2019-2020

IST 4001 Bitirme Projesi

Sınıflandırma için Makine Öğrenimi

Hazırlayanlar

İbrahim Berkay ÖZ

Hüseyin Cem ARAS

Danışman: Doktor Öğretim Üyesi Özgül VUPA ÇİLENGİROĞLU

Çalışma Ortağı: SANEM Plastik Tasarım Merkezi Müdürü Alim Fatih KILINÇ

İçerik

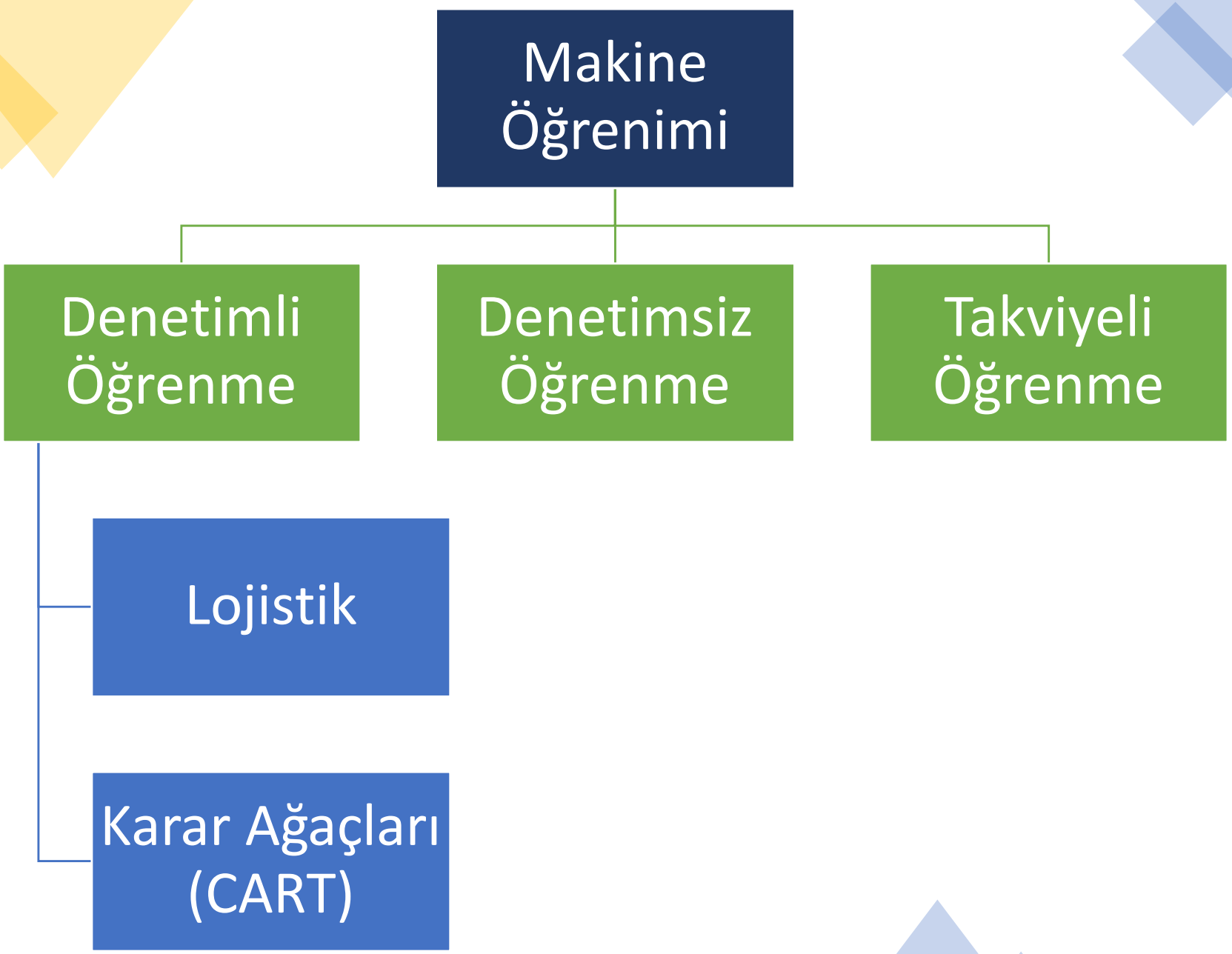
- Amaç
- Kullanılan Yöntemler
- Lojistik Regresyon Uygulaması
- CART Uygulaması
- Sonuç
- Kaynaklar

Amaç

- İzmir ili Torbalı ilçesinde plastik masa örtüsü üretimi gerçekleştiren SANEM PLASTİK firmasına ait bir fabrikadan alınan 2018-2019 yılları arasındaki veriler ile makine öğrenimi algoritmalarını kullanarak modeller oluşturup, modellere göre fabrikanın verimliliğini arttırmaktır.

Kullanılan Yöntemler

- Makine öğrenimini alt başlıkları içeresinde olan (Denetimli Öğrenme) sınıflandırma yöntemlerinden Lojistik Regresyon ve CART algoritmalarını kullanılmıştır.


Makine Öğrenimi
Denetimli Öğrenme
Denetimsiz Öğrenme
Takviyeli Öğrenme
Lojistik
Karar Ağaçları (CART)

Lojistik Regresyon

- Lojistik regresyon, normallik varsayımının bozulması nedeniyle doğrusal regresyon analizine alternatif olarak kullanılmaktadır.
- Doğrusal regresyon analizinde bağımlı değişkenin değeri, lojistik regresyon analizinde ise bağımlı değişkenin alabileceği değerlerden birinin gerçekleşme olasılığı tahmin edilmektedir.
- Temel olarak lojistik regresyonda bağımsız değişkenler ile iki ya da çok sınıflı kategorik bağımlı değişken arasındaki ilişkinin tanımlanması için matematiksel modelleme yapmak amaçlanmaktadır.

CART Algoritması

- CART algoritması, ağaç yapısına dayalı olarak sınıflandırma ve regresyon modellerinin türetilmesi için yaygın olarak kullanılan bir istatistiksel yöntemdir. CART ağaç modeli, tek değişkenli ikili kararların bir hiyerarşisini içerir.

- CART verileri iki alt kümeye ayırdığı için her bir alt küme içindeki durumlar, bir önceki alt kümeden daha homojen olacaktır. Bu ardışık süreç, homojenlik kriterine ulaşılıncaya veya diğer bazı durma kriterleri sağlanıncaya kadar kendini tekrar eder. Aynı kestirim değişkeni ağaçta farklı düzeylerde pek çok kez kullanılabilir. Ağacın yapısı önceden belirlenmemekte, verilerden türetilmektedir.

Uygulama

- İzmir ili Torbalı ilçesinde plastik masa örtüsü üretimi gerçekleştiren SANEM PLASTİK adlı fabrikadan alınan 2018-2019 yılları arasındaki veriler (n=198, train verisi=163) ile makine öğrenimi algoritmalarını kullanarak model oluşturulup, modele göre fabrikanın verimliliğini arttırmaktır.
- OEE, üretim, saat, kayıp metre, değişkenleri sürekli, hafta ve gün değişkenleri kategorik değişkenlerdir. Model için sürekli değişkenlerimiz, ortalamaya göre bölünüp 0&1 olarak kategorik değişken haline getirilmiştir.
- Modellerimizde OEE değişkeni bağımlı değişken iken diğer değişkenler bağımsız olarak ele alınmıştır.
- Modellerin amacı bağımsız değişkenlerin, OEE puanına nasıl etkileyeceğini araştırmaktır.

Değişkenlerin Tanımlayıcı İstatistikleri

```
     uretim
Min.    : 4750
1st Qu.:49889
Median :80145
Mean   :68018
3rd Qu.:86760
Max.   :94680
```

```
      OEE
Min.    : 29.00
1st Qu.: 82.00
Median : 89.00
Mean   : 85.15
3rd Qu.: 93.75
Max.   :100.00
```

```
      saat
Min.    : 0.200
1st Qu.: 1.000
Median : 1.950
Mean   : 2.576
3rd Qu.: 3.000
Max.   :23.400
```

```
 kayip_metre          gun    tecrube
Min.    : 1021    1:27   0: 83
1st Qu.: 3806    2:36   1:115
Median : 6943    3:38
Mean   : 8728    4:38
3rd Qu.:11651    5:34
Max.   :36471    6:25
```

NOT: Tecrübe değişkeni simülasyon ile elde edilmiştir.

Kategorik Değişkenlerin Tanımlayıcı İstatistikleri

uretim	OEE	saat	kayip_metre	gun	tecrube
0: 72	0: 68	0:132	0:124	1:27	0: 83
1:126	1:130	1: 66	1: 74	2:36	1:115
				3:38	
				4:38	
				5:34	
				6:25	

NOT: Ortalamanın altında olanlar 0, eşit ve üstünde olanlar 1 değerini almıştır.

Lojistik Regresyon Modeli (Full Model)

- Veri setimizin bütün değişkenler ile oluşturulan full Lojistik Regresyon modeli oluşturulmuştur.
- Çıktıda görüldüğü üzere üretim, saat anlamlı çıkmıştır. Kayıp metre değişkeni 0.05 sınırında olduğundan ‘.’ ile işaretlenmiştir.
- Yapılan Korelasyon analizine göre kayıp metre ve saat değişkeni ile ilişkili bulunmuştur (0.48). Bu yüzden kayıp metrenin modele alınmamasına karar verilmiştir.

```
> full.model = glm(OEE ~. - tecrube ,family=binomial, data=train)
> summary(full.model)

Call:
glm(formula = OEE ~ . - tecrube, family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7868  -0.4047   0.2040   0.6068   1.4250

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.14138    0.54440  -0.260  0.79509
uretim1       3.90538    0.80857   4.830 1.37e-06 ***
saat1        -2.21958    0.86006  -2.581  0.00986 **
kayip_metre1 -1.90456    1.00137  -1.902  0.05718 .
gun2          0.09838    0.73493   0.134  0.89351
gun3         -0.20531    0.75528  -0.272  0.78575
gun4         -0.03542    0.73268  -0.048  0.96144
gun5          1.07417    0.82050   1.309  0.19048
gun6          1.56849    0.87277   1.797  0.07231 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 214.49  on 162  degrees of freedom
Residual deviance: 120.85  on 154  degrees of freedom
AIC: 138.85

Number of Fisher Scoring iterations: 6
```

Lojistik Regresyon Modeli (En İyi Model)

- En iyi modeli oluşturmak için stepwise yöntemi kullanılmıştır.
- Bu durumda en iyi modelde üretim, saat değişkenleri %95 anlamlılık düzeyinde istatistiksel olarak anlamlı çıkmıştır.
- Daha sonra OR, CI ve Wald testine bakılmıştır.

```
> full.model = glm(OEE ~. - tecrube -kayip_metre ,family=binom
ial, data=train)
> best.model <- stepAIC(full.model, direction = "both",trace =
 FALSE)
> summary(best.model)

Call:
glm(formula = OEE ~ uretim + saat, family = binomial, data = t
rain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5993  -0.2757   0.2635   0.6622   1.2583

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.2805     0.3149   0.891    0.373
uretim1       3.0630     0.6392   4.792 1.65e-06 ***
saat1        -3.5317     0.6397  -5.521 3.38e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 214.49  on 162  degrees of freedom
Residual deviance: 131.67  on 160  degrees of freedom
AIC: 137.67
```

Lojistik Regresyon Modeli (En İyi Model)

- Kategorik değişkenlerle kurulan lojistik regresyon en iyi modelindeki Wald değerlerine bakıldığında, tüm p değerleri 0.05'den küçük olduğu için katsayılar %95 güven seviyesinde istatistiksel olarak anlamlı bulunmuştur.

```
> wald.test(b=coef(best.model),Sigma=vcov(best.model),Terms=1:2)
Wald test:
----------

Chi-squared test:
X2 = 31.2, df = 2, P(> X2) = 1.6e-07
```

Lojistik Regresyon Modeli (En İyi Model)

- Ayrıca OR ve güven aralıklarına bakıldığında, OR'a ait güven aralıklarının 1'i içermediği görülmektedir. Bu yüzden OR'ler anlamlıdır ve yorumlanabilir.

```
> exp(confint(best.model))
Waiting for profiling to be done...
                    2.5 %      97.5 %
(Intercept) 0.715917230  2.48285748
uretim1     7.032365088 93.58095767
saat1       0.006676091  0.08888818
```

Lojistik Regresyon Modeli (En İyi Model)

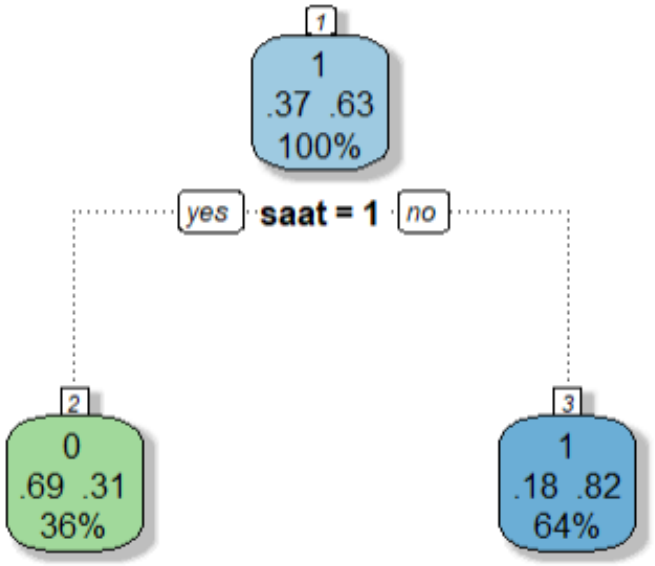
- Üretimi yüksek olanların, üretimi düşük olanlara göre Toplam Ekipman Etkinliğinin (OEE) yüzdesinin yüksek olma şansı yaklaşık olarak 21 katıdır.
- Saat değişkeninde ise saati az olanların, yüksek olanlara göre OEE yüzdesinin yüksek olma şansı ise 1/0.29 oranından yaklaşık olarak 3 katı olduğu elde edilmektedir.

```
> exp(coef(best.model))
(Intercept)      uretim1        saat1
 1.32378898 21.39165960  0.02925425
```

CART Algoritması (En İyi Model)

- CART ağacımız Gini katsayısına göre hesaplanıp, ilk olarak saat kategorik değişkenine ayrılmıştır.

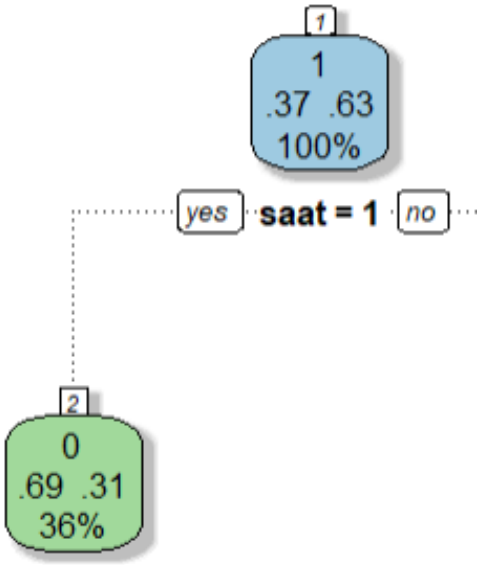
- Bu durumda OEE'nin yorumlanmasında en önemli değişken olarak saat değişkeni olduğu söylenebilir.



CART Algoritması (En İyi Model)

- Buna göre saati ortalama veya ortalamadan yüksek olanlar, sol tarafa ayrılmıştır. 163 birimlik "train" verisi üzerinde hesaplandığı için bu düğüme %36 oranla 59 birim düşmüştür.

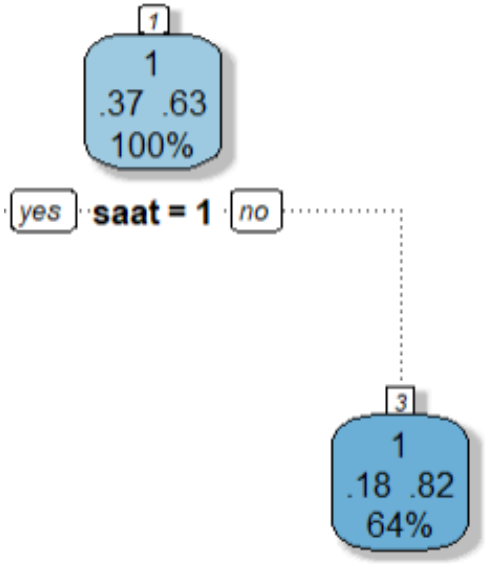
- 59 birimin %69'u yani 41 tanesi OEE'si 0 olan sınıfta yer almıştır. Kalan 18 tanesi ise OEE'si 1 olan sınıfta yer almıştır.



CART Algoritması (En İyi Model)

- Saati ortalamadan az olanlar ise, sağ tarafa ayrılmıştır. 163 birimin %64 oranla 104 birim bu düğüme düşmüştür.

- 104 birimin %18'i yani 19 tanesi OEE'si 0 olan sınıfta yer almıştır. Kalan 85 tanesi ise OEE'si 1 olan sınıfta yer almıştır.



YENİ MODEL

- Fabrika verileri ile kurulan lojistik regresyon modelinde saat ve üretim değişkenleri modelde anlamlı çıkmıştır. CART modelinde ise en önemli değişkenin saat olduğu tespit edilmiştir. Daha sonra CART modelinde ağaç derinliği kontrol edildiğinde 2. düğümün üretim değişkenine ayrıldığı gözlemlenmiştir. Fakat üretim değişkeni tek başına yeterli ayrıştırmayı yapamadığından ağaçta kırpılmıştır.
- Ancak modellemede değişken sayısını artırıp modelin nasıl değiştiğini görmek için verilerimize simülasyon çalışması ile makineyi çalıştıran kişinin tecrübesi (10 yıldan az&10 yıl ve daha fazla olarak belirlenmiş kategorik değişken) eklenmiştir.
- Bu yeni değişkene göre modeller kurulmuştur.

Lojistik Regresyon Modeli (Tecrübe, Üretim, Saat (TÜS))

- Burada en uygun modelde tecrübe, üretim, saat değişkenleri %95 anlamlılık düzeyinde istatistiksel olarak anlamlı çıkmıştır.
- Daha sonra OR, CI ve Wald testine bakılmıştır.

```
> model2 = glm(OEE ~uretim+ saat + tecrube ,family=binomia
l, data=train)
> summary(model2)

Call:
glm(formula = OEE ~ uretim + saat + tecrube, family = binom
ial,
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1667  -0.3807   0.1177   0.5445   1.9852

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.2441     0.4895  -2.541    0.011 *
uretim1       3.4907     0.7087   4.926 8.41e-07 ***
saat1        -4.0670     0.7303  -5.569 2.56e-08 ***
tecrube1      2.7232     0.5543   4.913 8.98e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
 ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 214.49  on 162  degrees of freedom
Residual deviance: 100.22  on 159  degrees of freedom
AIC: 108.22
```

Lojistik Regresyon Modeli (Tecrübe, Üretim, Saat (TÜS))

- Kategorik değişkenlerle kurulan en uygun lojistik regresyon modelindeki Wald değerlerine bakıldığında, tüm p değerleri 0.05'den küçük olduğu için katsayılar %95 güven seviyesinde istatistiksel olarak anlamlı bulunmuştur.

```
> wald.test(b=coef(model2),Sigma=vcov(model2),Terms=1:2)
Wald test:
----------

Chi-squared test:
X2 = 24.3, df = 2, P(> X2) = 5.4e-06
```

Lojistik Regresyon Modeli (Tecrübe, Üretim, Saat (TÜS))

- Ayrıca OR ve güven aralıklarına bakıldığında, OR'a ait güven aralıklarının 1'i içermediği görülmektedir. Bu yüzden OR'ler anlamlıdır.

```
> exp(confint(model2))
Waiting for profiling to be done...
                   2.5 %       97.5 %
(Intercept) 0.102099647   0.71419154
uretim1     9.408205782 161.53474388
saat1       0.003329586   0.06180391
tecrube1    5.466755363  48.97582068
```

Lojistik Regresyon Modeli (Tecrübe, Üretim, Saat (TÜS))

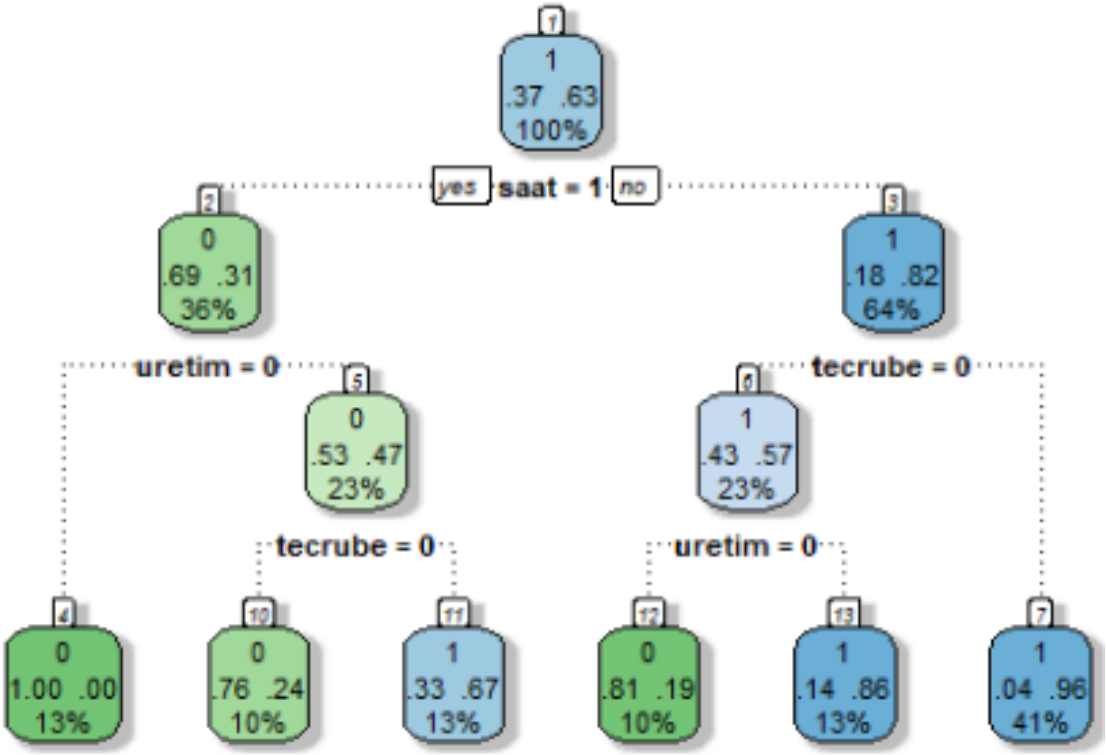
- 10 Yıllık Tecrübesi olanların, olmayanlara göre OEE yüzdesinin yüksek olma şansı yaklaşık olarak 15 kattır.
- Üretimi yüksek olanların, üretimi düşük olanlara göre OEE yüzdesinin yüksek olma şansı yaklaşık olarak 33 kattır.
- Saati az olanların saati yüksek olanlara göre OEE yüzdesinin yüksek olma şansı ise 1/0.17 yaklaşık olarak 6 kattır.

```
> exp(coef(model2))
(Intercept)     uretim1       saat1     tecrube1
 0.28821447 32.81030772  0.01712815  15.22833475
```

CART Algoritması (Tecrübe, Üretim, Saat, (TÜS))

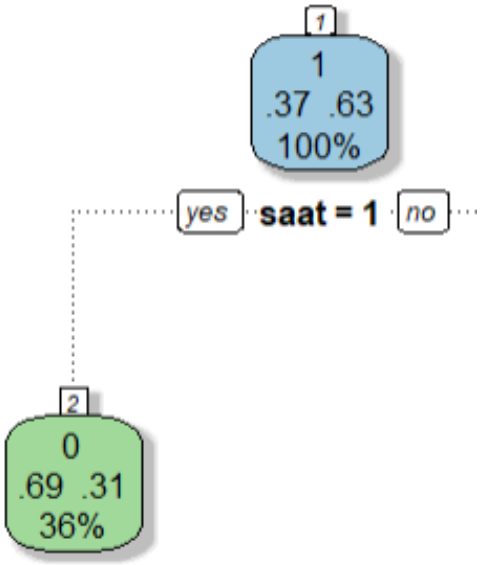
- CART ağacımız Gini katsayısına göre hesaplanıp, ilk olarak saat kategorik değişkeni ile ayrılmıştır.

- Bu durumda OEE'nin yorumlanmasında en önemli değişkenin saat olduğu söylenebilir.



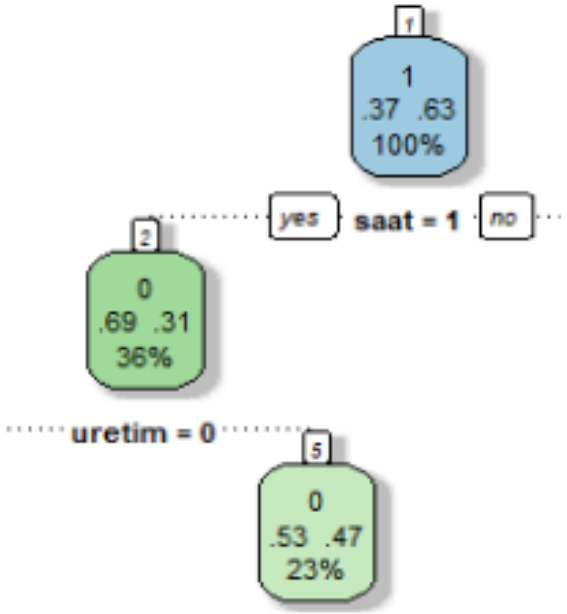
CART Algoritması (Tecrübe, Üretim, Saat, (TÜS))

- Buna göre saati ortalama veya ortalamadan yüksek olanlar, sol tarafa ayrılıp, 2. Düğümde yer almıştır.
- 163 birimlik “train” verisi üzerinde hesaplandığı için bu 2. düğüme %36 oranla 59 birim düşmüştür.
- 59 birimin %69’u yani 41 tanesi OEE’si 0 olan sınıfta yer almıştır. Kalan 18 tanesi ise OEE’si 1 olan sınıfta yer almıştır.



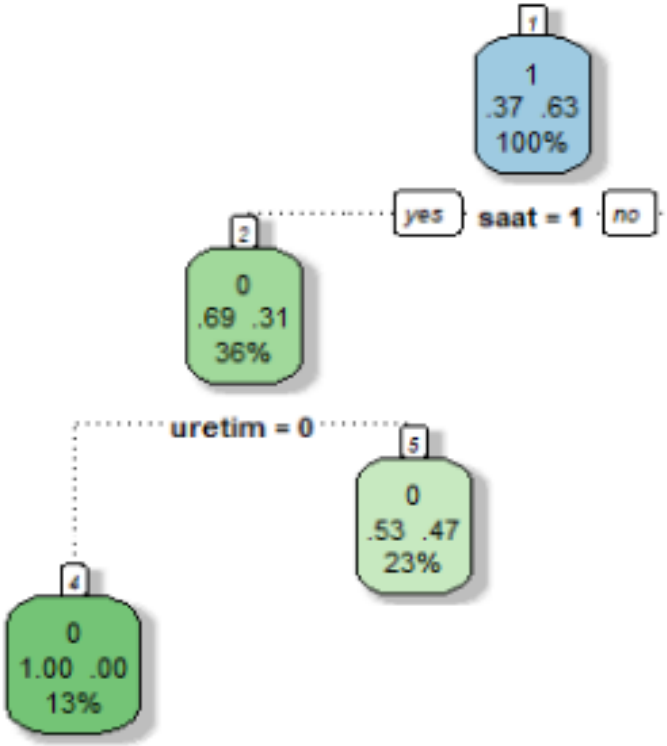
CART Algoritması (Tecrübe, Üretim, Saat, (TÜS))

- 2. düğüm daha sonra üretim değişkenine göre ikiye ayrılmıştır.
- Üretimi ortalama ya da ortalamadan yüksek olanlar sağ tarafa ayrılıp 5. düğümü oluşturmuşlardır. 5. düğüme toplam verinin %23'üne denk gelen 38 birim düşmüştür.
- Bu 38 birimin %53'ü yani 20 tanesi OEE'si 0 olan sınıfta yer almıştır. Kalan 18 tanesi ise OEE'si 1 olan sınıfta yer almıştır.



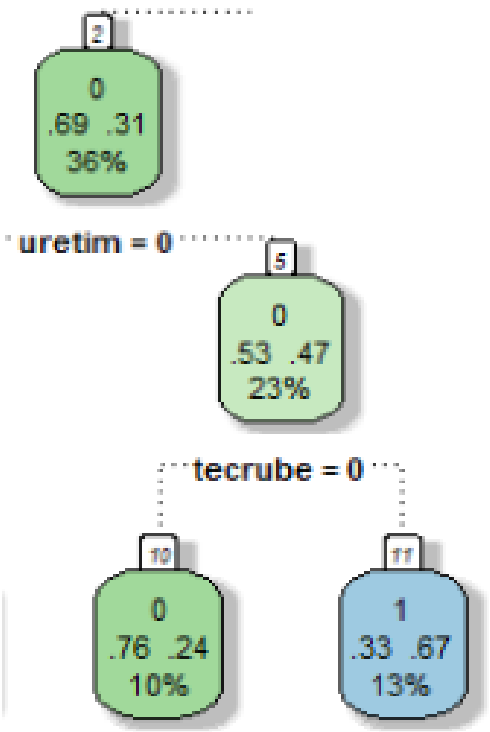
CART Algoritması (Tecrübe, Üretim, Saat, (TÜS))

- 2. düğümde üretimi ortalamadan az olanlar sol tarafa ayrılıp 4. düğümü oluşturmuşlardır.
- 4. düğüme toplam verinin %13'üne denk gelen 21 birim düşmüştür.
- Bu 21 birimin %100'ü yani tamamı OEE'si 0 olan sınıfta yer almıştır. OEE'si 1 olan sınıfta yer alan birim yoktur.



CART Algoritması (Tecrübe, Üretim, Saat, (TÜS))

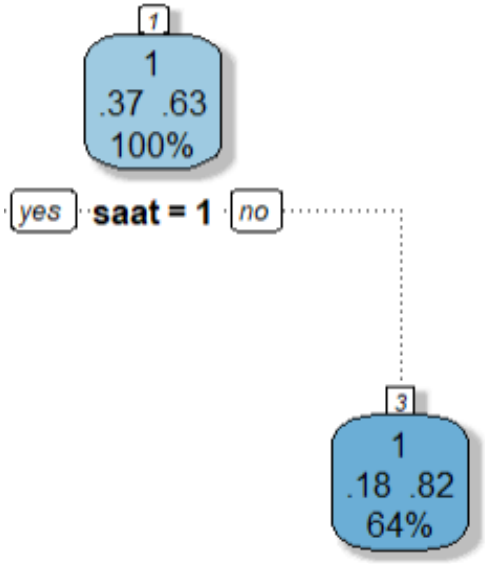
- 5. düğüm daha sonra tecrübe değişkenine göre ikiye ayrılmıştır.
- On yıllık tecrübesi olanlar sol tarafa ayrılıp 10. düğümü, olmayanlar sağ tarafa ayrılıp 11. düğümü oluşturmuşlardır.
- 10. düğüme toplam verinin %10'una denk gelen 16 birim düşmüştür. Bu 16 birimin %76'sı yani 12 tanesi OEE'si 0 olan sınıfta yer almıştır. Kalan 4 tanesi ise OEE'si 1 olan sınıfta yer almıştır.
- 11. düğüme ise toplam verinin %13'üne denk gelen 21 birim düşmüştür. Bu 21 birimin %33'ü yani 7 tanesi OEE'si 0 olan sınıfta yer almıştır. Kalan 14 tanesi ise OEE'si 1 olan sınıfta yer almıştır.



CART Algoritması (Tecrübe, Üretim, Saat, (TÜS))

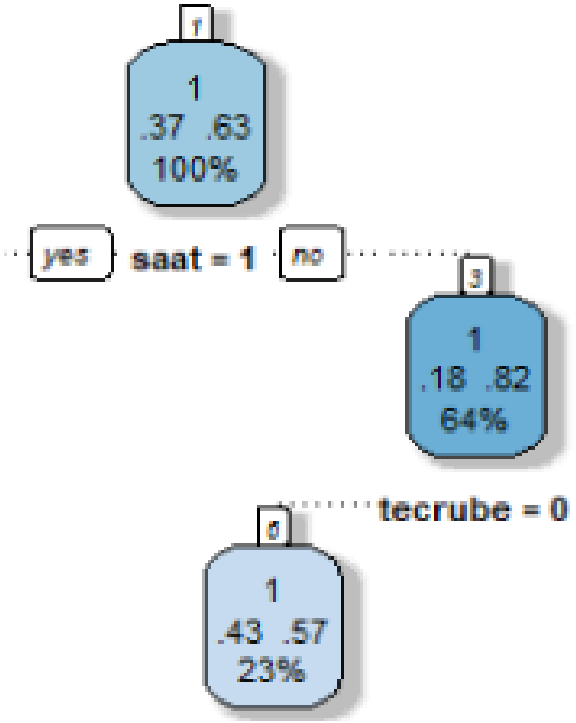
- Saati ortalamadan az olanlar ise, sağ tarafa ayrılmıştır. 163 birimin %64 oranla 104 birim bu düğüme yani 3. düğüme düşmüştür.

- 104 birimin %18'i yani 19 tanesi OEE'si 0 olan sınıfta yer almıştır. Kalan 85 tanesi ise OEE'si 1 olan sınıfta yer almıştır.



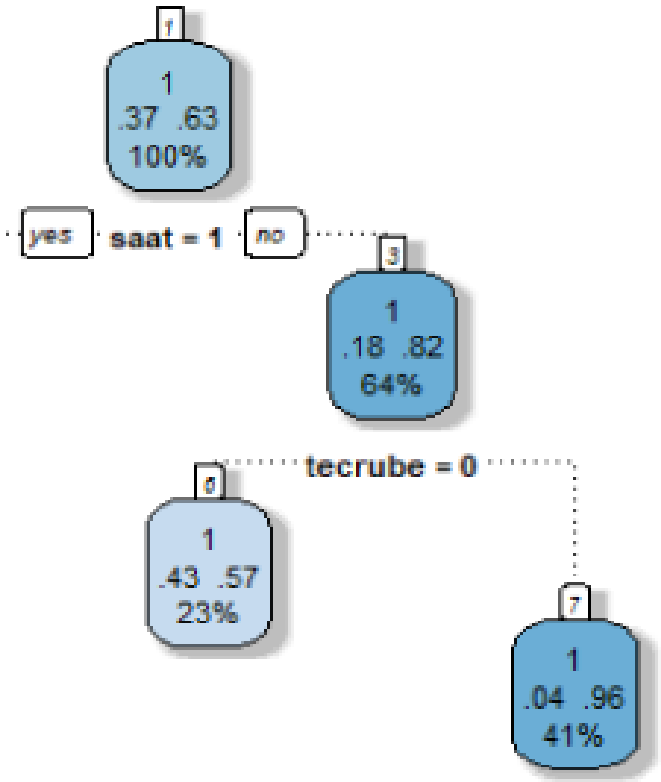
CART Algoritması (Tecrübe, Üretim, Saat, (TÜS))

- 3. Düğüm daha sonra tecrübe değişkenine göre ayrılmıştır.
- On yıllık tecrübesi olmayanlar sol tarafa ayrılıp 6. düğümü oluşturmuştur.
- 6. düğüme toplam verinin %23'üne denk gelen 38 birim düşmüştür. Bu 38 birimin %43'ü yani 16 tanesi OEE'si 0 olan sınıfta yer almıştır. Kalan 22 tanesi ise OEE'si 1 olan sınıfta yer almıştır.



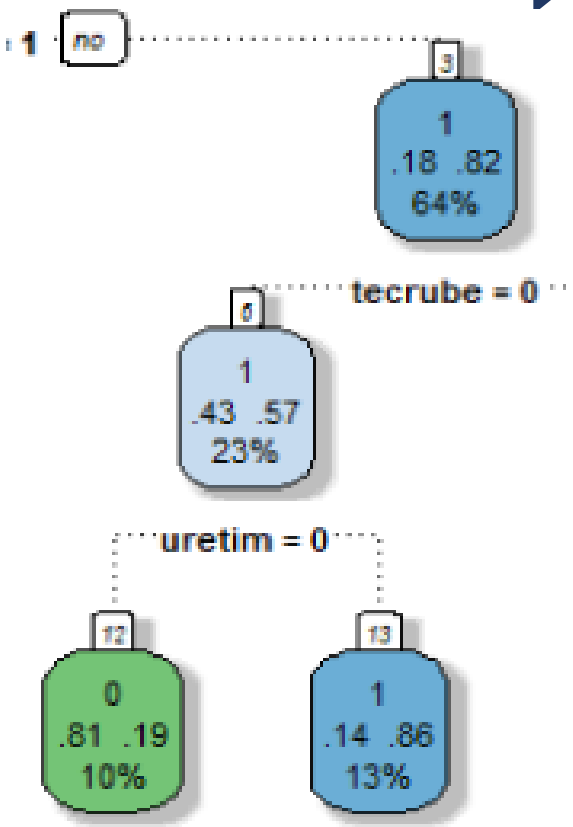
CART Algoritması (Tecrübe, Üretim, Saat, (TÜS))

- 3. düğümdeki on yıllık tecrübesi olanlar sağ tarafa ayrılıp 7. düğümü oluşturmuşlardır.
- 7. düğüme toplam verinin %41'ine denk gelen 67 birim düşmüştür.
-
- Bu 67 birimin %4'ü yani 3 tanesi OEE'si 0 olan sınıfta yer almıştır. Kalan 64 tanesi ise OEE'si 1 olan sınıfta yer almıştır.



CART Algoritması (Tecrübe, Üretim, Saat, (TÜS))

- 6. düğüm daha sonra üretim değişkenine göre ikiye ayrılmıştır.
- Üretimi az olanlar sol tarafa ayrılıp 12. düğümü, çok olanlar sağ tarafa ayrılıp 13. düğümü oluşturmuşlardır.
- 12. düğüme toplam verinin %10'una denk gelen 16 birim düşmüştür. Bu 16 birimin %81'i yani 13 tanesi OEE'si 0 olan sınıfta yer almıştır. Kalan 3 tanesi ise OEE'si 1 olan sınıfta yer almıştır.
- 13. düğüme ise toplam verinin %13'üne denk gelen 21 birim düşmüştür. Bu 21 birimin %14'ü yani 3 tanesi OEE'si 0 olan sınıfta yer almıştır. Kalan 18 tanesi ise OEE'si 1 olan sınıfta yer almıştır.



Sonuç

- İzmir ili Torbalı ilçesinde plastik masa örtüsü üretimi gerçekleştiren SANEM Plastik adlı fabrikadan 2018-2019 yılları arasında alınan 198 veri ile makine öğrenimi algoritmalarını kullanarak saat ve üretim değişkenlerinin OEE puanı üzerindeki etkileri, lojistik regresyon ve CART algoritmaları üzerinden modeller kurulmuştur.
- Kurulan lojistik regresyon modelinde saat ve üretim değişkenleri modelde anlamlı çıkmıştır. CART modelinde ise en önemli değişkenin saat olduğu tespit edilmiştir.
- Modellerimizin iki büyük kısıtlaması vardır. Biri değişken sayısının diğeri ise örneklem sayısının azlığıdır. Bu amaçla modellemede değişken sayısını artırıp modelin nasıl değiştiğini görmek için verilerimize simülasyon çalışması ile makineyi çalıştıran kişinin tecrübesi (10 yıldan az & 10 yıl ve daha fazla olarak belirlenmiş kategorik değişken) eklenmiştir.

Sonuç

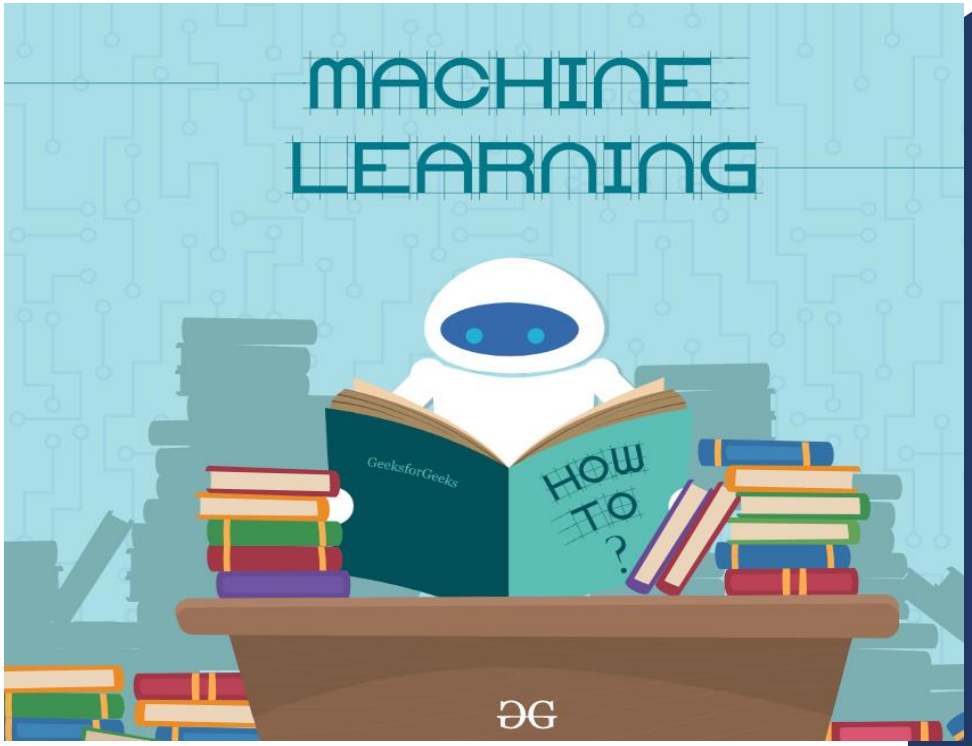
- Son eklenen değişken ile tecrübe, üretim, ve saat değişkenleri lojistik regresyon modelimizde anlamlı, CART modelinde ise yine en önemli değişkenin saat olduğu, daha sonra tecrübe en son olarak da üretim olduğu bulunmuştur.
- Bu çalışma ile üretim yapan fabrikaların verisinde makine öğrenimi kapsamında yer alan sınıflandırma modellerinden lojistik regresyon ve CART modellerinin uygulanabilirliği tespit edilmiştir. Ayrıca bu tip verilerde değişik sınıflandırma algoritmaları da kullanılabilir ve algoritmaların performansları da karşılaştırılabilir.
- Fabrikadaki uzmanlarla yapılan ortak çalışmalarla bu tespitin farklı veri türlerinde ve farklı modellerle yapılması planlanmalıdır. Böylece sanayi-üniversite işbirliği gerçekleşmesi sağlanabilir.

Teşekkür

- Bu çalışma kapsamında sanayi kolumuz olan SANEM Plastik AŞ. Yönetim Kuruluna ve Tasarım Merkezi Müdürü Alim Fatih KILINÇ'a sanayi-üniversite işbirliği kapsamında bize veri ve bilgi desteği sağladıkları için sonsuz teşekkürlerimizi sunarız.

Kaynakça

- Atalay M., Çelik E., 2017, Artificial İntelligence And Machine Learning Applications İn Big Data Analysis
- Özkaya A., 2012, Makine Öğrenmesi İle Ürün Sınıflandırma İncelemesi
- KALAYCI E., 2018, Comparison of machine learning techniques for classification of phishing web sites
- Abbas G., Shirali M., Moloud V., Amal S., 2018, Predicting the outcome of occupational accidents by CART and CHAID methods at a steel factory in Iran
- Yan-yan S., Ying L., 2015, Decision tree methods: applications for classification and prediction
- Antipov E., Pokryshevskaya E., 2009, Applying CHAID for logistic regression diagnostics and classification accuracy improvement



Teşekkür Ederiz