

# T.C. ALTINBAŞ ÜNİVERSİTESİ Lisansüstü Eğitim Enstitüsü Veri Analitiği Anabilim Dalı

# MÜŞTERİ KAYBI ÖNGÖRÜSÜ İÇİN KNIME: C4.5 DECISION TREE ALGORİTMASI İLE TELCO CHURN ANALİZİ

# EMRE ÖZYÜREK

VA520 Veri Analitiği İçin Makine Öğrenmesinde Gelişmeler

> Danışman Doç Dr Doğu Çağdaş Atilla

## ÖZET

Telco şirketleri, müşteri kaybetme oranlarını azaltmak ve işletmelerini sürdürülebilir kılmak için müşteri sadakatini artırmaya odaklanmaktadır. Bu amaçla, churn analitiği önemli bir araç haline gelmiştir. Bu projenin amacı, bir telco şirketinin müşteri churn oranını tahmin etmek için C4.5 karar ağacı algoritmasını kullanmaktır. Veri analizi ve modelleme süreci, KNIME Analytics Platform'u kullanarak gerçekleştirilmiştir. İlk adım olarak, telco şirketinin müşteri verileri toplanmış ve temizlenmiştir. Bu veriler, müşterilerin bilgileri, hizmet kullanımı, fatura bilgileri ve geçmiş churn durumları gibi çeşitli özellikleri içermektedir. Daha sonra, C4.5 karar ağacı algoritması kullanılarak model eğitilmiştir. Algoritma, veri setindeki değişkenlerin bilgi kazancını kullanarak en önemli ayrımcı özellikleri belirlemektedir. Veri üzerinde çeşitli özellik mühendisliği gerçekleştirilmiştir ve model eğitimi sürecinde, veri seti eğitim ve test kümelerine ayrılmıştır. Elde edilen model, müşteri churn olasılığını tahmin etmek için kullanılmıştır. Modelin performansı değerlendirilirken, doğruluk, hassasiyet, kesinlik ve F1 skoru gibi metrikler kullanılmıştır. Ayrıca, modelin karmaşıklığı ve yorumlanabilirliği de değerlendirilmiştir. Sonuçlar, C4.5 karar ağacı algoritmasının telco churn tahmininde etkili bir şekilde kullanılabileceğini göstermektedir. Model, müşterilerin churn riskini doğru bir şekilde tahmin etme yeteneği ile şirketin müşteri sadakatini artırmak için stratejik kararlar almasına yardımcı olabilir. Ayrıca, modelin karmaşıklığı düşük olduğundan, işletme tarafından kullanımı ve yorumlanması kolaydır. Bu proje, telco şirketlerine churn analitiğinde C4.5 karar ağacı algoritmasının kullanılmasının faydalarını ve potansiyelini göstermektedir. Daha fazla veri ve gelişmiş analiz yöntemleri kullanılarak, müşteri churn oranının daha da azaltılması ve işletme performansının iyileştirilmesi hedeflenmektedir.

Anahtar Kelime: Algoritma, Churn, Knime

# İÇİNDEKİLER

	<u>sayfa</u>
ÖZET	2
1. GİRİŞ	4
2. UYGULAMA	5
2.1. VERI	5
2.2. ÖN İŞLEME	5
2.2.1. Sütun ve satır filtreleme	6
2.2.2. Math Formula	7
2.3. DÖNÜŞTÜRME	7
2.3.1. Numeric Binner	8
2.3.2. Rule Engine	8
2.3.3. Category to Number modülü ile label encoding işlemi	9
2.4. SINIFLANDIRMA ALGORITMALARI	11
2.4.1. Decision Tree Learner modülü:	11
2.4.2. Decision Tree Predictor modülü:	12
2.4.3. Scorer modülü:	13
3. SONUC	15

## 1. GİRİŞ

Bu projenin amacı, Knime uygulamasında C4.5 decision tree algoritmasını kullanarak telco churn tahmini yapmaktır. C4.5 decision tree algoritması, veri madenciliği ve makine öğrenmesinde yaygın olarak kullanılan bir sınıflandırma algoritmasıdır. Bu algoritma, karar ağacı olarak bilinen bir yapı oluşturarak verileri sınıflandırır.

C4.5 algoritması, bilgi kazancı (information gain) ölçütünü kullanarak en iyi ayrımı sağlayan özellikleri belirler. Bu özellikler, veri setini en iyi şekilde bölerek homojen alt kümelere ayırmayı hedefler. Algoritma, veri setindeki değişkenlerin bilgi kazancını hesaplayarak, her bir değişkenin sınıflandırma performansını ölçer.

Karar ağacı modeli oluşturulurken, C4.5 algoritması ağaç büyüme yöntemlerini kullanır. Başlangıçta, tüm veri seti kök düğüm olarak kabul edilir. Ardından, en iyi ayrımı sağlayan özellik seçilir ve bu özelliğe göre veri seti alt kümelerine bölünür. Bu adım, ağacın bir seviyesini oluşturur. İşlem alt kümeleri için tekrarlanır ve ağaç büyüdükçe daha fazla ayrım gerçekleştirilir.

C4.5 algoritması, veri setindeki eksik değerleri de ele alır. Eksik değerler, ağaç büyümesi sırasında uygun şekilde yönetilir ve tahminlerdeki yanlılıkları en aza indirmek için stratejik olarak kullanılır.

Karar ağacı modeli oluşturulduktan sonra, "Decision Tree Predictor" modülü kullanılarak test veri kümesine uygulanır. Bu modül, her bir müşteri için karar ağacı modelini kullanarak churn olasılığını tahmin eder. Tahminler, belirlenen sınıf eşik değeri üzerinden sınıflandırılır.

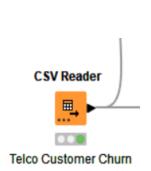
Son olarak, "Scorer" modülü kullanılarak tahmin sonuçları değerlendirilir. Bu modül, tahmin edilen churn sonuçları ile gerçek churn durumlarını karşılaştırarak çeşitli performans metriklerini hesaplar. Doğruluk, hassasiyet, kesinlik ve F1 skoru gibi metrikler, modelin churn tahmin yeteneğini değerlendirmek için kullanılır.

C4.5 decision tree algoritması, veri setindeki değişkenlerin önemini ölçerek ve sınıflandırma kurallarını bir karar ağacı olarak temsil ederek veri analizi ve tahminleme süreçlerinde etkili bir araç sağlar. Knime platformu, bu algoritmanın uygulanmasını kolaylaştırarak kullanıcılara veri işleme, model oluşturma ve sonuç analizi gibi teknik işlemleri görsel bir sekilde gerçekleştirme imkanı sunar.

#### 2. UYGULAMA

#### 2.1. Veri

Uygulama kısmına geçtiğimizde, projenin veri setini yüklemek için "CSV Reader" modülünü kullanacağız. Bu modül, telco veri setini CSV dosyası formatında okuyarak ve KNIME platformunda kullanılabilir bir veri tablosu oluşturacaktır.

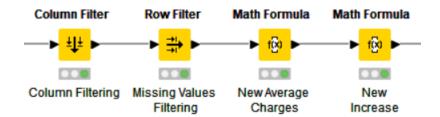


Row ID	S custom	S gender	SeniorC	tenure	S Partner	S Churn
Row0	7590-VHVEG	Female	0	1	Yes	No
Row1	5575-GNVDE	Male	0	34	No	No
Row2	3668-QPYBK	Male	0	2	No	Yes
Row3	7795-CFOCW	Male	0	45	No	No
Row4	9237-HQITU	Female	0	2	No	Yes
Row5	9305-CDSKC	Female	0	8	No	Yes
Row6	1452-KIOVK	Male	0	22	No	No
Row7	6713-OKOMC	Female	0	10	No	No
Row8	7892-POOKP	Female	0	28	Yes	Yes
Row9	6388-TABGU	Male	0	62	No	No
Row10	9763-GRSKD	Male	0	13	Yes	No
Row11	7469-LKBCI	Male	0	16	No	No
Row12	8091-TTVAX	Male	0	58	Yes	No
Row13	0280-XJGEX	Male	0	49	No	Yes
Row14	5129-JLPIS	Male	0	25	No	No
Row15	3655-SNQYZ	Female	0	69	Yes	No

Yapılan istatislik işlemi sonucunda sonucunda, 21 sütunun 17'sinin kategorik, 3'ünün nümerik, 1'inin kategorik ama kardınal (ordınal) ve 2'sinin nümerik ama kategorik (categorical) tipte olduğu görülmüştür.

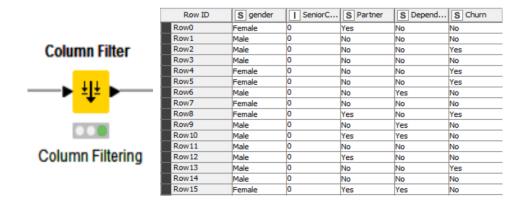
### 2.2. Ön İşleme

Veri setinizde anlamsız, gereksiz veya fazlalık olan nitelikleri temizlemek, veri ön işleme aşamasının önemli bir adımıdır. Bu adımda, veri setindeki gereksiz veya anlamsız nitelikleri filtreleyerek, modelin performansını artırabilir ve gereksiz bilgilerin modeli etkilemesini önleyebilirsiniz.

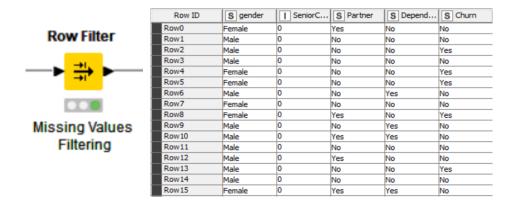


#### 2.2.1. Sütun ve satır filtreleme

Column Filter modülü, veri setindeki belirli sütunları filtrelemenize olanak sağlar. Bu modülü kullanarak, analiz için gereksiz veya anlamsız olduğunu düşündüğünüz 'customerID' sütununu çıkarıyoruz.

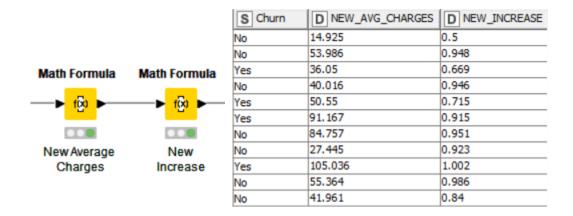


Row Filter modülü, veri setindeki belirli satırları filtrelemenize olanak tanır. Örneğin, veri setinizde boş veya eksik değerler içeren satırları çıkarabilir veya belirli bir sınıf etiketine sahip olan satırları filtreleyebiliriz. Veri setimizde yapılan istatistik analizi sonucunda "TotalCharges" sütununda 11 eksik değerli satır tespit edildi ve bu satırlar Row Filter modülü kullanılarak çıkarıldı. Bu adım, veri setinizin eksik değerlerden arındırılmasını sağlar ve modelinizin doğru ve güvenilir sonuçlar üretmesine yardımcı olur.



#### 2.2.2. Math Formula

Math Formula modülü kullanarak "\$TotalCharges\$ / (\$tenure\$ + 1)" formülüyle "NEW\_AVG\_CHARGES" adında ve "\$NEW\_AVG\_CHARGES\$ / \$MonthlyCharges\$" formülünü kullanarak "NEW\_INCREASE" adında 2 yeni sütunu veri setimize ekliyoruz.



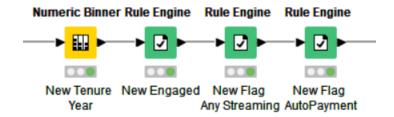
Yeni sütunlar, veri setimizdeki farklı değişkenler arasında ilişkileri incelemek veya yeni özellikler oluşturmak için kullanılacaktır.

#### 2.3. Dönüştürme

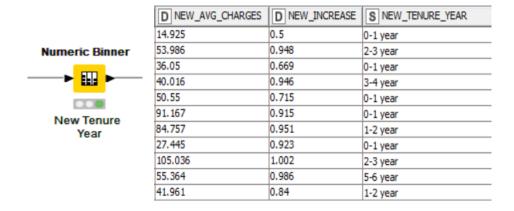
Dönüştürme aşamasında, veri setimizdeki sayısal niteliklerin bazılarında normalizasyon ve bazılarında sepetleme (binning) işlemlerini uygulayabiliriz. Bu işlemler, veri setimizin özelliklerini daha uygun hale getirerek modelinizin performansını artırabilir.

Normalizasyon, sayısal niteliklerin değerlerini belirli bir aralığa ölçeklendirme işlemidir. Bu, farklı ölçeklere veya birimlere sahip olan nitelikleri karşılaştırılabilir hale getirir. Normalizasyon işlemi genellikle veri setindeki tüm sayısal niteliklere uygulanır.

Sepetleme (binning), sayısal niteliklerin değerlerini belirli aralıklara bölme işlemidir. Bu, sürekli bir sayısal niteliği kategorik bir niteliğe dönüştürerek modelin daha iyi anlamasını sağlar. Sepetleme işlemi, bazı sayısal niteliklerde daha uygun olabilir ve modelin daha iyi sonuçlar üretmesine yardımcı olabilir.



#### 2.3.1. Numeric Binner



Numeric Binner modülünü kullanarak "tenure" niteliğini 6 ölçeğe ayırarak "NEW\_TENURE\_YEAR" adında yeni bir sütun oluşturulmuştur. Bu işlem, "tenure" niteliğini daha kategorik bir şekilde temsil etmek ve analiz için daha uygun hale getirmek amacıyla yapılır. Aşağıda oluşturulan ölçek verilmiştir.

#### 2.3.2. Rule Engine

Rule Engine kullanarak "NEW\_ENGAGED", "NEW\_FLAG\_ANY\_STREAMING" ve "NEW\_FLAG\_AutoPayment" adında üç yeni sütun oluşturabilirsiniz. Bu sütunlar, belirli kurallara dayalı olarak mevcut veri setinin niteliklerinden türetilir.

<sup>&</sup>quot;Expression" bölümüne girilen aşağıdaki kodlar;

Contract = "One year" => 1

\$Contract\$ = "Two year" => 1

TRUE => FALSE

"NEW\_ENGAGED" sütununu oluşturacak ve "Contract" niteliğine bağlı olarak değerler atayacaktır.

\$StreamingTV\$ = "Yes" => 1

\$StreamingMovies\$ = "Yes" => 1

TRUE => FALSE

Bu kod, "NEW\_FLAG\_ANY\_STREAMING" sütununu oluşturacak ve "StreamingTV" ve "StreamingMovies" niteliklerine bağlı olarak değerler atayacaktır.

\$PaymentMethod\$ IN ("Bank transfer (automatic)", "Credit card (automatic)") => 1

TRUE => FALSE

Kodu ile de, "NEW\_FLAG\_AutoPayment" sütununu oluşturacak ve "PaymentMethod" niteliğine bağlı olarak değerler atayacaktır.

Bu adımları takip ederek, belirli kurallara dayalı olarak "NEW\_ENGAGED", "NEW\_FLAG\_ANY\_STREAMING" ve "NEW\_FLAG\_AutoPayment" adında üç yeni sütunu veri setinize ekleyebilirsiniz. Bu sütunlar, veri setinizin belirli niteliklerine bağlı olarak 1 veya 0 değerlerini alacaktır.

#### 2.3.3. Category to Number modülü ile label encoding işlemi



Category to Number modülünü kullanarak Label Encoding işlemini gerçekleştiriyoruz. Bu işlem, Churn dışındaki tüm kategorik sütunları sayısal değerlere uygulanıp, dönüştürerek analiz için daha uygun hale getirecektir. Ayrıca, Label Encoding, sütundaki farklı kategorik değerlere sıralı sayısal değerler atayarak gerçekleştirilir. Bu yöntem, bazı makine öğrenimi algoritmalarında kullanılabilir, ancak sıralama ilişkisi yaratmaz. Eğer kategorik değerler arasında anlamlı bir sıralama ilişkisi varsa, bu durumda Ordinal Encoding veya One-Hot Encoding gibi diğer dönüşüm yöntemlerini kullanmak daha uygun olabilir.

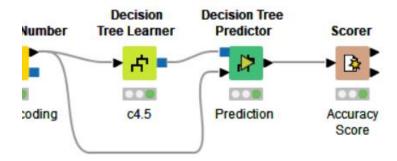
Partner	Depend	PhoneS	Multiple	Interne	OnlineS	OnlineB	Device	TechSu	Streami	Streami	Contra	Paperle	Payme	NEW_T
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	1	0	1	1	1	0	0	0	1	1	1	1
1	0	1	1	0	1	0	0	0	0	0	0	0	1	0
1	0	0	0	0	1	1	1	1	0	0	1	1	2	2
1	0	1	1	1	0	1	0	0	0	0	0	0	0	0
1	0	1	2	1	0	1	1	0	1	1	0	0	0	0
1	1	1	2	1	0	0	0	0	1	0	0	0	3	3
1	0	0	0	0	1	1	0	0	0	0	0	1	1	0
0	0	1	2	1	0	1	1	1	1	1	0	0	0	1
1	1	1	1	0	1	0	0	0	0	0	1	1	2	4
0	1	1	1	0	1	1	0	0	0	0	0	0	1	3

Yapılan işlemler sonucunda başlangıçta 21 olan sütun sayısı 42'ye çıkmıştır. Bu artış, veri setindeki önişleme ve dönüşüm adımlarıyla ilgili olarak gerçekleşmiştir. İşlem adımları sonucunda yeni sütunlar eklenmiş ve bazı sütunlar üzerinde dönüşümler yapılmıştır.

Önişleme ve dönüşüm adımları, veri setinin analiz veya makine öğrenimi modellerine uygun hale getirilmesi için yapılan işlemlerdir. Bu adımlar veri setinin temizlenmesi, eksik verilerin doldurulması, kategorik verilerin sayısal formata dönüştürülmesi, normalizasyon veya ölçeklendirme gibi işlemleri içerebilir. Bu adımlar genellikle veri setinin kalitesini artırır ve analiz veya modelleme için daha iyi sonuçlar elde etmeyi amaçlar.

Bu yeni sütunlar, veri setinin daha ayrıntılı bir temsilini veya analiz için daha fazla bilgiyi içerebilir. Her bir sütun, veri setinin farklı özelliklerini veya dönüşümlerini temsil edebilir. Bu ek sütunlar, veri setinin analiz veya modelleme sürecinde daha kapsamlı bir görüntü sağlamak veya daha iyi sonuçlar elde etmek için kullanılabilir.

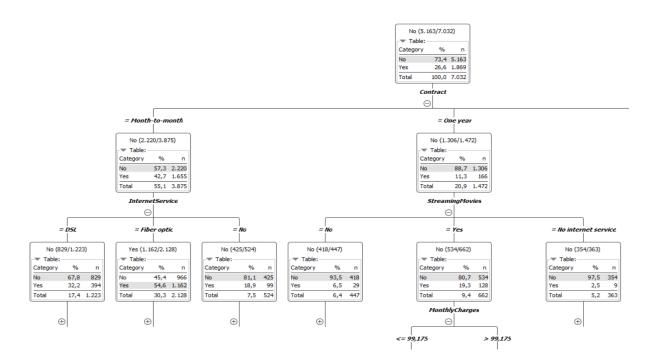
#### 2.4. Sınıflandırma Algoritmaları



Projenizde Knime kullanarak C4.5 decision tree algoritmasıyla telco churn tahmini yapmayı hedefliyorsunuz. Decision tree algoritması, verilerin ağaç yapısıyla temsil edildiği ve bir dizi karar kuralı kullanılarak sınıflandırma veya regresyon yapılmasına olanak sağlayan bir makine öğrenimi algoritmasıdır. Knime platformunda, decision tree modelinin oluşturulması, tahmin yapılması ve sonuçların değerlendirilmesi için özel modüller bulunmaktadır.

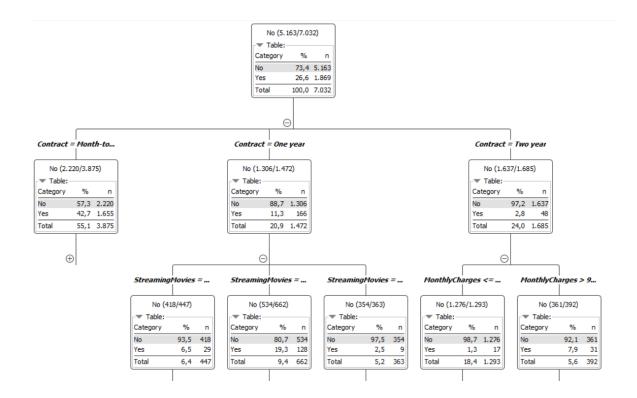
#### 2.4.1. Decision Tree Learner modülü:

Decision Tree Learner modülü, veri setiniz üzerinde decision tree modelini eğitmek için kullanılır. Bu modül, girdi olarak kategorik veya sayısal nitelikleri kabul eden bir veri seti alır. Algoritma, veri setindeki nitelikleri kullanarak karar kuralı oluşturur ve ağacın yapısını belirler. Özellikle C4.5 decision tree algoritmasını kullanıyorsanız, bu modül C4.5 algoritmasının özel bir uygulamasını içerir. Decision Tree Learner modülünü yapılandırarak, hedef değişkeninizi (churn) belirleyebilir ve eğitim parametrelerini ayarlayabilirsiniz.



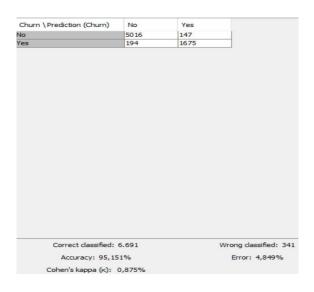
#### 2.4.2. Decision Tree Predictor modülü:

Decision Tree Predictor modülü, eğitilen decision tree modelini kullanarak tahmin yapmak için kullanılır. Bu modül, decision tree modelini ve tahmin yapılacak veri setini girdi olarak alır. Model, veri setindeki özniteliklere dayalı olarak tahminler üretir ve çıktı olarak tahmin edilen değerleri sağlar. Tahmin edilen değerler genellikle binary bir sınıflandırma olan churn durumunu (churn veya non-churn) içerir. Predictor modülü, decision tree modelinin kullanımını kolaylaştırır ve tahmin sonuçlarını diğer modüllerle birlikte kullanmanıza olanak tanır.



#### 2.4.3. Scorer modülü:

Scorer modülü, decision tree modelinin performansını değerlendirmek için kullanılır. Bu modül, tahmin edilen değerler ve gerçek değerler arasındaki karşılaştırmaları yapar ve farklı değerlendirme metriklerini hesaplar. Churn durumu üzerinde tahmin yapmak istediğimiz için, Scorer modülü churn etiketlerini ve tahmin edilen churn değerlerini karşılaştırarak doğruluk, hassasiyet, geri çağırma gibi metrikleri hesaplar. Bu metrikler, modelin performansını ölçmek ve geliştirmek için kullanılabilir.



Scorer modülü ile yapılan değerlendirmeye göre, projenizin sonuçları aşağıdaki gibi ortaya çıkmıştır:

Toplamda 7032 satır veriye sahibiz. 6691 satır doğru şekilde sınıflandırılmıştır (correct classified). Elde edilen doğruluk oranı (accuracy) %95,151'dir. Bu, modelin doğru sınıflandırma oranını gösterir. Cohen's Kappa değeri 0,875%'dir. Bu değer, modelin tahmin performansının rastlantısal performansa göre ne kadar iyilik gösterdiğini ölçer. Yüksek bir Kappa değeri, modelin iyi bir performansa sahip olduğunu gösterir. 341 satır yanlış şekilde sınıflandırılmıştır (wrong classified). Hata oranı (error) %4,849'dir. Bu değer, modelin yanlış sınıflandırma oranını temsil eder.

## 3. SONUÇ

Projemizin sonuçlarına göre, kullanmış olduğunuz önişleme adımları ve C4.5 decision tree algoritmasıyla başarılı bir telco churn tahmini gerçekleştirildi. Modelimizi değerlendirmek gerekirse:

Doğruluk (accuracy) oranı %95,151 olarak oldukça yüksektir, bu da modelimizin genel olarak doğru sınıflandırmalar yaptığını göstermektedir. Ancak, yanlış sınıflandırılan 341 satır ve %4,849 hata oranı dikkate alınmalıdır.

Modelimizi daha da iyileştirmek için aşağıdaki adımları uygulayabiliriz:

- 1- Daha fazla veri toplamak: Modelinizi daha güçlü hale getirmek için daha fazla veri toplamayı düşünebiliriz. Daha büyük bir veri seti, modelin genellemesini ve daha doğru tahminler yapmasını sağlayabilir.
- 2- Nitelik seçimi: Veri setinizdeki 21 sütun üzerinde yapılan analiz sonucunda, bazı niteliklerin gereksiz veya anlamsız olduğu tespit edilmiş olabilir. Nitelik seçimi yaparak, modelin daha iyi bir performans göstermesini sağlayabiliriz. İstatistiksel analiz ve özellik önem sıralaması gibi yöntemlerle, en önemli nitelikleri belirleyebilir ve modelimizi bu niteliklerle eğitebiliriz.
- 3- Model parametreleri ayarlama: Decision tree algoritmasının parametrelerini ayarlayarak modelinizi iyileştirebiliriz. Örneğin, ağacın derinliği, düğüm bölünmesi için kullanılan kriterler veya minimum örnekleme sayısı gibi parametreleri deneyebiliriz. Farklı parametre değerlerini kullanarak modelin performansını değerlendir ve en iyi sonuçları elde etmek için optimize edebiliriz.

- 4- Farklı algoritmalar deneme: C4.5 decision tree algoritması başarılı olsa da, farklı makine öğrenimi algoritmalarını da deneyebiliriz. Örneğin, random forest, gradient boosting veya support vector machines gibi farklı algoritmaların performansını değerlendirerek, daha iyi sonuçlar elde edebiliriz.
- 5- Hiperparametre optimizasyonu: Modelimizi daha iyi bir şekilde ayarlamak için hiperparametre optimizasyonu tekniklerini kullanabiliriz. Grid search veya random search gibi yöntemlerle farklı parametre kombinasyonlarını deneyerek, en iyi performansı elde etmeyi hedefleyebiliriz.

Sonuç olarak, projemizin sonuçları olumlu olsa da, modelinizi daha da geliştirmek için yukarıdaki önerilere dikkat etmemiz önemlidir. Veri setimizin niteliklerini ve model parametrelerini iyileştirmek, daha geniş bir veri seti toplamak ve farklı algoritmaları denemek, modelimizin performansını daha da artırabilir ve telco churn tahmininde daha doğru sonuçlar elde etmemizi sağlayabilir.