. Aşırı öğrenmenin doğruluk oranına etkisi **şekil 4.2** de görülebilir.



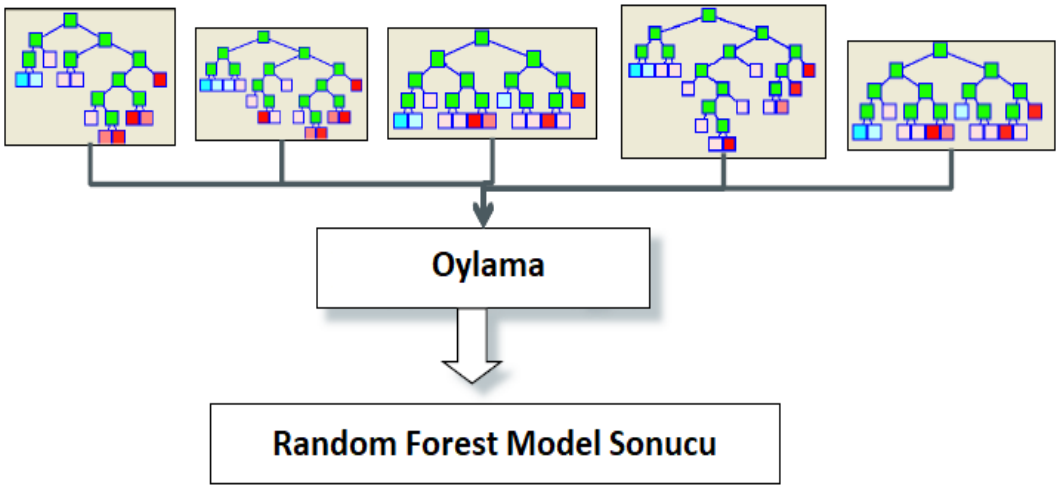
Şekil 4.2 Aşırı öğrenme durumunda sonucun değişimi

4.3 Random Forest

Karar ağaçlarının en büyük problemlerinden biri aşırı öğrenme yani veriyi ezberlemesidir. Bu durum makine öğrenmesinde over fitting olarak adlandırılır. Random forest veri setimizde bulunan öz niteliklerden rastgele alarak farklı veri seti kümeleri ile 10'larca hatta 100'lerce karar ağacı oluşturur ve bunları birleştirerek bir model kurar. Yani 100'lerce farklı budanmış karar ağaçlarını farklı veri alt setleri ile oluştur. Modelde yer alan her bir karar ağacı bağımsız olarak oluşturulur. Her bir karar ağacı sınıflandırma ve tahmin işlemini yine bağımsız yapar. Tüm karar ağaçlarının yaptığı tahminler bir oylama mekanizmasına tabi tutularak, en yüksek oy alan değer, modelin sonucu olarak verilir.

Birden çok karar ağacı ile oylama yapılarak oluşturulan sonuç, karar ağaçlarında gördüğümüz en büyük problem olan aşırı öğrenmenin önüne geçmiş olur. Ayrıca her bir karar ağacında kullandığımız farklı veri setlerinden dolayı da aykırı veri (outlier) problemi minimum seviyeye inmiş olur. Random forest yöntemi bir bagging yöntemidir. Bagging yönteminde birden fazla yöntem paralel olarak farklı veri seti kümeleri ile eğitilmekte ve tüm modellerin oluşturduğu sonuç oylamaya tabi tutularak nihai sonuç ortaya çıkarılmaktadır. Bizim çalışmamızda da karar ağaçlarının yanı sıra

bir bagging yöntemi olan random forest modeli de kullanılmıştır. Basit bir random forest modeli **şekil 4.3** de verilmiştir.



Şekil 4.3 Örnek bir random forest modeli

4.4 Naif Bayes

Bayes teoremine dayanan bir sınıflandırma tekniğidir. Sınıflandırma yaparken tüm sınıf etiketlerini bağımsız olarak değerlendirilir. Örneğin bir meyvenin erik olma ihtimalini rengi, şekli ve boyutlarının her birini ayrı ayrı değerlendirerek tahmin etmeye çalışır. Renk yeşil ise +1, şekil yuvarlak ise +1 ve boyut 2 cm'den küçük ise +1 şeklinde bağımsız değerlendirme yapar.

Naif bayes yöntemi uygulanırken bulmak istediğimiz her sınıfın bağımsız olarak gerçekleşme olasılığını ve veride verilen tüm niteliklerin değerlerine ait iki sınıftan birinde olma olasılıkları birbirinden bağımsız hesaplamamız gerekmektedir. Bunu yaparken aşağıdaki **denklem 4.1** ile verilen formülü kullanılmaktadır.

$$P(\mathbf{X} \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) = P(x_1 \mid C_i) \times P(x_2 \mid C_i) \times \ldots \times P(x_n \mid C_i) \qquad (4.1)$$

4.4.1 Avantajları

- Gerçekleştirmesi kolaydır.
- Çoğu durumda iyi sonuçlar üretir.

4.4.2 Dezavantajları

- Sınıf bilgisi verildiğinde nitelikler bağımsızdır.
- Gerçek hayatta değişkenler birbirine bağlıdır.
- Değişkenler arası ilişki modellenemez.

4.5 Lojistik Regresyon

Lojistik regresyon iki olası sonucu olan bir olayı incelememizi sağlayan istatistiksel bir yöntemdir. Lojistik regresyon yönteminde bir veya daha fazla bağımsız değişken kullanılabilir.

Lojistik regresyon modeli 1845'li yıllarda nüfus artışı için yapılan matematiksel çalışmalar sırasında ortaya çıkmıştır (Gürcan, 1998). Lojistik regresyon analizi terimi, bağımlı değişkene uygulanan logit dönüşümünden gelir. Bu durum aynı zamanda hem tahminde hem de yorumlamada bazı farklılıklara neden olmaktadır (Hair vd. 2006).

Lojistik regresyon yönteminde bilinmesi gereken bazı temel terimler vardır. Bunlar odds, odds ratio ve lojittir.

- **Odds:** Başarı ya da görülme olasılığının "p", başarısızlık veya görülmeme olasılığına "1-p" oranıdır. Yani p / (1-p).
- **Odds Ratio** : iki odds'un birbirine olasılığıdır yani iki değişken arasındaki ilişkiyi vermektedir.
- **Lojit:** Odds ratio'nun doğal logoritmasıdır. Bu sayede asimetrik olan odds ratio simetrik hale dönüştürülmüş olur.

Bizim yaptığımız çalışmada bir sınıflandırma çalışması olduğu ve sonuç olarak churn olacak ya da olmayacak diye iki sınıftan oluştuğu için lojistik regresyon modelinin de kullanılmasına karar verilmiştir.

4.6 XGBoosting Algoritması

Gradient boosting algoritmasının üzerine kurulu bir algoritmadır. Random forestdan farklı olarak bagging yerine boosting yöntemini kullanır. Boosting yöntemi birden fazla zayıf öğreticiyi sıralı bir şekilde kullanarak birbirlerinin hatalarından öğrenmeyi temel almaktadır (Anonymous 2019). Bagging de ise bu işlem sıralı değil bağımsız olarak yapılır ve her model birbirinden habersiz olarak eğitilir.

XGBoosting yönteminde eğitilen her model için tüm veri seti kullanılırken, random forest yönteminde her model farklı bir veri seti alt kümesi ile eğitilmektedir. İki yöntemin arasındaki farklar aşağıdaki **çizelge 4.1**'de verilmiştir (Yüceoğlu 2018).

Çizelge 4.1 Bagging ve boosting yöntemleri arasındaki farklar

	Boosting	**Bagging**
Çalışma yöntemi	İteratif	Paralel
Sonuçların değerlendirlmesi	Ağırlıklı Ortalama	Ortalama (voting)
Her model veri seti kullanımı	Bütün veri kümesi	Rassal örneklem
Aşırı öğrenme dayanıklılığı	Zayıf	Güçlü
Öğrenme hızı	Yavaş	Hızlı
Diğer avantaj	Yanlılık (bias) azaltma	Varyans azaltma

4.7 Eksik Veriler (Missing Values)

Makine öğrenmesi algoritmalarında kullandığımız veri setlerinde bulunan boş kayıtlar modelin oluşturulması, eğitilmesi ve tahminleme çalışmalarında tutarsız sonuçların oluşmasına sebep olmaktadır. Bu yüzden eksik olan verilerin tespit edilip belirli kurallara göre işlenmesi gerekmektedir. Genel olarak eksik olan verilere aşağıdaki adımlar uygulanır.

- Veri setindeki satır veya kolonun silinmesi. Eğer bir kolon yani nitelik belirli bir yüzde üzerinde eksik (missing) değer içeriyor ise bu niteliği veri setinden komple çıkarmamız gerekir ya da eğer veri setindeki bir satır birçok nitelik için eksik veri içeriyor ise bu satırı veri setinden çıkarmamız gerekir.
- Sabit bir değer ile doldurulması.
- İlgili kolonda yer alan diğer satırlardaki verilerin ortalama veya ortanca değerleri ile doldurulması.
- Süreklilik arz eden sayısal değerleri içeren bir niteliğin kategorik bir nitelik haline getirilmesi. Çok geniş aralıkta sayısal ifadeler yer alan bir niteliğin belirli aralıklara göre kategorik hale getirilmesidir. '0'dan küçük', '0-10' , '10-50', '50-100', '100'den büyükler' şeklinde düşünülebilir.
- Eksik olan veriyi yüksek ilişkili (high correlation) olduğu başka bir niteliğin değerine göre tahmin etmek.

4.8 Sonuçlarının Değerlendirilmesi

Çalışmamızda kullanacağımız modellerin üreteceği sonuçları değerlendirmek için yaygın olarak kullanılan birtakım ölçütler kullanılmıştır. Bu ölçütler aşağıda verilmiştir ayrıca bu ölçütlerin hesaplamasında karışıklık matrisi (confusion matrix) kullanılmaktadır. Bu matriste satırlarda yer alan değerler veri setimizdeki gerçek değerleri sütunlar ise modelimizin çalışması sonrası oluşan sınıflandırma / tahmin değerlerini içerir **(Çizelge 4.2).**

Çizelge 4.2 Karışıklık Matrisi

		Tahmin Edilen Sınıf (Predicted Class)	
		Sınıf = 1	Sınıf = 0
Gerçek Sınıf Değeri (Actual Class)	Sınıf = 1	True positive (TP)	False negative (FN)
	Sınıf = 0	False positive (FP)	True negative (TN)

4.8.1 Doğruluk (Accuracy)

Model çalışması sonucunda doğru olarak sınıflandırılmış / tahmin edilmiş tüm örnek sayılarının veri setindeki tüm örnek sayısına oranıdır **(Denklem 4.2)** .

$$\text{Doğruluk} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (4.2)$$

4.8.2 Hata oranı (Error Rate)

Model çalışması sonucunda yanlış olarak sınıflandırılmış / tahmin edilmiş tüm örnek sayılarının veri setindeki tüm örnek sayısına oranıdır **(Denklem 4.3)**.

$$\text{Hata Oranı} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (4.3)$$

4.8.3 Duyarlılık (Sensitivity)

Model çalışması sonucunda doğru olarak sınıflandırılmış / tahmin edilmiş pozitif örnek sayılarının pozitif tüm örnek sayısına oranıdır **(Denklem 4.4).**

$$\text{Duyarlılık} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4.4)$$

4.8.4 Kesinlik (Precision)

Model çalışması sonucunda doğru olarak sınıflandırılmış / tahmin edilmiş pozitif örnek sayılarının pozitif sınıflandırılmış tüm örnek sayısına oranıdır **(Denklem 4.5)**.

$$\text{Kesinlik} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4.5)$$

4.8.5 F-Ölçütü (F-Measure)

Model çalışması sonucunun değerlendirilmesi için duyarlılık ve kesinlik değerleri tek başına anlam ifade etmez bu yüzden bu iki değerin harmonin ortalamalarını alınarak f-ölçütü bulunur **(Denklem 4.6)**.

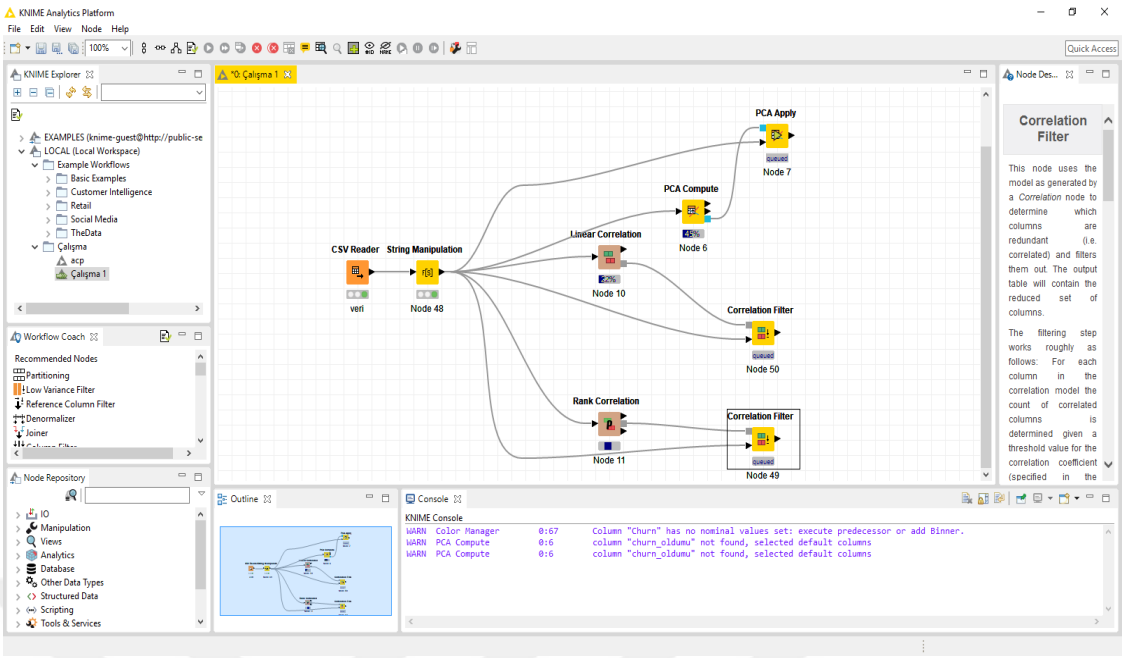
$$\text{F} - \text{Ölçütü} = \frac{2 \text{ x Kesinlik x Duyarlılık}}{\text{Kesinlik} + \text{Duyarlılık}} \quad (4.6)$$

4.9 Knime Analytics

KNIME Analytics Platform veri bilimi uygulamaları ve hizmetleri oluşturmak için açık kaynaklı bir yazılımdır. KNIME, verileri anlamayı ve veri bilimi iş akışlarını ve yeniden kullanılabilir bileşenleri içerir. Bu bileşenler sayesin de farklı kaynaklardan veri okuma ve okunan veri üzerinden temizleme, değiştirme, birleştirme ve bunun gibi birçok işlemi yapabilme imkânı sunar. Ayrıca okunan veri üzerinden birçok makine öğrenmesi algoritmasının çalıştırılabilmesi ve sonuçlarını istenilen bir sisteme kayıt edilmesine de olanak sağlar.

Daha önce yapılan çoğu makine öğrenmesi çalışmasında WEKA uygulaması kullanılmış olmasına rağmen biz çalışmamızda hem KNIME uygulamasının sonuçlarını karşılaştırmak için hemde akademik çalışmalar dışında birçok özel sektörde kullanım alanı bulan KNIME ile çalışmayı tercih ettik. KNIME uygulaması WEKA tarzı uygulamalara göre daha kullanıcı dostu bir arayüz sunmakta ve veri ile ilgili çok fazla sayıda bileşen içermektedir. Profesyonel bir ETL ürününde bulabileceğiniz tüm özellikler KNIME içerinde olduğu gibi veri madenciliği ve makine öğrenmesi konularında ihtiyacınız olan tüm bileşenlerede sahip bir uygulamadır.

Bu çalışmamızdaki temel amaç algoritmaların matematiksel ve programsal içeriği değil, kullanımları sonucu oluşan sonuçların incelenmesi olduğu ve büyük veriler ile çalışmamız gerektiği için çalışmamızda KNIME uygulaması ve içerisindeki hazır bileşenler kullanılmıştır. KNIME uygulamasına ait ekran görüntüsü **şekil 4.4** de verilmiştir.



Şekil 4.4 Knime analytics örnek ekran görüntüsü

Çalışmada alınan sonuçların ve sürelerin referans olarak alınabilmesi için modelin eğitildiği bilgisayar ve uygulama özellikleri aşağıda verilmiştir.

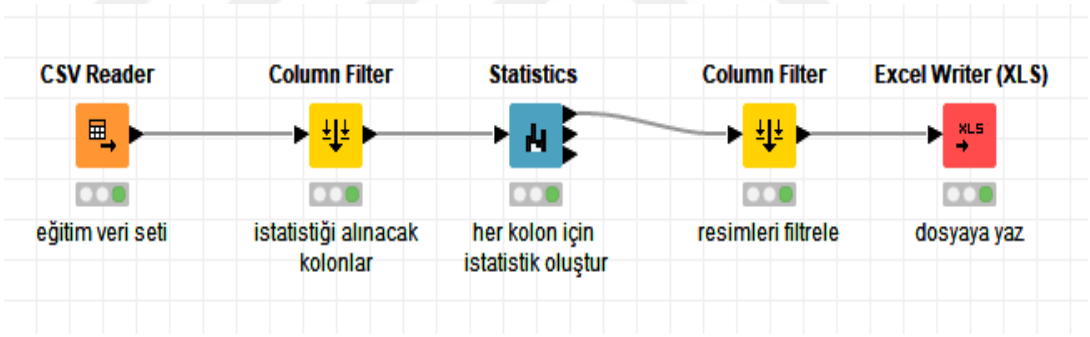
- **İşlemci**: AMD Opteron Processor 6174 2.2Ghz (4 core)
- **RAM**: 16 GB
- **İşletim sistemi**: Windows Server 2008 R2 64 bit
- **Knime sürüm**: 3.7.1
- **Sanallaştırma**: VMWare

5. UYGULAMA

Bu bölümde veri seti üzerinde yapılan işlemlerden bahsedilmiş ve veri setini oluşturan niteliklerin her biri hakkında istatistiksel bilgiler verilmiştir. Bu çalışmada kullandığımız karar ağaçları, random forest, naif bayes, lojistik regresyon ve xgboosting algoritmaları ile model oluşturarak bu modellerin test edilmesi KNIME uygulaması üzerinde yapılmış ve sonuçları paylaşılmıştır.

5.1 Veri Seti İşlemleri

Yöntem bölümünde başlıkların verildiği veri seti üzerinde ilk olarak eksik verilerin analizi ve bunların sabit değerler ile değiştirilmesi işlemi yapılacaktır. Verilerin veri tabanından alınıp CSV olarak kaydedilmesi ve eksik verilerin değiştirilmesi işlemi KNIME üzerinde aşağıda anlatılacağı gibi yapılmıştır. Veriler üzerinden ilk olarak istatistiksel bilgileri çıkarıldığı çalışma KNIME üzerinden **şekil 5.1**'deki paket ile yapılmış ve eksik verilerin sayısı da bu çalışmada **çizelge 5.1**'de verilmiştir.



Şekil 5.1 Knime analytics üzerinden tüm veriye ait istatistiksel bilgileri oluşturan paket

Yukarıda **şekil 5.1**'de verilen paketin çalışması sonrası oluşan istatistiksel bilgilerin tutulduğu tablo csv olarak kaydedilmiş ve bu csv dosya içerindeki verileri **çizelge 5.1'**de verilmiştir. İstatistiksel bilgileri içerisinde bulunan histogram diyagramların csv dosyaya atılmaması için resim içerikler filtrelenmiştir.

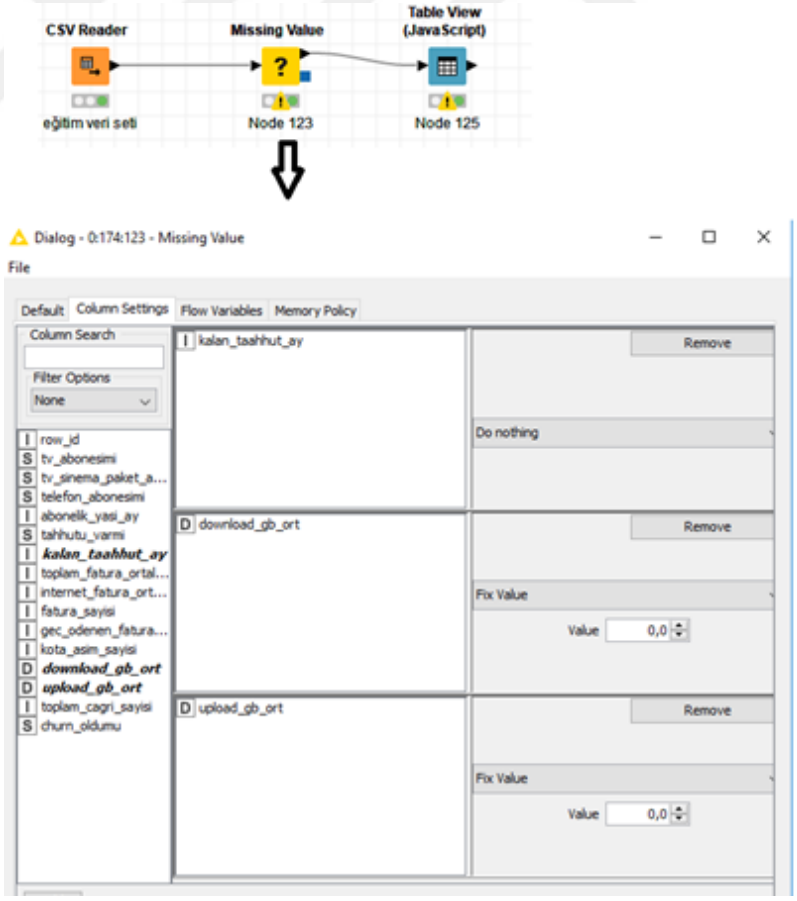
Çizelge 5.1 Eğitim verisi istatistiksel bilgileri

Kolon	En Düşün	En Yüksek	Ortalama	Standart Sapma	Varyans	Çarpıklık	Eksik Bilgi	Ortanca	Satır Sayısı
Abonelik Yaşı Ay	-6	154	31,38	24,91	620,67	1,20	0	25	1364037
Kalan Taahhüt Ay	0	37	10,31	8,19	67,10	0,12	448466	11	1364037
Toplam Fatura Ortalaması	0	21057	61,03	85,57	7322,95	59,53	0	47	1364037
İnternet Fatura Ortalaması	0	406	19,13	13,68	187,05	8,26	0	19	1364037
Fatura Sayısı	0	36	6,08	1,34	1,80	-2,00	0	6	1364037
Geç Ödenen Fatura Sayısı	0	15	1,97	2,35	5,51	0,82	0	1	1364037
Kota Aşım Sayısı	0	9	0,20	0,98	0,95	5,39	0	0	1364037
Download GB	0	12061,7	46,22	68,35	4671,54	25,08	9416	30,2	1364037
Upload GB	0	2049,1	4,36	10,86	117,97	32,12	9416	2,3	1364037
Toplam Çağrı Sayısı	0	2313	0,77	2,80	7,82	417,11	0	0	1364037

Eğitim verisi üzerinde herhangi bir işlem yapmadan önceki istatistiksel bilgiler yukarıdaki gibi oluşmuştur. Dikkat edilirse kalan “taahhut "ay”, “download gb ort” ve “upload gb ort” kolonlarında eksik verilerin olduğu ayrıca “toplam fatura ortalamasi”, “download gb ort”, “upload gb ort” ve “toplam cagri sayisi” kolonlarında max değerinin standart sapmadan çok yüksek olduğu gözlemlenebilmektedir. Bunlar modellerimizi oluşturma aşamasında veri setimiz içerisinde bulunmasını istemediğimiz verilerdir ve bu sorunlu veriler düzeltilmeden model eğitme aşamasına geçmek sağlıksız bir modelin oluşmasına sebep olacaktır. Yöntem ve materyal bölümlerinde de bahsettiğimiz eksik veri temizleme ve aykırı(outliers) verilerin tespit edilmesi işlemleri bu bölümde yapılmış ve sonuçları değerlendirilmiştir. Eksik verilerin ve aykırı verilerin temizlenmesi sonrası veri setimiz daha sağlıklı model oluşturmak için hazır olacaktır.

5.1.1 Eksik verilerin temizlenmesi

Bu bölümde yukarıda istatistiğini çıkardığımız veri setimizde bulunan eksik verileri değerlendirilmesi ve temizlenmesi işlemi yapılacaktır. Yukarıdaki istatistik bilgiler tablomuzda (**çizelge 5.1**) "kalan taahhut ay", "download gb ort" ve "upload gb ort" alanlarında eksik veri olduğunu söylemiştik. Bu alanlardan "kalan taahhüt ay" alanında eksik veri olması taahhüt vermeyen abonelerden dolayı normaldir bu yüzden taahhütü olmayan aboneleri yanlış değerlendirmemek için bu alan üzerinde bir eksik verileri doldurma veya temizleme ile ilgili işlem yapılmamıştır. Download ve upload bilgilerinin bulunduğu alanda eksik veri olması aslında abonenin hiçbir internet aktivitesi gerçekleştirmediğini göstermektedir bu yüzden bu alanlardaki eksik veriler 0 ile değiştirilmiştir. Bu işlemleri yapmak için KNIME analytics üzerinden bulunan "Missing Value" node'u kullanılmış ve KNIME paketi **şekil 5.2**'deki gibi oluşturulmuştur.



Şekil 5.2 Eksik verilerin temizlenmesi için knime analytics paket yapısı

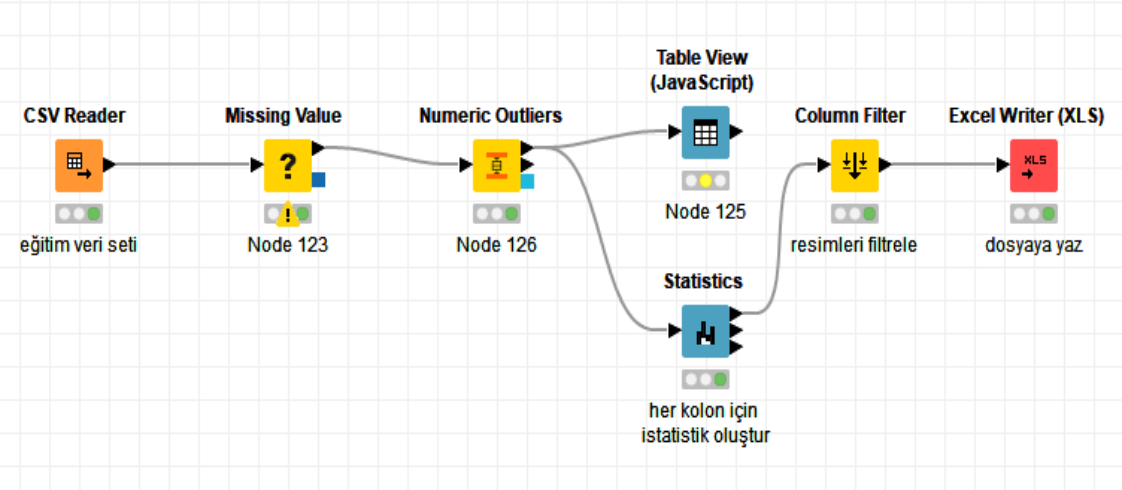
5.1.2 Aykırı (Outliers) verilerin tespiti ve temizlenmesi

Önceki bölümlerde de bahsettiğimiz aykırı verilerin varlığı veri setini kullanarak oluşturduğumuz modellerin doğru kararlar verememesine sebep olmaktadır. Bu yüzden aykırı veriler üzerinde temizleme işlemleri yapmak gerekir. Aykırı veriler üzerinde genel olarak aşağıdaki işlemler uygulanabilir.

- Aykırı değeri kendine en yakın kabul edilebilir değer ile değiştir.
- Eksik veri olarak doldur.
- Aykırı değer içeren tüm satırı veri setinden sil.

Bizim veri setimize ait istatistik incelendiğinde toplam "fatura ortalamasi", "download gb ort", "upload gb ort" ve "toplam cagri sayisi" kolonlarında max değerinin standart sapmadan çok yüksek olduğu gözlemlenmiştir. Bu kolonlar ve diğer tüm sayısal kolonlar üzerinde aykırı değerler tespit edilecek ve aykırı değer içeren bu satırların veri setinden silinmesi sağlanacaktır. Aykırı veri temizleme işlemi daha önce yapılan eksik veri temizleme işlemi ardından yapılmıştır. Modeli eğitme, doğrulama ve test etme aşamalarında bu ön işleme aşamalarından geçmiş olan veri seti kullanılmıştır.

Aykırı veri temizleme işlemi KNIME Analytics üzerindeki önceki paketimize (**şekil 5.2**) "Numeric Outliers" nodu eklenerek **şekil 5.3**'deki gibi oluşturulmuştur. Daha önceki adımda yapılan eksik verilerin doldurulması ve bu adımda yapılan aykırı veri içeren satırların silinmesi ardından tekrar tüm veri için istatistik tablosu oluşturulmuş ve **çizelge 5.3'**de verilmiştir.



Şekil 5.3 Eğitim setindeki aykırı (outliers) verilerin temizlenmesi

Aykırı veri temizleme aşamasında veri setinde bulunan bir niteliğe ait alt ve üst değer aralıkları çıkarılmaktadır. Alt ve üst limit değerlerinin dışında kalan değerlere sahip olanlar ise aykırı olarak değerlendirilmektedir. **Şekil 5.3**'deki paketin çalışması sonrası her bir nitelik için alt ve üst değerler oluşturulmuş ve **çizelge 5.2'**de verilmiştir.

Çizelge 5.2 Veri setine ait aykırı değerlerin alt ve üst aralık değerleri

Kolon Adı	Satır Sayısı	Aykırı Kayıt Sayısı	Alt Sınır	Üst Sınır
Abonelik Yaşı Ay	1.364.037	39.546	-34,5	89,5
Kalan Taahhüt Ay	915.571	0	-23	41
Toplam Fatura Ortalaması	1.364.037	132.955	5	93
İnternet Fatura Ortalaması	1.364.037	48.664	-2	38
Fatura Sayısı	1.364.037	119.751	4,5	8,5
Geç Ödenen Fatura Sayısı	1.364.037	16	-6	10
Kota Aşım Sayısı	1.364.037	71.270	0	0
Download GB	1.364.037	63.817	-74,75	146,45
Upload GB	1.364.037	85.087	-6,15	11,85
Toplam Çağrı Sayısı	1.364.037	114.612	-1,5	2,5

Veri setindeki alanlardan aykırı verilerin tespiti **çizelge 5.2'**deki değerlere göre yapılmaktadır. "Lower bound" değerinden küçük ve "upper bound" değerinden büyük değerlere sahip olan alanların bulunduğu satırlar tamamen silinmiştir, bu sayede veri setimizde bulunan niteliklerden hiç birisi aykırı veri içermeyecek hale getirilmiştir.

Veri setimize uygulanan eksik veri doldurulması ve aykırı veri içeren satırların silinmesi işlemleri sonrası oluşturulan istatistik tablosu **çizelge 5.3**'deki gibi olmuştur. Veri temizliği öncesi veri setine ait **çizelge 5.1'de** verilen istatistik bilgileri ile veri temizleme aşaması sonrası oluşan ve **çizelge 5.3'**de verilen bilgileri incelendiğinde temizlik işleminin istatistiksel bilgileri ne kadar değiştirdiği görülebilmektedir.

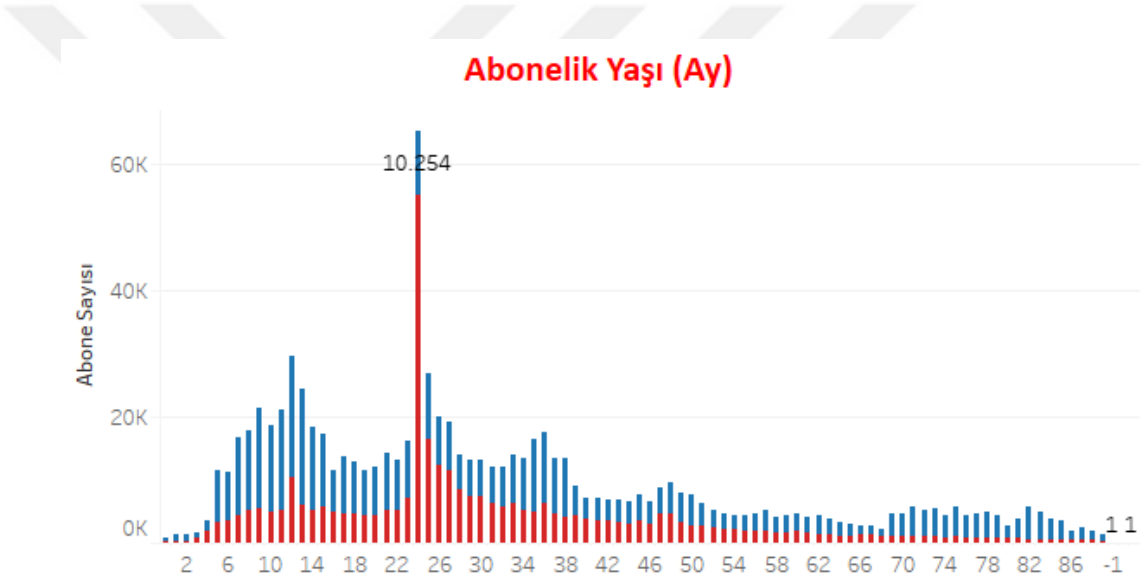
Çizelge 5.3 İşlemler sonrası oluşturulan istatistik tablosu

Kolon	En Düşün	En Yüksek	Ortalama	Standart Sapma	Varyans	Çarpıklık	Eksik Bilgi	Ortanca	Satır Sayısı
Abonelik Yaşı Ay	-6	89	32,32	21,08	444,52	0,90	0	26	857991
Kalan Taahhüt Ay	0	35	10,18	7,56	57,13	0,12	294226	11	857991
Toplam Fatura Ortalaması	5	93	45,65	13,77	189,60	0,60	0	44	857991
İnternet Fatura Ortalaması	0	38	18,43	6,71	44,96	-0,25	0	20	857991
Fatura Sayısı	5	8	6,36	0,56	0,31	0,58	0	6	857991
Geç Ödenen Fatura Sayısı	0	8	2,11	2,41	5,80	0,67	0	1	857991
Kota Aşım Sayısı	0	0	0,00	0,00	0,00	0,00	0	0	857991
Download GB	0	146,4	37,31	33,54	1124,81	0,98	0	29	857991
Upload GB	0	11,8	2,90	2,67	7,14	1,06	0	2,2	857991
Toplam Çağrı Sayısı	0	2	0,32	0,60	0,36	1,69	0	0	857991

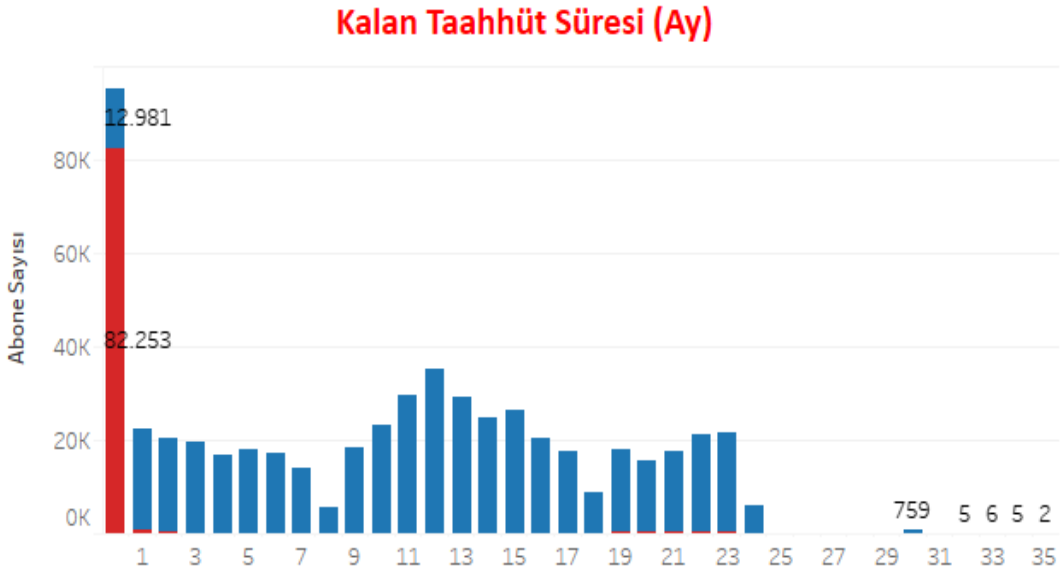
Yukarıdaki **çizelge 5.3** ile verilen tablo incelendiğinde bizim izin verdiğimiz "kalan taahhüt ay" alanı dışında eksik veri kalmadığı gözlenmiş ayrıca min ve max değerlerinin standart sapmaya göre oldukça ideal değerler aldığı gözlemlenmiştir. Aykırı veri

temizliği öncesi 1.364.037 olan kayıt sayısı bu işlemde veri setinden atılan satırlar sonrası 857.991 adet olmuştur. Veri setimiz %38 oranında küçülmüş olsa da yeni veri setimiz daha sağlıklı bir model oluşturabilmemizi sağlayacaktır.

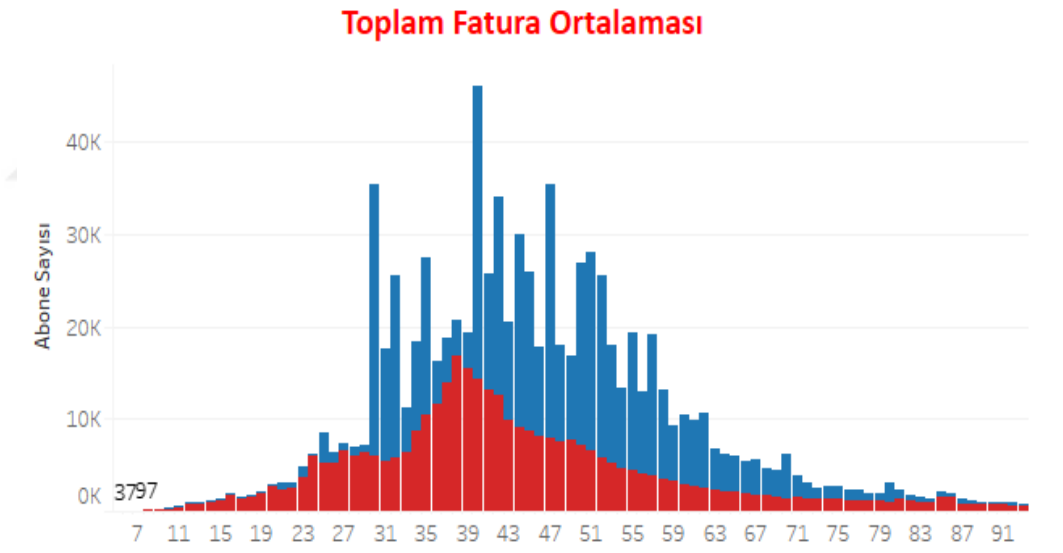
Eksik verilerin doldurulması ve aykırı verileri içeren satırların silinmesi sonrası veri setimize ait dağılım grafiği **şekil 5.4-5.13**'de görüldüğü gibi değişmiştir. Yaptığımız temizlik çalışmalarından değer tipi olarak sadece sayısal değere sahip olan alanlar etkilenmiştir. "tv abonesimi" gibi boolean değere sahip alanlar etkilenmemiş bu yüzden dağılım grafiğinden önemli bir değişme olmamıştır. Bu yüzden **şekil 5.4-5.13**'de sadece sayısal değer içeren alanlara ait dağılım listelerinin yeni hali verilmiştir.



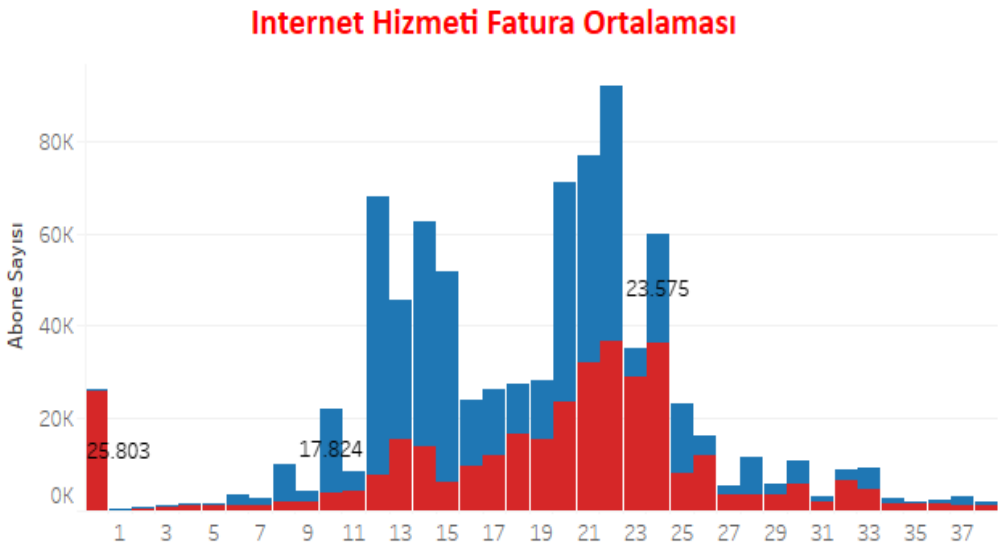
Şekil 5.4 Veri işleme sonrası abonelik yaşı bilgisine göre abone iptal sayılar yeni grafik



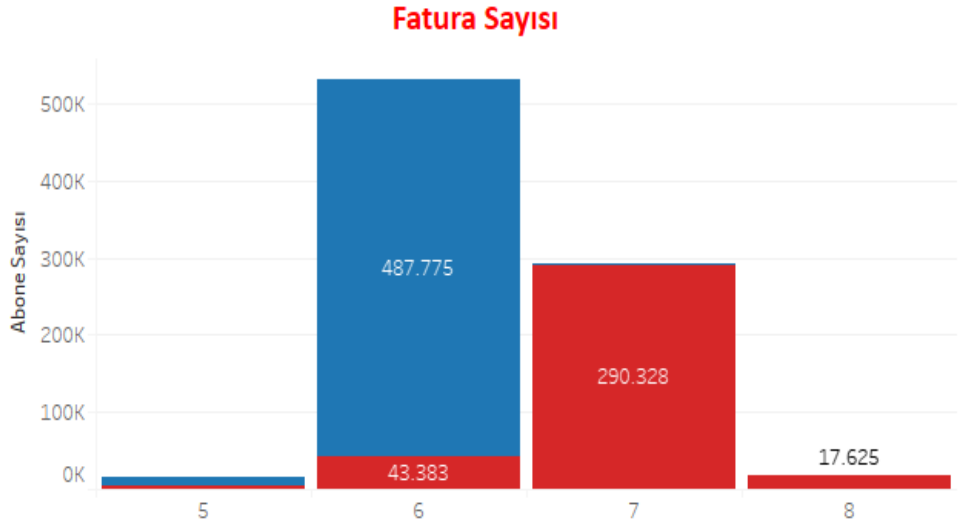
Şekil 5.5 Veri işleme sonrası taahhüt veren abonelerin kalan taahhüt bilgisine göre abone iptal sayıları



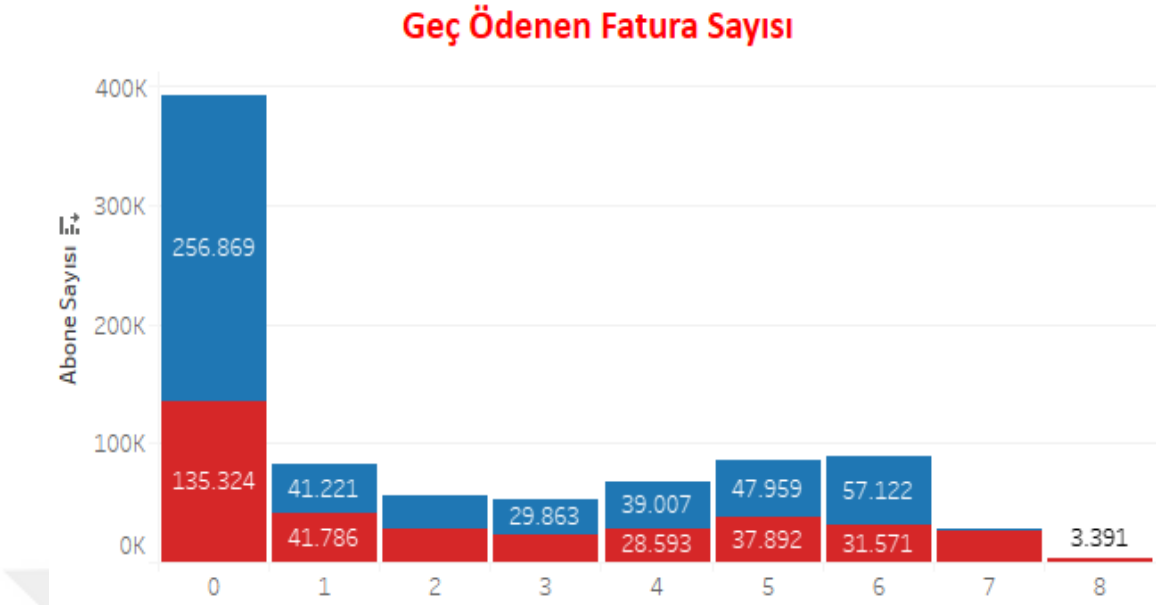
Şekil 5.6 Veri işleme sonrası abonenin tüm hizmetlerine ait son 6 ay fatura ortalama bilgisine göre abone iptal sayıları



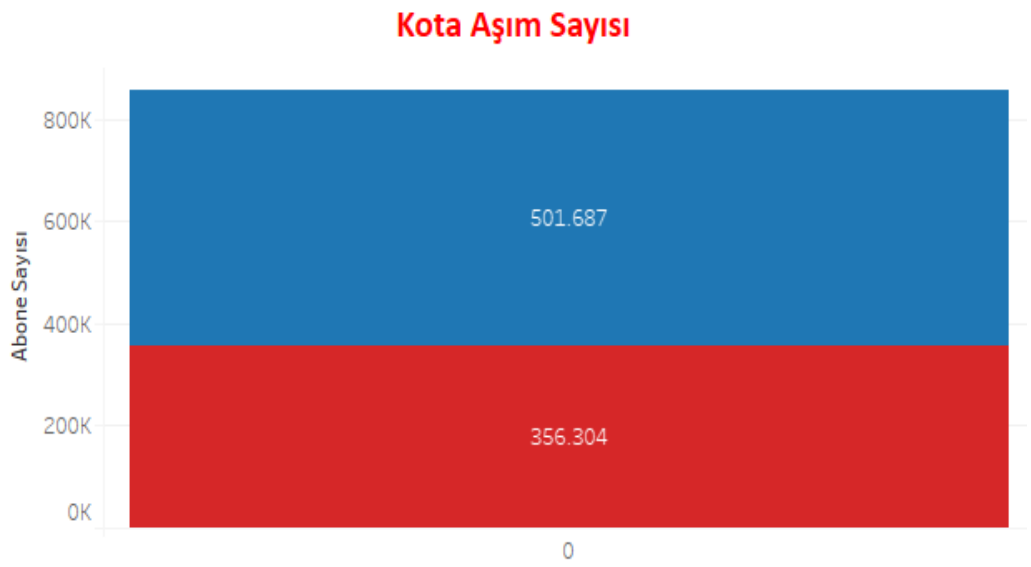
Şekil 5.7 Veri işleme sonrası abonenin internet hizmetine ait son 6 ay fatura ortalama bilgisine göre abone iptal sayıları



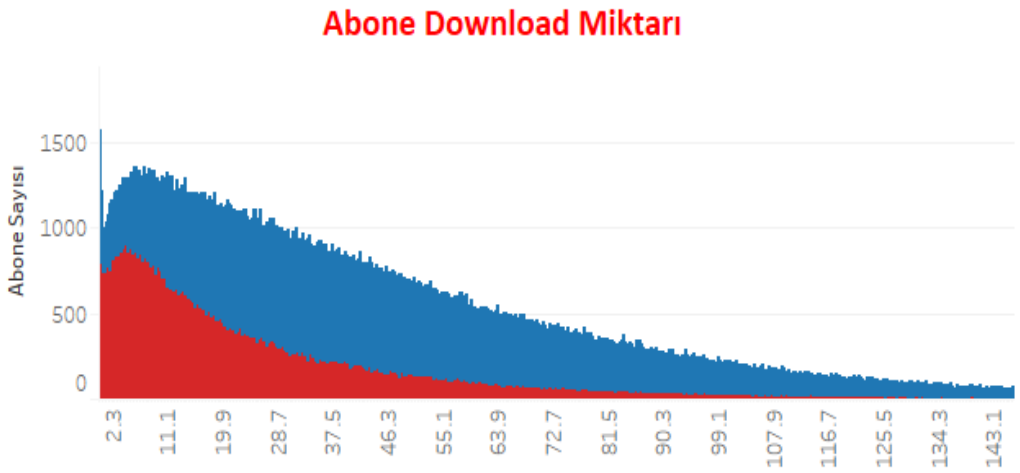
Şekil 5.8 Veri işleme sonrası aboneye son 6 ayda çıkan fatura sayısı bilgisine göre abone iptal sayıları



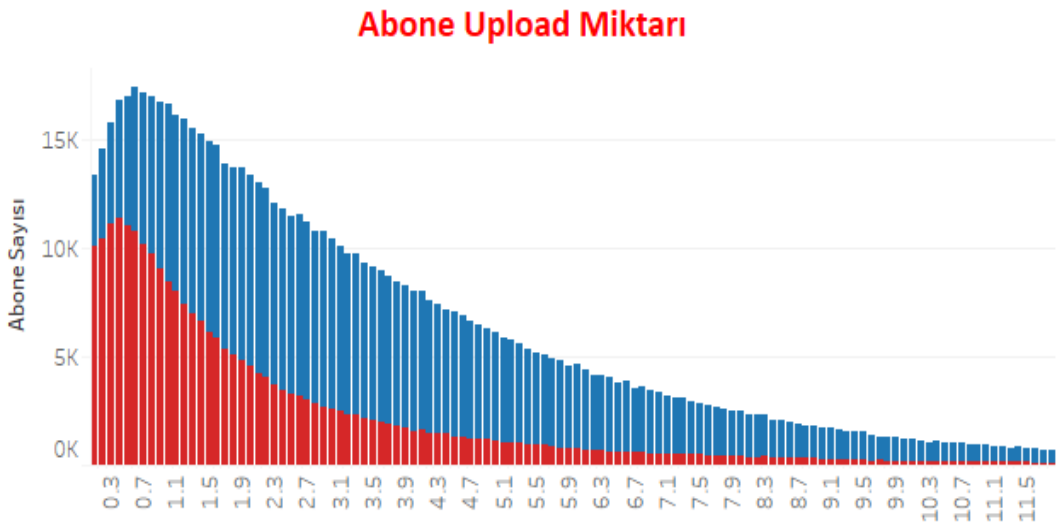
Şekil 5.9 Veri işleme sonrası abonenin son 6 ayda geç ödediği fatura sayısı bilgisine göre abone iptal sayıları



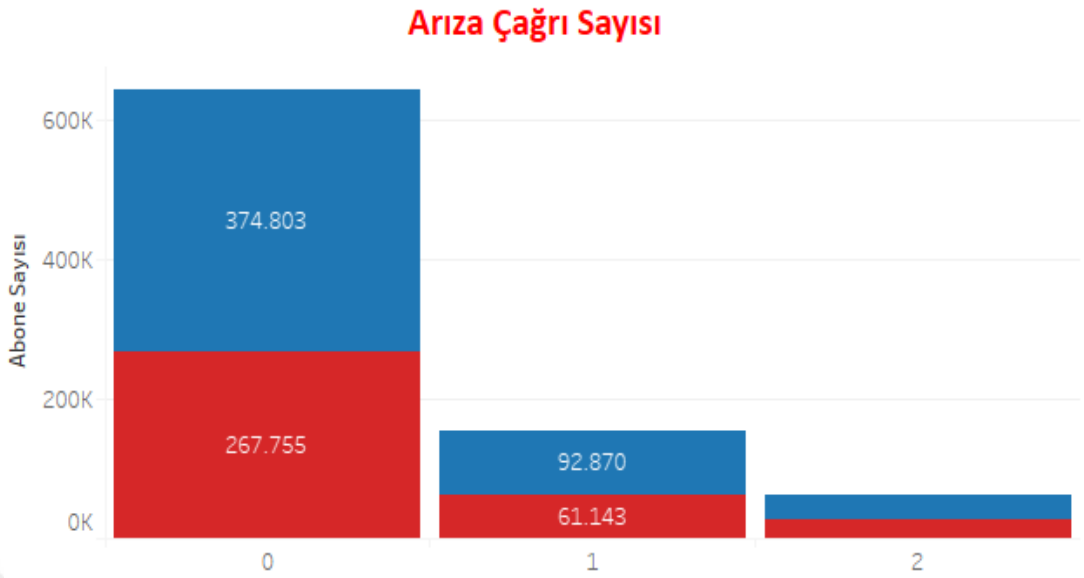
Şekil 5.10 Veri işleme sonrası abonenin son 6 ayda yaptığı kota aşım sayısına bilgisine göre abone iptal sayıları



Şekil 5.11 Veri işleme sonrası abonenin son 6 ayda yaptığı ortalama download miktarı bilgisine göre abone iptal sayıları



Şekil 5.12 Veri işleme sonrası abonenin son 6 ayda yaptığı ortalama upload miktarı bilgisine göre abone iptal sayıları

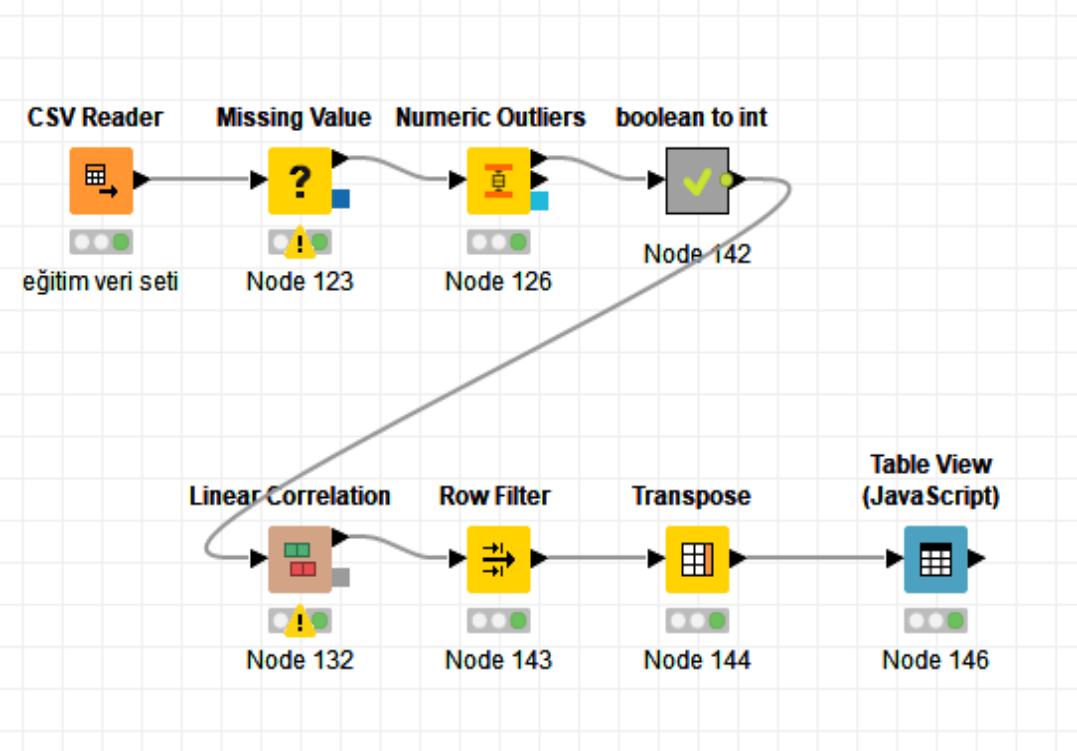


Şekil 5.13 Veri işleme sonrası abonenin son 6 ayda açtığı çağrı sayısı bilgisine göre abone iptal sayıları

5.1.3 Veri seti içinde korelasyonların bulunması

Makine öğrenmesi yöntemlerinde modeli eğitmek için kullandığımız alanların birbiri ile olan korelasyon yani ilişkileri önemlidir. Eğer sınıf etiketi olarak kullandığımız alan ile hiç ilişkisi olmayan bir alanı, modeli eğitmek için kullanırsak hem eğitme ve tahmin maliyetimizi gereksiz yere artırmış oluruz hem de model kalitesini düşürmüş oluruz. Bu yüzden bizim çalışmamızda sınıf etiketi olarak kullandığımız ve abonenin durumunu ifade eden "churn oldumu" alanı ile diğer alanlar arasındaki ilişki bulanarak bu ilişkinin '–0.1' ile '0.1' arasında olması durumlarında o kolonun veri setinden çıkarılması sağlanmıştır.

Korelasyon değerlerinin bulunması için yine KNIME Analytics içerisinde bulunan "Linear Correlation" node'u kullanılmış ve oluşturulan pakete ait ekran görüntüsü **şekil 5.14**'deki gibi olmuştur.



Şekil 5.14 Alanlar arasındaki korelasyonu bulan knime paketi

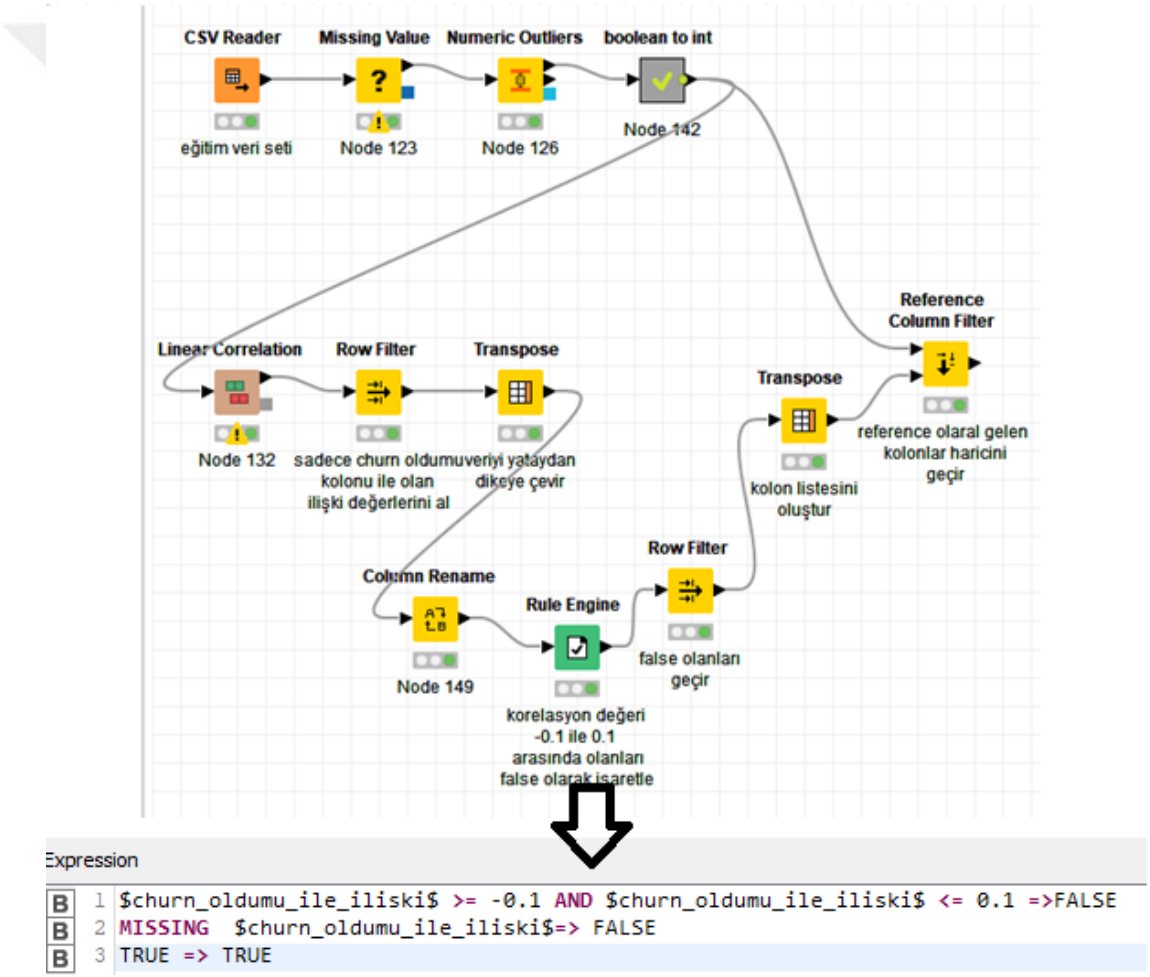
Şekil 5.14'de verilen paket çalıştığında "churn oldumu" alanı ile diğer alanlar arasındaki ilişki (korelasyon) hesaplanmış ve bu ilişki oranları **çizelge 5.4**'de verilmiştir

Çizelge 5.4 "Churn oldumu" alanı ile diğer alanlar arasındaki korelasyon değerleri

Alanlar	Churn Oldumu
Taahütü Var Mı?	-0,7328
Kalan Taahhüt Ay	-0,5349
Tv Abonesi Mi?	-0,4574
Download GB	-0,4477
Telefon Abonesi Mi?	-0,4233
Upload GB	-0,4122
TV Sinema Paket Abonesi Mi?	-0,3267
Toplam Fatura Ortalaması	-0,1369
Abonelik Yaşı Ay	-0,0599
Toplam Çağrı Sayısı	0,0039
İnternet Fatura Ortalaması	0,0931
Geç Ödenen Fatura Sayısı	0,1253
Fatura Sayısı	0,8108
Kota Aşım Sayısı	

Çizelge 5.4 ile verilen tablo incelendiğinde kota aşım sayısı ile sınıf etiketimiz arasında hiçbir ilişki bulunmadığı görülmüş, ayrıca abonelik yaşı, toplam çağrı sayısı ve internet fatura ortalaması alanları ile sınıf etiketimiz arasında çok düşük bir ilişki tespit edilmiştir. Bu yüzden bu dört alan veri setimizden çıkarılarak yeni veri setimiz model eğitimi için hazır hale getirilmiştir.

Veri setimizden bu alanları çıkarmak için KNIME analytics üzerinde bu alanları dinamik olarak tespit eden ve eğitim listesinden çıkaran **şekil 5.15**'deki paket hazırlanmıştır.



Şekil 5.15 Veri setinden ilişkisi düşük olan kolonların çıkarılması

5.2 Modelin Oluşturulması

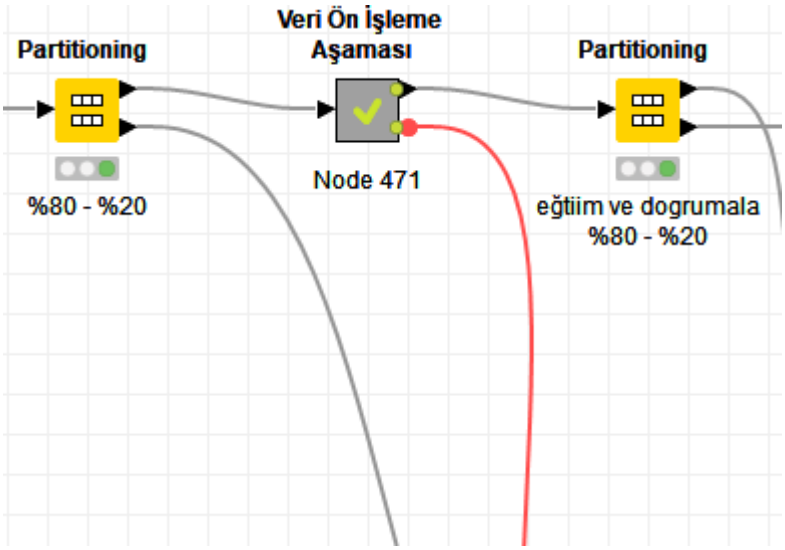
Önceki adımlarda modelimizi oluşturmak için kullanacağımız veri seti eğitim, doğrulama ve test için oluşturacağımız modeller için nasıl hazır hale getireceğimiz anlatılmıştır. Modeli oluşturmadan önce veri setimiz 2 farklı gruba bölünmüştür. Bu iki gruba grup 1 ve grup 2 denmiştir. Grup 1 toplam veri setinin %80'ini içerirken grup 2 de ise %20'lik kısım kalmıştır. Grup 1 de bulunan veri seti veri ön işleme süreçlerine sokulduktan sonra yine farklı 2 gruba, ve yine %80 - %20 oranında bölünmüştür. Bu yeni gruplara da grup 3 ve grup 4 ismi verilmiştir. Grup 3'deki veriler ile modelimizi eğitilirken grup 4'deki veriler ile doğrulama işlemi yapılmıştır. İlk bölme işleminde oluşan grup 2 ile ise veri ön işleme süreçleri çalıştırıldıktan sonra modelin testi yapılarak doğrulamadaki sonuçlara yakın sonuçlar alıp almadığımızı kontrol edilmiştir. Veri setini her gruba ayırma işlemi sırasında veri setinde sınıf etiketimiz olan ve abonenin durumunu gösteren "churn oldumu" kolonuna ait dağılım oranını bozulmadan gruplara ayırma işlemi yapılmıştır. Bu işle yine KNIME ile yapılmıştır.

Veri setini gruplara ayırdıktan sonra grup 1'deki veri seti kendi içerisinde ön işleme süreçlerine sokulmuş ve bu süreçler sonrası oluşan eksik veri, aykırı veri ve korelasyon modelleri oluşturularak test adımında kullanılmak üzere kayıt edilmiştir. Yani test işleminde kullanılan grup 2'deki veri ön işleme süreçlerindeki parametreler grup 1'deki verilere ait hesaplanan parametrelerdir. Grup 1'in %80-%20 oranlarında bölünerek oluşturulan grup 3 veri seti ile oluşturulan ve eğitilen modeller grup 2 ve grup 4 veri seti ile çalıştırılarak doğrulama ve test işlemleri gerçekleştirilmiştir. Gruplara göre veri seti boyutları ve sınıf etiketimiz olan "churn oldumu" kolonundaki dağılım aşağıdaki çizelgede verilmiştir. Bu çizelgeyi inceleyecek olursak grup 1 veri setinin ön işleme sürecine sokulması ardından oluşturulan grup 3 ve grup 4'ün sınıf etiketinde 1 değerinin dağılımı ile grup 2 aynıdır. Ayrıca toplam veriden oluşan grup 1 ve grup 2'nin de ön işleme sürecine girmeden önceki sınıf etiketinde 1 değerinin dağılımı toplam veri seti ile aynı olmuştur. Buda sınıf etiketi dağılımının değiştirilmeden veri setini istediğimiz gibi gruplara başarılı bir şekilde böldüğümüzün göstergesi olmuştur. Gruplara göre veri sayıları ve sınıf etiketinin 0 olma oranları **çizelge 5.5**'de verilmiştir.

Çizelge 5.5 Veri gruplarına göre satır sayısı ve sınıf etiketi (churn oldumu) alanında 0 değerinin oranı

Satır sayısı	Ham veri		Ön işlemeden sonra	
	Toplam	Sınıf Etiketi dağılım oranı	Toplam	Sınıf Etiketi dağılım oranı
Toplam Veri	1.364.037	55,23%	857.991	
grup 1	1.091.229	55,23%	686.429	58,49%
grup 2	272.808	55,23%	171.562	58,39%
grup 3	549.143	58,49%	-	
grup 4	137.286	58,49%	-	

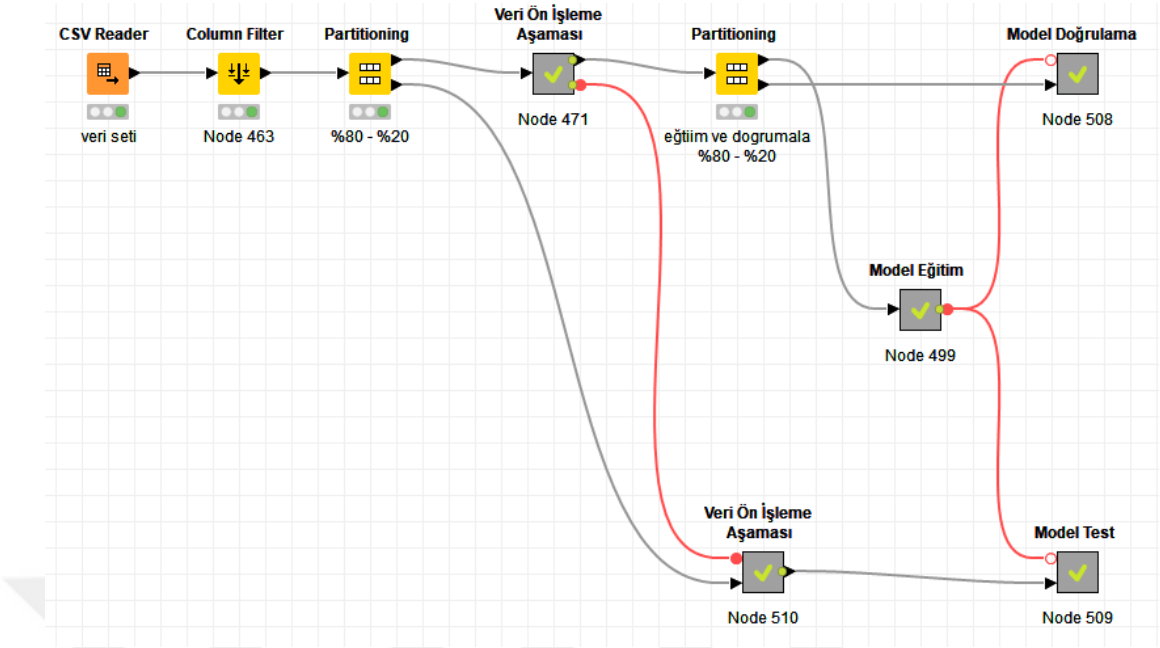
Veriyi üç farklı gruba "churn oldumu" kolonundaki dağılım bozulmadan bölmek için yine KNIME Analytics içerisinde bulunan "Partitioning" node'unu kullanılmıştır. Bu node içerisinde bulunan "stratified sampling" seçeneği seçilmiş ve dağılımını koruyacağımız kolon olarak da "churn oldumu" kolonu seçilmiştir. Bu node otomatik olarak veri setimizde seçtiğimi sınıf etiketine göre dağılımı koruyarak veriyi bölmektedir. İlgili node kullanımı **şekil 5.16**'da verilmiştir.



Şekil 5.16 Veri setini "churn oldumu" kolonunun dağılımı aynı kalacak şekilde gruplara bölme

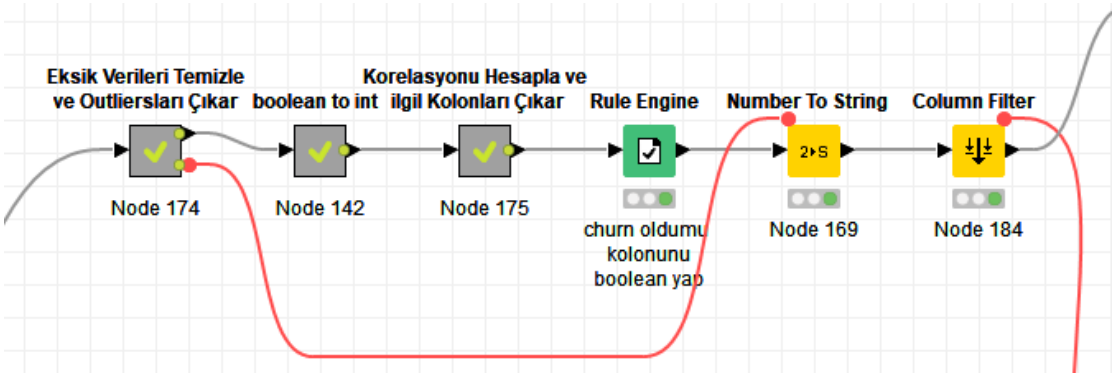
Veriyi üç farklı gruba böldükten sonra modeli eğitmek için 549.143 satır, modeli doğrulamak için 137.286 satır ve modeli test etmek için ise 171.562 satırdan oluşan veri setlerimiz oluşmuştur. Veri setlerinde doğrulama ve test için kullanılacak olanların arasındaki en temek fark veri ön işleme süreçleridir. Doğrulama için oluşturduğumuz grup 4 veri seti eğitim için oluşturduğumuz grup 3 veri seti ile beraber ön işleme süreçlerine girmiştir. Ön işleme süreçleri daha önce anlatılan eksik verilerin temizlenmesi, aykırı verilerin işlenmesi ve alanlar arası korelasyon oranlarının tespiti işlemleridir. Test veri seti ise veri ön işleme sürecine tek olarak dahil olmuş ve istatistiksel parametreleri grup 1'in ön işleme sürecinden almıştır.

Eğitim veri seti kullanılarak altı adet model oluşturulmuş ve oluşturulan model diğer veri setleri ile çalıştırılarak sonuçlar incelenmiştir. Modeller karar ağacı, random forest, naif bayes, lojistik regresyon ve xgboosting algoritmaları ile kullanılarak oluşturulmuştur. Karar ağacı ve random forest modelleri ayrı ayrı "gini index" ve "information gain ratio" ölçütleri kullanılarak modellenmiş bu sayede toplam yedi farklı model üzerinden eğitim, doğrulama ve test işlemleri yapılmıştır. Bu yedi modelin eğitilmesi ve test edilmesi yine KNIME Analytics üzerinde bulunan nodeler ile yapılabilmektedir. Model oluşturma ve test işlemi tamamlandıktan sonra KNIME üzerinden bulunan "scorer" node'u kullanılarak başarı oranları çıkarılmıştır. Bu işlemleri yapacak ve sonucu üretecek olan paket KNIME üzerinde tasarlanmış ve **şekil 5.17**'de verilmiştir.

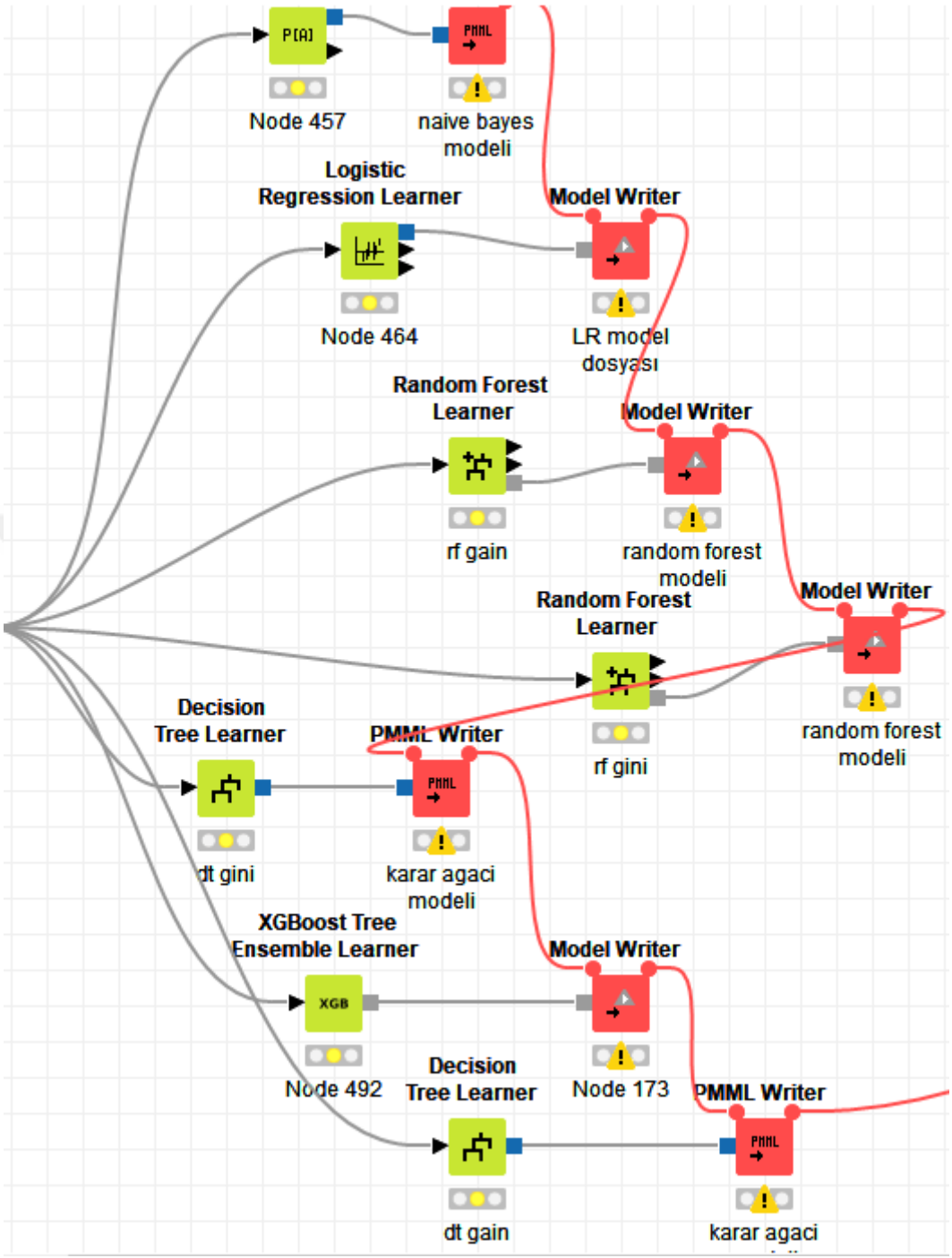


Şekil 5.17 Modelleri eğitmek, doğrulamak ve test etmek için hazırlanan knime paketi

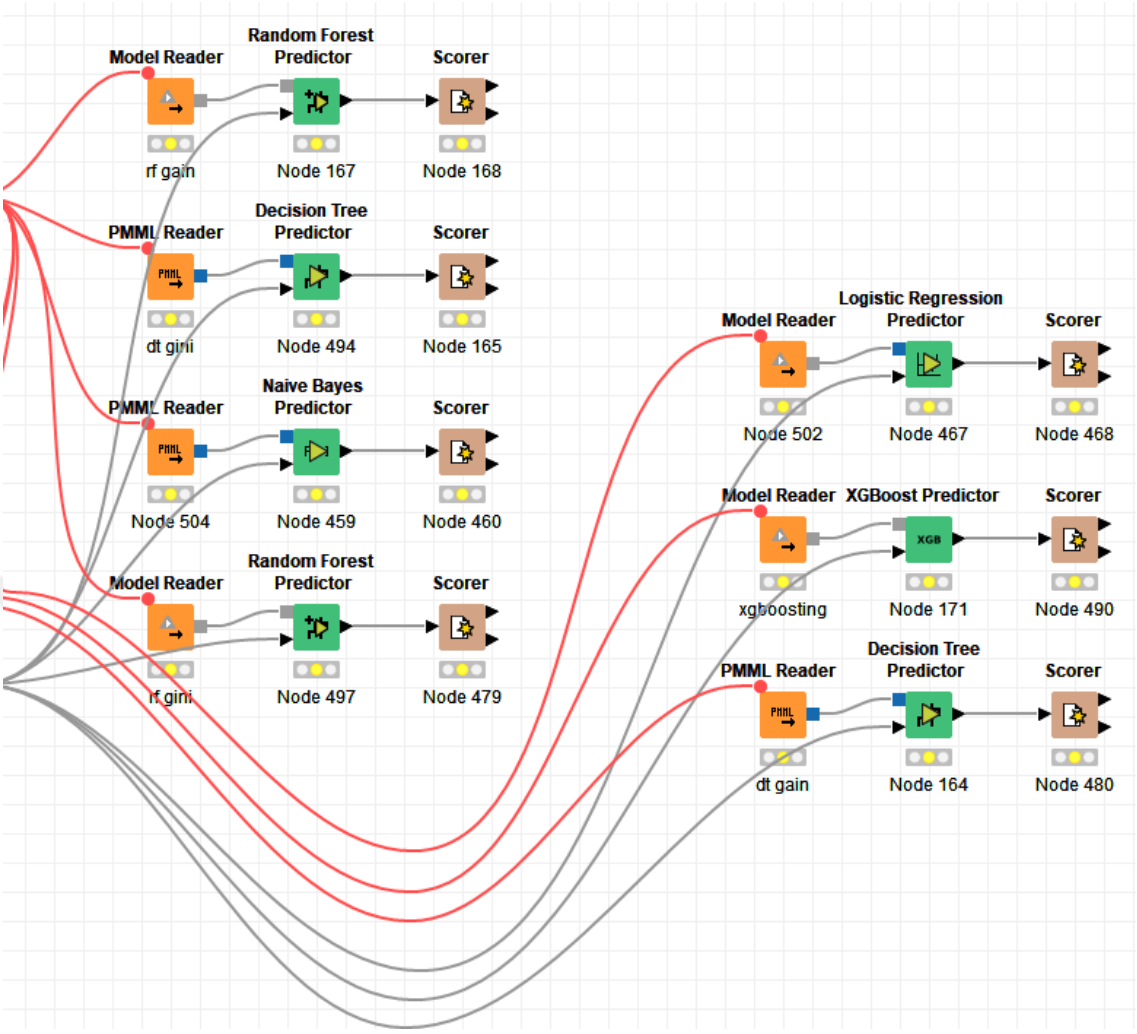
Şekil 5.17'de verilen pakette karmaşıklığı azaltmak adına her bir model bir metanode içerisine alınmıştır. Bu metanodelar sırası ile "Veri ön işleme aşaması", "Model Eğitim", "Model Doğrulama" ve "Model Test" olarak oluşturulmuştur. Bu metanodelarda ait ekran görüntüleri aşağıda şekilde **5.18-5.20**'de verilmiştir.



Şekil 5.18 Veri ön işleme aşaması metanode'u

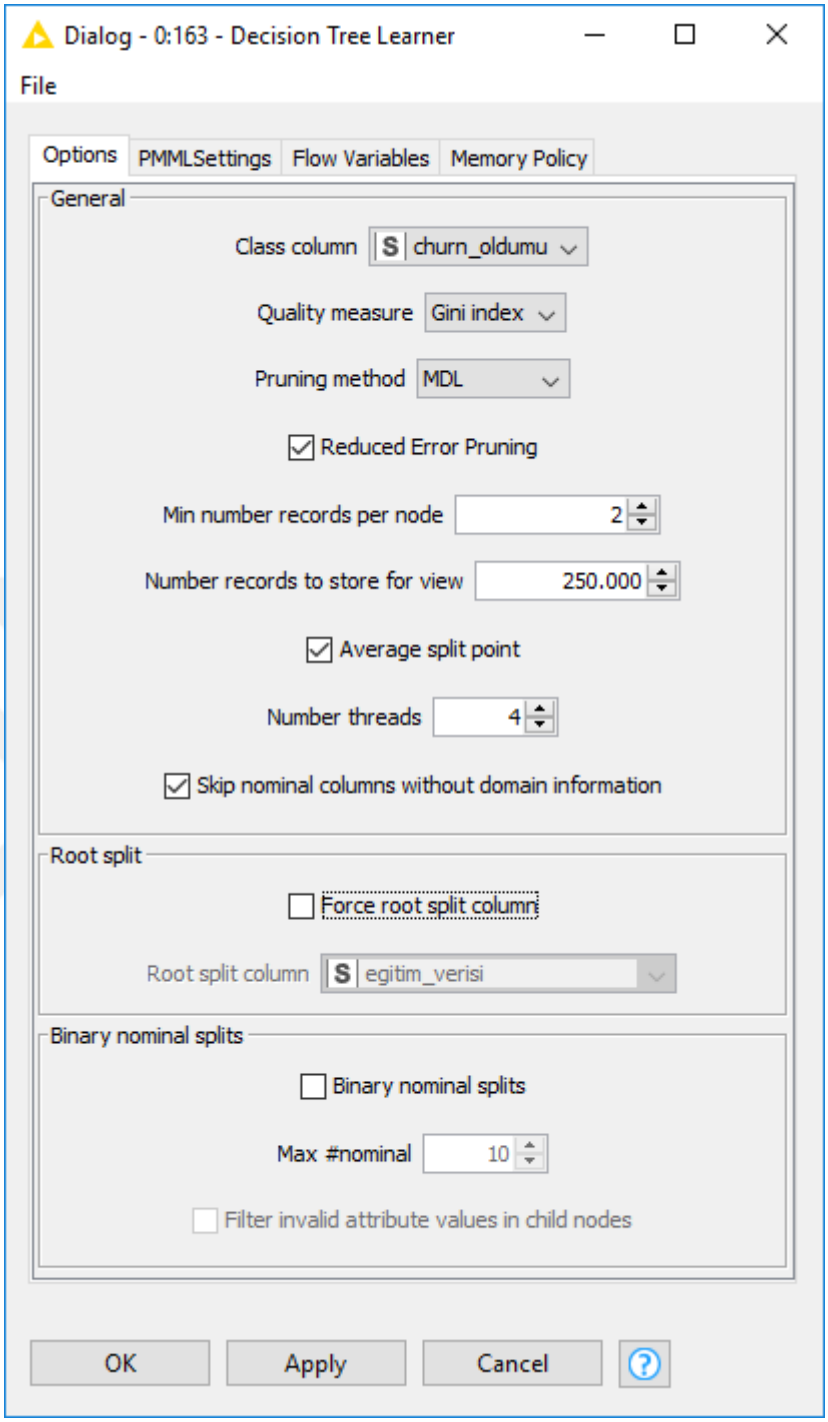


Şekil 5.19 Model eğitim metadonu'u

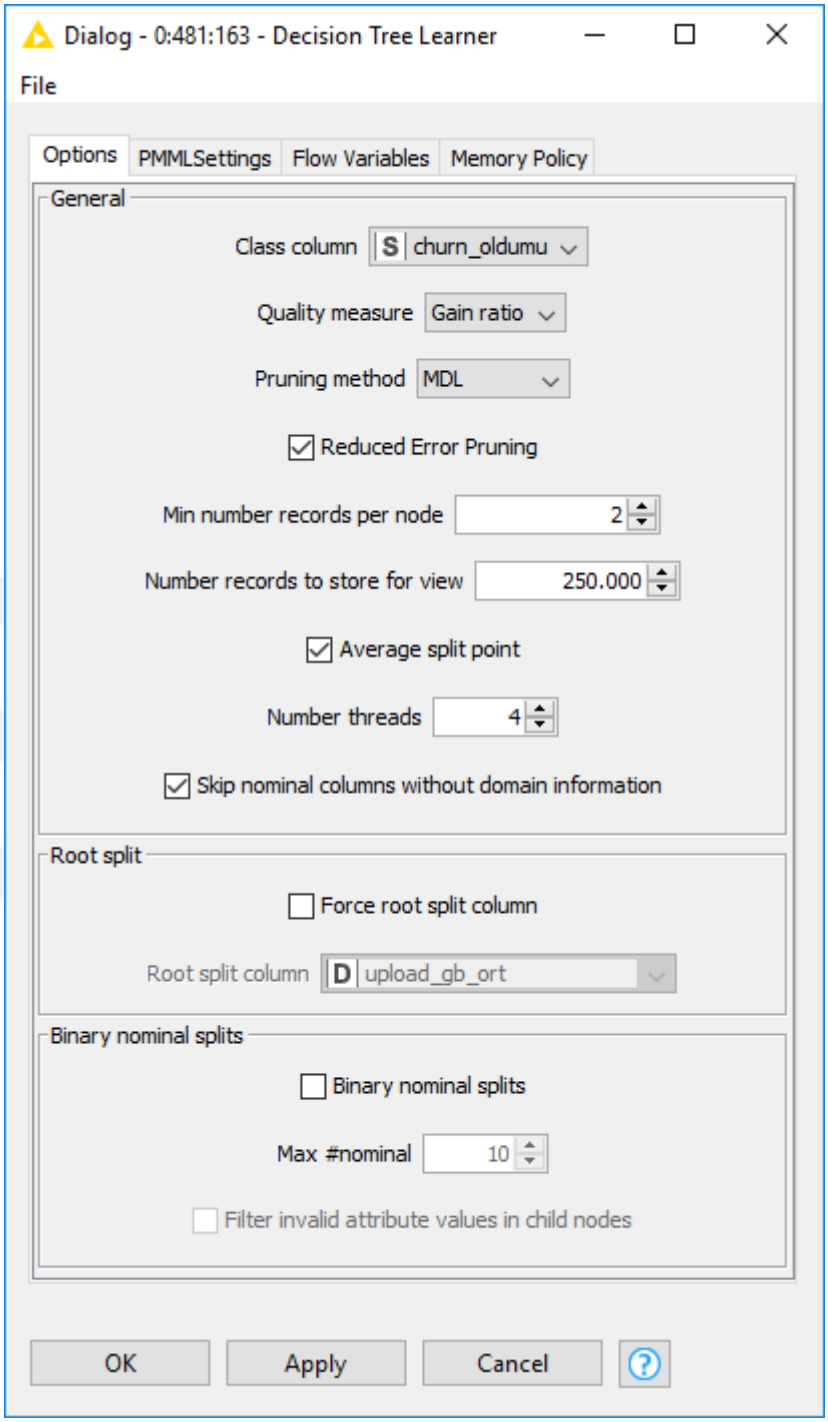


Şekil 5.20 Model doğrulama metanode'u

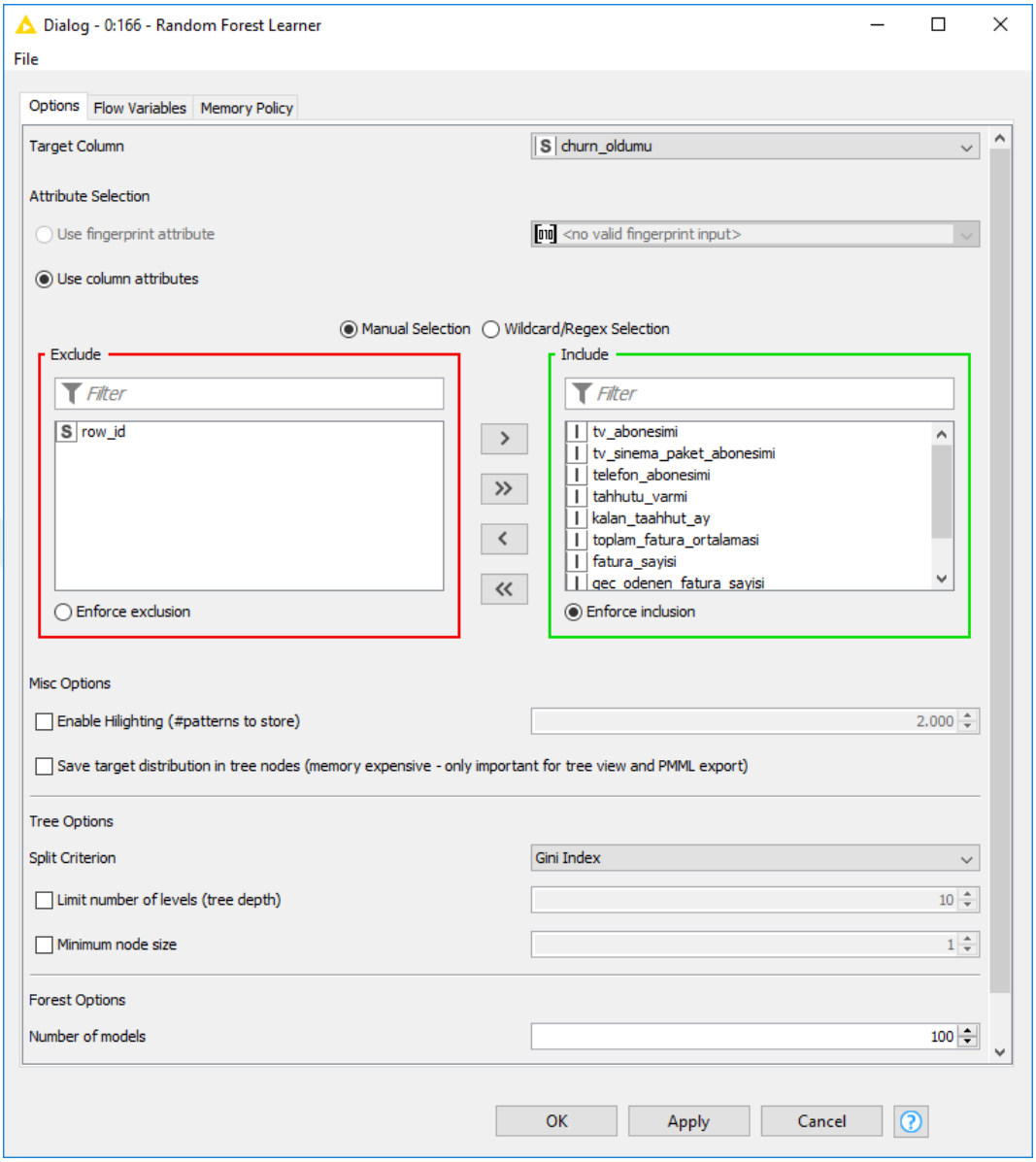
Model eğitim bölümünde kullanılan algoritmalardan karar ağacı algoritmasında aşırı öğrenmenin önüne geçebilmek için MDL ağaç budama algoritması kullanılmıştır (Sikonja ve Kononenko 1998), (Mehta vd. 1995). Random forest algoritmasında modelde 100 adet ağaç olacak şekilde ayar yapılmıştır. XGBoosting modelinde de 100 adet iterasyon yapılmıştır. Her yedi yöntem içinde ayarlar **şekil 5.21-5.27**'deki gibi yapılmıştır.



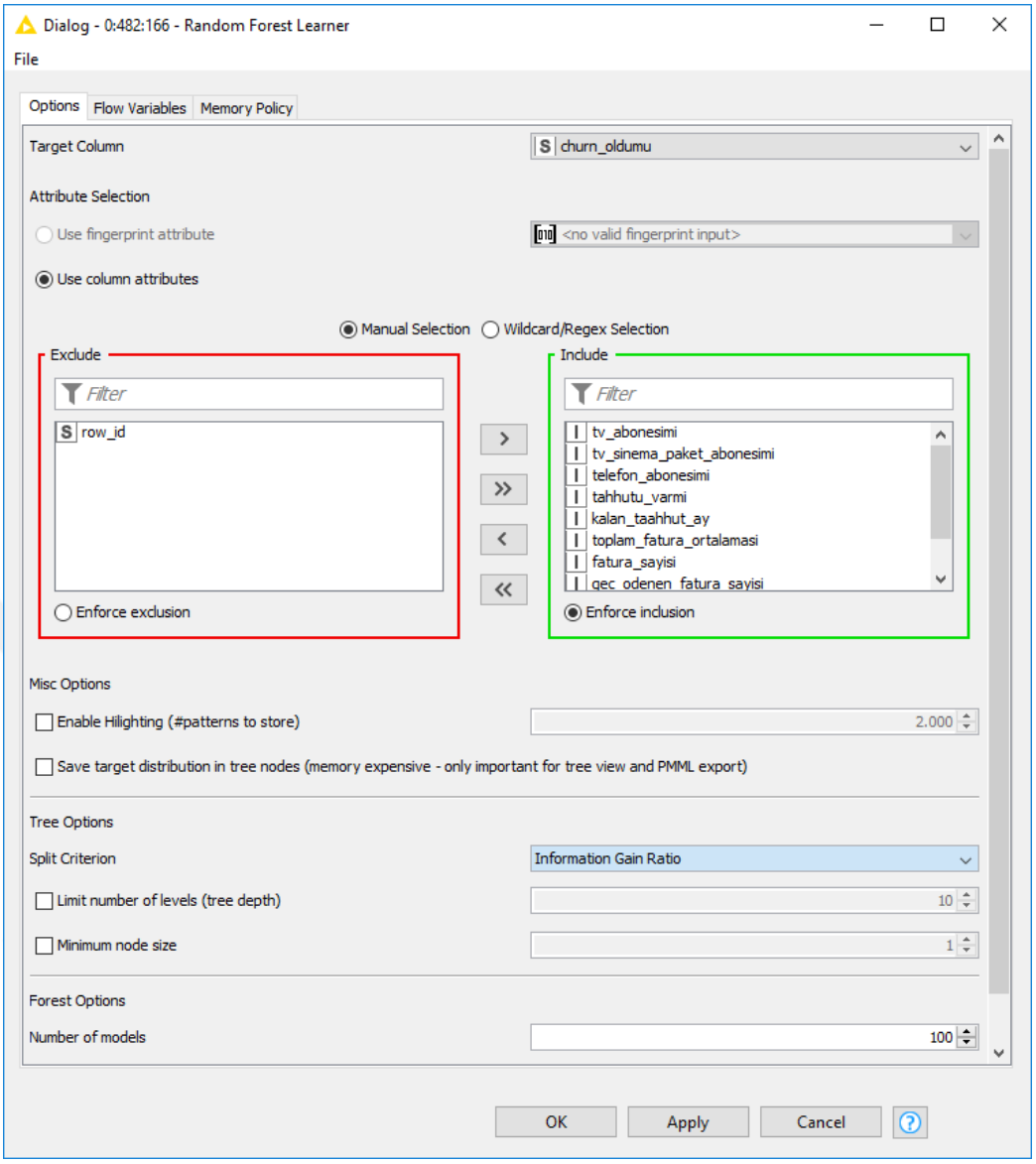
Şekil 5.21 Knime ile karar ağacı oluşturma node ayarları (gini index ile)



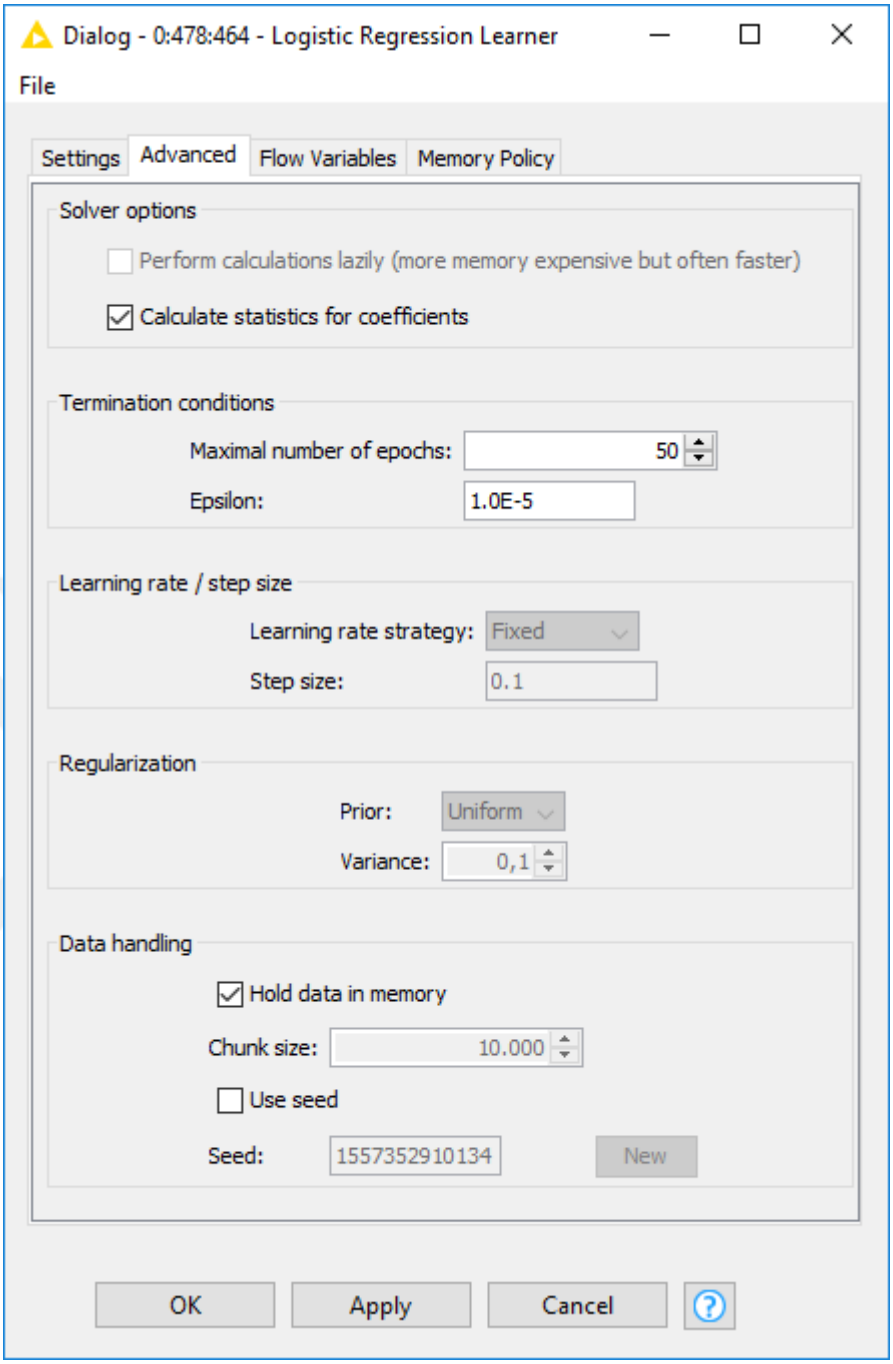
Şekil 5.22 Knime ile karar ağacı oluşturma node ayarları (gain ratio ile)



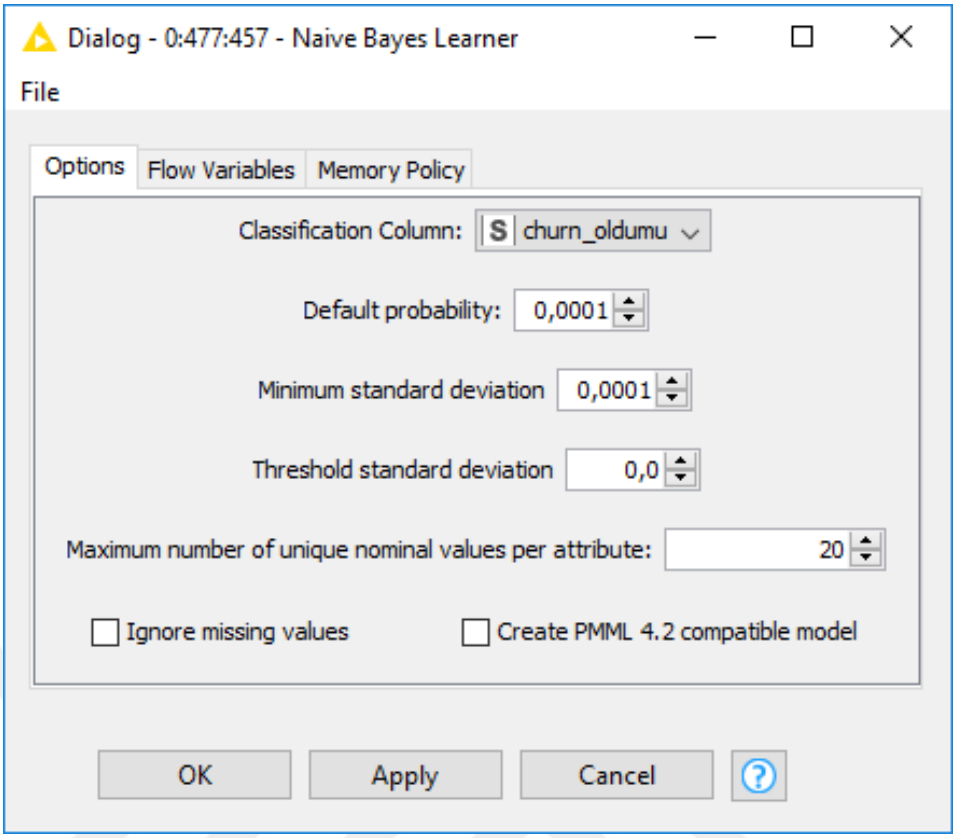
Şekil 5.23 Knime ile random forest model oluşturma node ayarları (gini index ile)



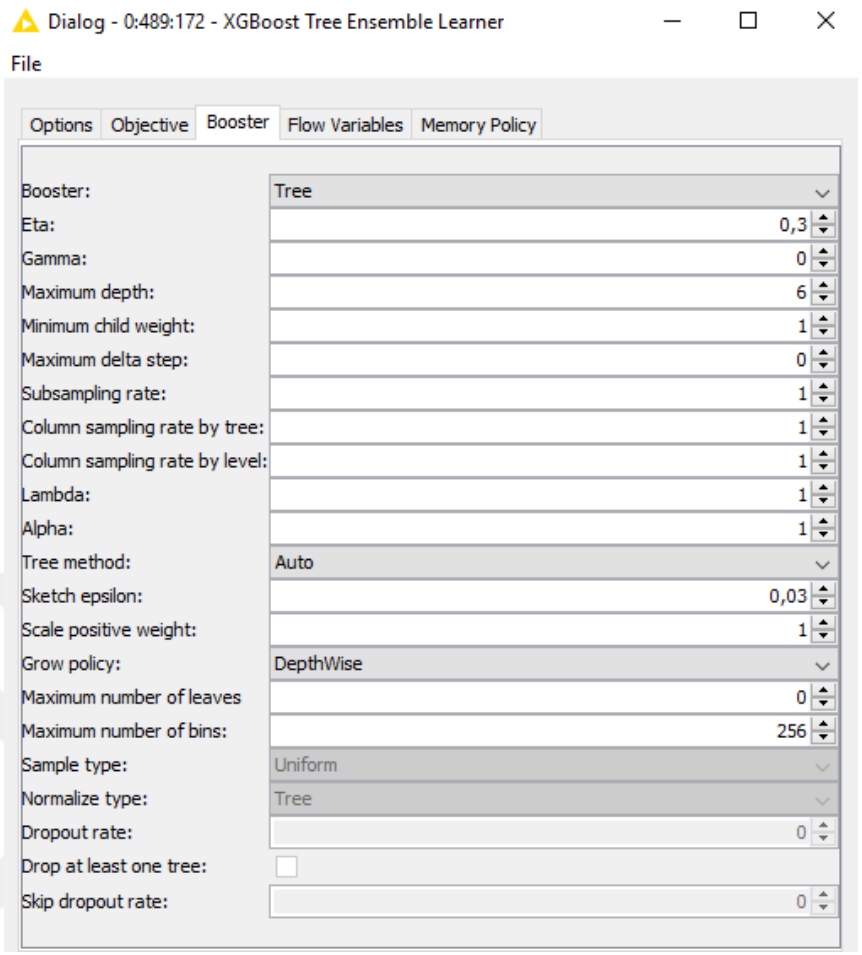
Şekil 5.24 Knime ile random forest model oluşturma node ayarları (information gain ratio ile)



Şekil 5.25 Knime ile lojistik regresyon model oluşturma node ayarları



Şekil 5.26 Knime ile naif bayes model oluşturma node ayarları



Şekil 5.27 Knime ile XGBoosting model oluşturma node ayarları

5.3 Oluşan Modellerin Değerlendirilmesi

Veri setimizi eğitim, doğrulama ve test için 3 farklı gruba ayrılmıştır. Eğitim verisi ile oluşturduğumuz modellerimizi kullanarak ilk olarak doğrulama verisi ile tahmin çalışması yapıp sonuçlar incelenmiştir. Daha sonra ise test verisi ile tekrar tahminleme çalışması yapılarak doğrulama verisi ile yapılan çalışmanın sonuçları ile karşılaştırılmış ve modellerimizin sağlıklı çalışıp çalışmadığı incelenmiştir.

Doğrulama ve test verileri ile çalıştırılan modellerin sonuçlarını incelemek için yöntem bölümünde bahsedilen karışıklık matrisi kullanılmıştır. Karışıklık matrisi kullanılarak da doğruluk, hata oranı, duyarlılık, kesinlik ve f-ölçütü gibi doğruluk istatistikleri olarak isimlendirilen değerler hesaplanarak karşılaştırma yapılmıştır.

Doğrulama verisi ile çalıştırılan modellere ait karışıklık (confusion matrix) aşağıda **çizelge 5.6'**de verilmiştir. Doğruluk verisi ile çalıştırılan modellere ait tüm ölçütlerde karışıklık matrisi ile hesaplanarak **çizelge 5.7**'da verilmiştir.

Çizelge 5.6 Doğrulama verisi ile çalıştırılan tüm modellere ait karışıklık matrisi

			Tahmin Edilen Sınıf(Predicted Class)	
			Sınıf = 1	Sınıf = 0
Naif Bayes	Gerçek Sınıf Değeri (Actual Class)	Sınıf = 1	52.686	4.296
		Sınıf = 0	7.546	72.758
Lojistik Regresyon	Gerçek Sınıf Değeri (Actual Class)	Sınıf = 1	51.351	5.631
		Sınıf = 0	5.461	74.843
Karar Ağacı-Gini	Gerçek Sınıf Değeri (Actual Class)	Sınıf = 1	76.490	3.814
		Sınıf = 0	4.284	52.698
Karar Ağacı-Gain	Gerçek Sınıf Değeri (Actual Class)	Sınıf = 1	52.701	4.281
		Sınıf = 0	3.879	76.425
Random Forest-Gini	Gerçek Sınıf Değeri (Actual Class)	Sınıf = 1	52.184	4.798
		Sınıf = 0	3.971	76.333
Random Forest-Gain	Gerçek Sınıf Değeri (Actual Class)	Sınıf = 1	52.187	4.795
		Sınıf = 0	3.978	76.326
XGBoosting	Gerçek Sınıf Değeri (Actual Class)	Sınıf = 1	52.994	3.988
		Sınıf = 0	3.669	76.635

Çizelge 5.7 Doğrulama verisi ile çalıştırılan modellere ait ölçütler (doğruluk istatistikleri)

Algoritma	Doğruluk	Hata Oranı	Sınıf	Duyarlılık	Kesinlik	F-Ölçütü
Naif Bayes	91,37%	8,63%	1	92,46%	87,47%	89,90%
			0	90,60%	94,42%	92,47%
Lojistik Regresyon	91,92%	8,08%	1	90,12%	90,39%	90,25%
			0	93,20%	93,00%	93,10%
Karar Ağacı - Gini	94,10%	5,90%	1	95,25%	94,70%	94,97%
			0	92,48%	93,25%	92,86%
Karar Ağacı - Gain	94,06%	5,94%	1	92,49%	93,14%	92,81%
			0	95,17%	94,70%	94,93%
Random Forest - Gini	93,61%	6,39%	1	91,58%	92,93%	92,25%
			0	95,06%	94,09%	94,57%
Random Forest - Gain	93,61%	6,39%	1	91,59%	92,92%	92,25%
			0	95,05%	94,09%	94,57%
XGBoosting	94,42%	5,58%	1	93,00%	93,52%	93,26%
			0	95,43%	95,05%	95,24%

Doğrulama verisi ile çalıştırılan modellerin sonuçları yer alan yukarıdaki çizelgeler ile karşılaştırma yapabilmek ve modelin sağlık durumunu anlayabilmek için test verisi ile de aynı modeller çalıştırılmış ve sonuçlar aşağıdaki **çizelge 5.8'**de verilmiştir.

Çizelge 5.8 Test verisi ile çalıştırılan tüm modellere ait karışıklık matrisi

			Tahmin Edilen Sınıf(Predicted Class)	
			Sınıf = 1	Sınıf = 0
Naif Bayes	Gerçek Sınıf Değeri (Actual Class)	Sınıf = 1	66.743	4.651
		Sınıf = 0	8.576	91.592
Lojistik Regresyon	Gerçek Sınıf Değeri (Actual Class)	Sınıf = 1	65.180	6.214
		Sınıf = 0	6.075	94.093
Karar Ağacı - Gini	Gerçek Sınıf Değeri (Actual Class)	Sınıf = 1	96.139	4.029
		Sınıf = 0	4.585	66.809
Karar Ağacı - Gain	Gerçek Sınıf Değeri (Actual Class)	Sınıf = 1	66.907	4.487
		Sınıf = 0	4.146	96.022
Random Forest - Gini	Gerçek Sınıf Değeri (Actual Class)	Sınıf = 1	66.130	5.264
		Sınıf = 0	4.164	96.004
Random Forest - Gain	Gerçek Sınıf Değeri (Actual Class)	Sınıf = 1	66.117	5.277
		Sınıf = 0	4.155	96.013
XGBoosting	Gerçek Sınıf Değeri (Actual Class)	Sınıf = 1	67.206	4.188
		Sınıf = 0	3.883	96.285

Test verisi ile çalıştırılan modellerin ürettiği sonuçlara göre oluşturulan karışıklık matrisi üzerinden doğruluk, hata oranı, duyarlılık, kesinlik ve f-ölçütü gibi doğruluk istatistikleri hesaplanmış ve aşağıdaki **çizelge 5.9'**da verilmiştir.

Çizelge 5.9 Test verisi ile çalıştırılan modellere ait ölçütler (doğruluk istatistikleri)

Algoritma	Doğruluk	Hata Oranı	Sınıf	Duyarlılık	Kesinlik	F-Ölçütü
Naif Bayes	92,29%	7,71%	1	93,49%	88,61%	90,98%
			0	91,44%	95,17%	93,27%
Lojistik Regresyon	92,84%	7,16%	1	91,30%	91,47%	91,39%
			0	93,94%	93,81%	93,87%
Karar Ağacı - Gini	94,98%	5,02%	1	95,98%	95,45%	95,71%
			0	93,58%	94,31%	93,94%
Karar Ağacı - Gain	94,97%	5,03%	1	93,72%	94,16%	93,94%
			0	95,86%	95,54%	95,70%
Random Forest - Gini	94,50%	5,50%	1	92,63%	94,08%	93,35%
			0	95,84%	94,80%	95,32%
Random Forest - Gain	94,50%	5,50%	1	92,61%	94,09%	93,34%
			0	95,85%	94,79%	95,32%
XGBoosting	95,30%	4,70%	1	94,13%	94,54%	94,34%
			0	96,12%	95,83%	95,98%

6. SONUÇ

Bu çalışmada telekomünikasyon sektöründe, aldığı hizmeti iptal etme olasılığı yüksek olan abonelerin makine öğrenmesi yöntemlerinden olan 5 farklı sınıflandırma yöntemi ile tespit etmeye çalışılmış ve sonuçlar karşılaştırmak üzere üretilmiştir. Çalışmada naif bayes, lojistik regresyon, karar ağaçları, random forest ve xgboosting yöntemleri kullanılmıştır. Tüm yöntemlere ait oluşturulan modeller doğrulama ve test verileri ile ayrı ayrı çalıştırılmış ve sonuçları karşılaştırılmak üzere aşağıdaki **çizelge 6.1**'de bir araya getirilmiştir.

Çizelge 6.1 Doğrulama ve test verisi ile çalıştırılan modellerin doğruluk istatistikleri

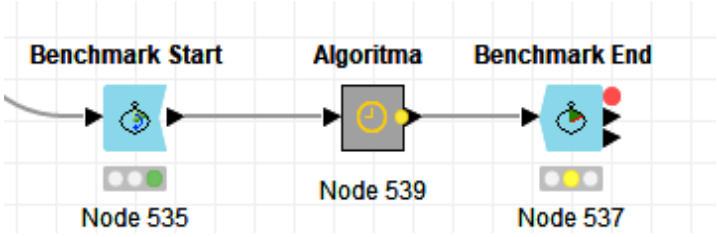
Algoritma	Doğruluk		Hata Oranı		Sınıf	Duyarlılık		Kesinlik		F-Ölçütü	
	Doğrulama	Test	Doğrulama	Test		Doğrulama	Test	Doğrulama	Test	Doğrulama	Test
Naif Bayes	91,37 %	92,29 %	8,63 %	7,71 %	1	92,46%	93,49 %	87,47 %	88,61 %	89,90 %	90,98 %
					0	90,60%	91,44 %	94,42 %	95,17 %	92,47 %	93,27 %
Lojistik Regresyon	91,92 %	92,84 %	8,08 %	7,16 %	1	90,12%	91,30 %	90,39 %	91,47 %	90,25 %	91,39 %
					0	93,20%	93,94 %	93,00 %	93,81 %	93,10 %	93,87 %
Karar Ağacı (Gini)	92,64 %	93,81 %	6,19 %	5,02 %	1	94,01%	94,98 %	93,46 %	94,46 %	93,73 %	94,72 %
					0	90,73%	92,18 %	91,48 %	92,90 %	91,10 %	92,54 %
Karar Ağacı (Gain)	92,60 %	93,80 %	6,20 %	5,03 %	1	90,73%	92,31 %	91,38 %	92,76 %	91,05 %	92,54 %
					0	93,92%	94,86 %	93,46 %	94,54 %	93,69 %	94,70 %
Random Forest (Gini)	93,61 %	94,50 %	6,39 %	5,50 %	1	91,58%	92,63 %	92,93 %	94,08 %	92,25 %	93,35 %
					0	95,06%	95,84 %	94,09 %	94,80 %	94,57 %	95,32 %
Random Forest (Gain)	93,61 %	94,50 %	6,39 %	5,50 %	1	91,59%	92,61 %	92,92 %	94,09 %	92,25 %	93,34 %
					0	95,05%	95,85 %	94,09 %	94,79 %	94,57 %	95,32 %
XGBoosting	94,42 %	95,30 %	5,58 %	4,70 %	1	93,00%	94,13 %	93,52 %	94,54 %	93,26 %	94,34 %
					0	95,43%	96,12 %	95,05 %	95,83 %	95,24 %	95,98 %

Yukarıdaki çizelge incelendiğinde doğrulama ve test verisi ile çalıştırılan modellerin iki veri seti içinde birbirine yakın sonuçlar ürettiği gözlemlenmiştir. Bu durum modelimizin sağlıklı olarak eğitilip oluşturulduğunu göstermiştir. Hem doğrulama hem de test verisi üzerinde inceleme yaptığımız zaman modellerin hepsinin birbirine yakın sonuçlar ürettiğini görsekte en başarısız model naif bayes ile üretilen model olurken en başarılı yöntem ise xgboosting yöntemi olmuştur. XgBoosting yönteminde sonra ise random forest yöntemi gelmektedir. Bu iki yöntemin başarılı olması bagging ve boosting yaklaşımları olmuştur. Bizim ulaştığımız başarı oranları ile daha önce yapılan çalışmalardaki başarı oranlarını karşılaştıracak olursak **Çizelge 6.2**'deki gibi bir sonuç oluşmaktadır. Bu çizelgedede görebildiğimiz gibi başarı oranımız incelediğimiz birçok çalışmadan daha yüksektir.

Çizelge 6.2 Sonuçların Karşılaştırılması

Çalışma	**Skor**
Bizim çalışmamız (En iyi sonuç)	95,30%
Anonymous 2019	79,95%
Coşkun ve Baykal 2011	86,36%
Günay 2018	91,00%
Kaynar vd. 2017	91,35%

Sonuçların yanı sıra kullandığımız tüm yöntemlerin eğitim ve sonra test verisi ile çalışması için gerekli olan süreler de hesaplanarak aşağıdaki **çizelge 6.2**'de verilmiştir. Algoritmaların çalışma süreleri KNIME üzerinde bulunan "benchmark start" ve "benchmark end" node'ları kullanılarak bulunmuştur. Bu nodeların kullanımı **şekil 6.1**'deki gibidir.



Şekil 6.1 KNIME üzerinde çalışma sürelerinin bulunması için kullanılan benchmark node'ları

Çizelge 6.3 Algoritmalar eğitim ve test çalışma süreleri

Algoritma	Eğitim Süresi(sn)	Çalışma süresi(sn)
XGBoosting	545,10	42,81
Random Forest (Gain)	517,52	54,71
Random Forest (Gini)	516,57	54,57
Karar Ağacı (Gain)	456,63	46,27
Karar Ağacı (Gini)	103,61	46,91
Lojistik Regresyon	45,50	48,50
Naif Bayes	9,12	39,77

Yukarıdaki çizelge incelendiğinde en uzun eğitim süresi beklendiği gibi bir boosting yöntemi olan XGBoosting algoritmasına aittir. Bagging yöntemi ile bir den fazla ağaç üreten random forest algoritması da 2 ve 3 sırada yer almaktadır. En kötü sonucu üreten naif bayes algoritması en kısa sürede eğitilen ve çalıştırılan algoritma olmuştur.

Çalışmamızın en önemli ve zor olan kısmı modelleri uygulamak değil, modellerde kullanacağımız verinin kaynak sistemlerden üretilmesi ve üretilen bu verinin veri işleme yöntemleri ile temizlenerek kullanılabilir hale getirilmesi sürecidir. Çalışmamızdaki tüm süreç tamamen ücretsiz olan KNIME uygulaması üzerinde yapılmıştır. KNIME gibi bir ürün kullanmamızdan dolayı çalışmamızda da asıl önemli kısım olan veriyi tanımlama ve hazırlama kısımlarına daha çok zaman ve efor sarf edilebilmiştir. Ayrıca KNIME uygulamasının makine öğrenmesi yöntemlerinin kullanılacağı çalışmalarda başarılı bir şekilde kullanılabileceği de görülmüştür.

Bu çalışma sonrası üretilen modeller şirket aktif aboneleri üzerinde çalıştırılarak iptal olma olasılığı yüksek olanların bulunması ve bunlarla ilgili aksiyon alınması için bir temel teşkil etmektedir. Bu ileriye dönük çalışmanın yapılabilmesi ve sonuçlarının paylaşılabilmesi için izin alınması ve KVKK kapsamında uygun verinin ilgili şirket tarafından paylaşılması gerekmektedir. Bu yüzden bu ileriye dönük yapılabilecek olan çalışma bizim çalışmamızda yer almamıştır.

KAYNAKLAR

Anonymous. 2018. Customer Churn – Logistic Regression with R. Web Sitesi: http://www.treselle.com/blog/customer-churn-logistic-regression-with-r/, Erişim Tarihi: 01.02.2019.

Başarslan, M. S. 2017. Telekomünikasyon Sektöründe Müşteri Kayıp Analizi. Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, Düzce Üniversitesi, Düzce.

Burez, J., Van den Poel, D.,2009. Handling class imbalance in customer churn prediction, Elsevier, (36), 4626-4636

Coşkun, C., Baykal, A. 2011. Veri madenciliğinde sınıflandırma algoritmalarının bir örnek üzerinde karşılaştırılması. Akademik Bilişim Konferansı, 2-4 Şubat 2011, İnönü Üniversitesi, Malatya.

Coussement, K., Lessmann, S., ve Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. Decision Support Systems, 27-36.

Eulor T. 2005. Churn Predictionin Telecommunications Using Mining Mart, 5th IEEE International Conference, Data Mining (ICDM), 99-106, Texas, USA

Günay, M., Ensari, T.,2018, Makina Öğrenmesiyle Müşteri Kayıplarının Analizi, 26th Signal Processing and Communications Applications Conference

Gürcan, M. (1998). Lojistik regresyon analizi ve bir uygulama. Yüksek lisans tezi (Yayınlanmamış). Ondokuz Mayıs Üniversitesi, Fen Bilimleri Enstitüsü, Samsun.

Hair, J. F., Black, W. C., Babin, B., Anderson, R. E., Tatham, R. L. (2006). Multi-variate data analysis (6th ed). Upper Saddle River, NJ: Prentice-Hall.

Kaynar, O., Tuna, M. F., Görmez, Y. ve Deveci, M. A. (2017). Makine öğrenmesi yöntemleriyle müşteri kaybı analizi. C.Ü. İktisadi ve İdari Bilimler Dergisi.

Kim K, Jun C., Lee J. 2014. Improved churn prediction in telecommunication industry by analyzing a large network. Elsevier, 41(15), 6575-6584.

Kotler, P., Keller K. L., 2015. Marketing Management, 5.Baskı

Lawrence, J. A., Pasternack, J. B. A. 2002. Applied Management Science Modeling: Spreadsheet Analysis, and Communication for Decision Making, Second Edition, USA. (syf: 350:352).

Mehta, M., Rissanen, J., Agrawal, R. 1995. MDL-based Decision Tree Pruning

Odabaş, Ö. 2017. Veri Madenciliği Teknikleri ile Telekom Sektöründe Ayrılan Müşteri Analizi. T.C. İstanbul Ticaret Üniversitesi, Yüksek Lisans Tezi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, İstanbul

Sikonja, M.R., Kononenko, I. 1998. Pruning regression trees with MDL

Spanoudes, P., Nguyen, T. 2017. Deep Learning in Customer Churn Prediction: Unsupervised Feature Learning on Abstract Company Independent Feature Vectors, arXiv, 1703.03869.

Ulucan, A. 2007. Yöneylem Araştırması İşletmecilik Uygulamalı Bilgisayar Destekli Modelleme, Siyasal Kitabevi, 2. Baskı, 341, Ankara.

Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G., ve Chatzisavvas, K.Ch. 2015. A comparison of machine learning techniques for customer churn prediction. Elsevier, 55, 1-9

Verbekea W, Martens D., Baesens B. 2014. Social network analysis for customer churn prediction. Elsevier, 14(C), 431-446

Yüceoğlu B. 2018. Web Sitesi: http://www.veridefteri.com/wp-content/uploads/2018/06/Boosting.pdf, Erişim Tarihi: 16.05.2019.

ÖZGEÇMİŞ

Adı Soyadı : Mehmet Sabri KUNT
Doğum Yeri : Sarıkaya / YOZGAT
Doğum Tarihi : 03.10.1981
Medeni Hali : Evli
Yabancı Dili : İngilizce

Eğitim Durumu

Lise : Sarıkaya Lisesi
Lisans : Selçuk Üniversitesi Mühendislik Mimarlık Fakültesi Bilgisayar Mühendisliği Bölümü (2005)
Yüksek Lisans : Ankara Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı (2019)

Çalıştığı Kurumlar

Türksat A.Ş. 16.01.2017 – Devam ediyor
ÖSYM 15.03.2016 – 31.12.2016
Türk Telekom, 10.2008 –03.2016
Teta Elektronik, 01.2008 – 10.2008
Simetri Yazılım, 12.2005 – 12.2007