



BURSA TEKNİK ÜNİVERSİTESİ
MÜHENDİSLİK FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ

VERİ MADENCİLİĞİNE GİRİŞ

Bank Marketing Veri Seti
Rule-Based Methods

20360859045

BURAK ASLAN

2024

İÇİNDEKİLER

- 1. Veri Setinin Amacı**
- 2. Değişkenler Tablosu**
- 3. Kullanılan Sınıflandırma Metodu**
- 4. Orange Programı ile Oluşturulan Şema**
- 5. CN2 Rule ile Oluşturulan Kural Tablosu**
- 6. ROC Eğrisi**
 - 6.1. ROC Eğrileri Karşılaştırması (Decision Tree vs CN2 Rule)**
- 7. Test and Score Tablo**
- 8. Confusion Matrix Tablosu**
 - 8.1. Performans Karşılaştırması**
- 9. Gain Rati ve Gini Değerleri**
- 10. Indirect Method**
 - 10.1. Karar Ağaçlarından Yararlanarak Kural Oluşturma**
- 11. Veri Seti Üzerinden Yapılan Akademik Çalışma**
- 12. Kaynakça**

1. Veri Setinin Amacı

Veriler, bir Portekiz bankacılık kurumunun doğrudan pazarlama kampanyaları ile ilgilidir. Pazarlama kampanyaları telefon aramalarına dayanmakta. Genellikle, bir müşteriye birden fazla kez ulaşılması gerekti ve bu sayede ürünün (banka vadeli mevduat) abonelik durumu ('evet') veya ('hayır') olup olmayacağı tespit edildi.

Örneklerin %10' unu ve 16 öznitelik içeren "**bank.csv**" dosyasından rastgele seçilmiştir.

Sınıflandırma hedefi, müşterinin bir vadeli mevduat (**y değişkeni**) aboneliği yapip yapmayacağını (evet/hayır) tahmin etmektir.

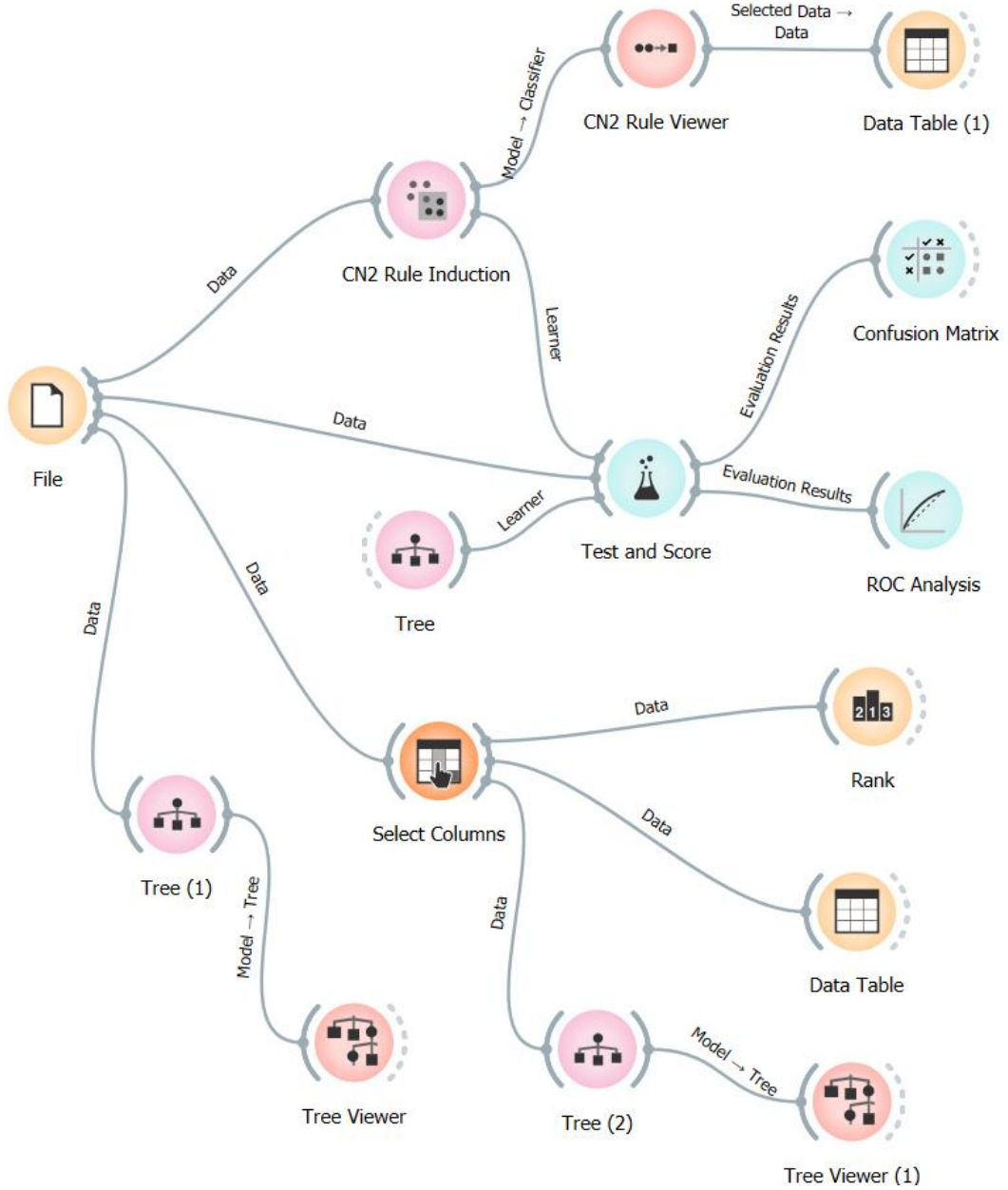
2. Değişkenler Tablosu:

Variable Name	Role	Type	Demographic	Description
age	Feature	Integer	Age	
job	Feature	Categorical	Occupation	type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','manager','employed','services','student','technician','unemployed','unknown')
marital	Feature	Categorical	Marital Status	marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' widowed)
education	Feature	Categorical	Education Level	(categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
default	Feature	Binary		has credit in default?
balance	Feature	Integer		average yearly balance
housing	Feature	Binary		has housing loan?
loan	Feature	Binary		has personal loan?
contact	Feature	Categorical		contact communication type (categorical: 'cellular','telephone')
day_of_week	Feature	Date		last contact day of the week
month	Feature	Date		last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
duration	Feature	Integer		last contact duration, in seconds (numeric). Important note: this attribute highly target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call and thus, if the duration is 0, it is the last contact of the campaign. After the end of the call y is obviously known. Thus, this input should only be used for purposes and should be discarded if the intention is to have a realistic prediction model.
campaign	Feature	Integer		number of contacts performed during this campaign and for this client (numeric, equals to 1 for each instance)
pdays	Feature	Integer		number of days that passed by after the client was last contacted from a previous time (numeric; -1 means client was not previously contacted)
previous	Feature	Integer		number of contacts performed before this campaign and for this client (numeric)
previous_outcome	Feature	Categorical		outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
y	Target	Binary		has the client subscribed a term deposit?

3. Kullanılan Sınıflandırma Metodu

Sınıflandırma kuralları çıkarımı için Direct Methodlardan CN2 Metodunu kullandım. (Rule Based Methods)

4. Orange Programı ile Oluşturulan Şema



5. CN2 Rule ile Oluşturulan Kural Tablosu:

	IF conditions	THEN class	Distribution	Probabilities [%]
0	duration≤76.0 AND job=blue-collar →	y=no	[154, 0]	99 : 1
1	duration≤76.0 AND job=entrepreneur →	y=no	[21, 0]	96 : 4
2	duration≤76.0 AND job=housemaid →	y=no	[14, 0]	94 : 6
3	duration≤76.0 AND job=management →	y=no	[150, 0]	99 : 1
4	duration≤91.0 AND age≥54.0 →	y=no	[92, 0]	99 : 1
5	duration≤212.0 AND job=entrepreneur →	y=no	[74, 0]	99 : 1
6	duration≤212.0 AND job=unemployed →	y=no	[61, 0]	98 : 2
7	duration≤212.0 AND default≠no →	y=no	[37, 0]	97 : 3
8	contact=unknown AND job=student →	y=no	[13, 0]	93 : 7
9	contact=unknown AND job=unemployed →	y=no	[18, 0]	95 : 5
10	contact=unknown AND job=unknown →	y=no	[6, 0]	88 : 12
11	contact=unknown AND balance≥5366.0 →	y=no	[55, 0]	98 : 2
12	contact=unknown AND balance≥5346.0 →	y=yes	[0, 1]	33 : 67
13	contact=unknown AND month=jan →	y=no	[1, 0]	67 : 33
14	contact=unknown AND month=nov →	y=yes	[0, 2]	25 : 75
15	contact=unknown AND campaign≥10.0 →	y=no	[36, 0]	97 : 3
16	contact=unknown AND poutcome=sucess →	y=no	[1, 0]	67 : 33

Compact view Restore original order

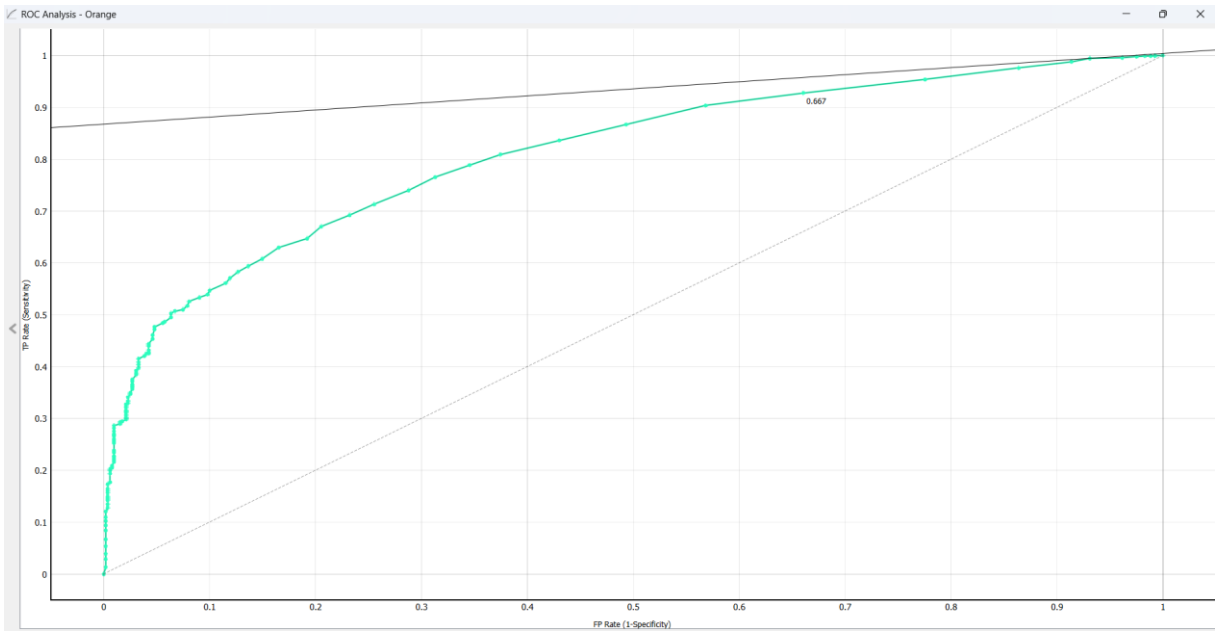
13 | 4521

CN2 ile oluşturulan bazı kurallar. Distribution sınıf dağılımını göstermekte. Probabilities olasılık değerlerini göstermekte. Seçili kuralda %93,3 no sınıfının, %6,7 yes sınıfının seçildiği gösterilmekte.

“(contact=unknown) AND (job=student) → no” kuralının tablo üzerinden gösterimi

	y	contact	job	duration	month	poutcome	loan	education	marital
1	no	unknown	student	230	may	unknown	no	secondary	single
2	no	unknown	student	197	jun	unknown	no	secondary	married
3	no	unknown	student	115	jun	unknown	no	unknown	married
4	no	unknown	student	289	may	unknown	no	tertiary	single
5	no	unknown	student	332	may	unknown	no	secondary	single
6	no	unknown	student	6	aug	unknown	no	secondary	single
7	no	unknown	student	198	may	unknown	no	unknown	single
8	no	unknown	student	227	may	unknown	no	unknown	single
9	no	unknown	student	133	may	unknown	no	secondary	single
10	no	unknown	student	28	jun	unknown	no	secondary	married
11	no	unknown	student	622	may	unknown	no	secondary	single
12	no	unknown	student	95	jun	unknown	no	secondary	single
13	no	unknown	student	359	jun	unknown	no	unknown	single

6. ROC Eğrisi:

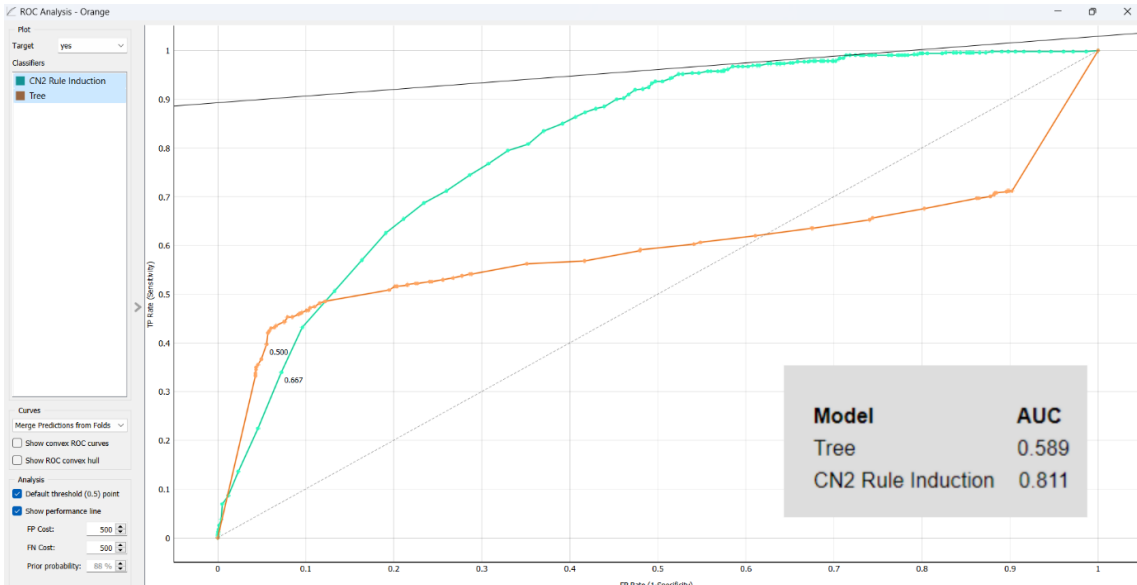


ROC eğrisi, TP oranını (y ekseninde) FP oranına (x ekseninde) karşı karakterize eder.

İyi bir sınıflandırıcı sol uç köşeye mümkün olduğunca yakın olması gerekir.

ROC eğrisinin altında kalan alan (AUC) CN2 Rule Induction'a göre 0,811 bu yeterince iyi bir sınıflandırma olduğunu gösteriyor.

6.1. ROC Eğrileri Karşılaştırması (Decision Tree vs CN2 Rule)

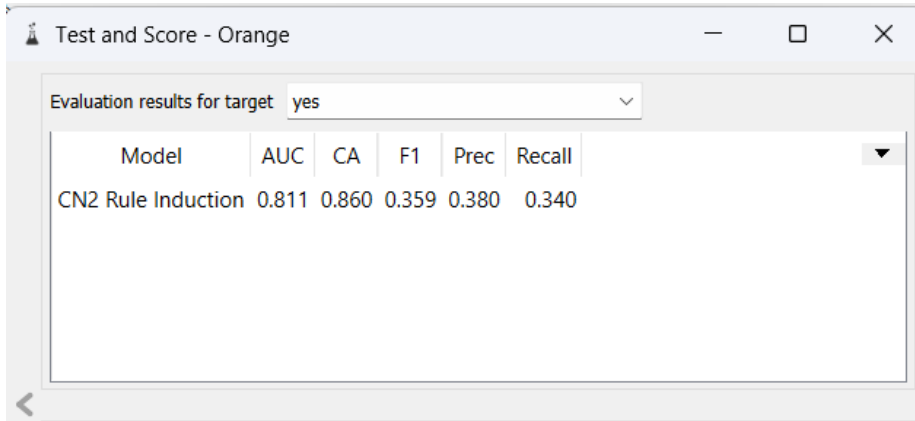


■ CN2 Rule Induction

■ Decision Tree

Grafik altında kalan alanları göz önünde bulunduracak olursak CN2 modeli Karar ağacından daha iyi bir model olduğunu söyleyebiliriz.

7. Test and Score Tablo



Model	AUC	CA	F1	Prec	Recall
CN2 Rule Induction	0.811	0.860	0.359	0.380	0.340

AUC, CA (Class Accuracy), F1(F-Measure), Precision ve Recall (sensitivity) değerleri ölçülmüştür.

AUC değeri (0.811) 1 değerine yeterince yakın ve yeterince iyi sınıflandırdığı anlamına gelmekte.

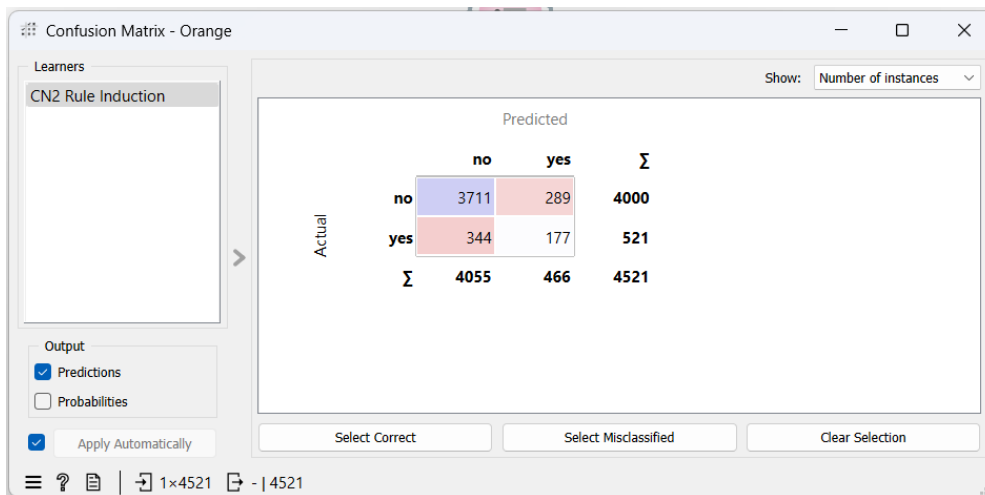
Accuracy değerimiz de 1 değerine çok yakındır.

Precision: Modelin pozitif olarak tahmin ettiği örneklerin ne kadarının gerçekten pozitif olduğunu gösterir. Burada gerçekte %38 pozitif olduğunu göstermekte.

Recall (Sensitivity): Gerçek pozitif örneklerin ne kadarının model tarafından doğru tahmin edildiğini gösterir. Burada %34 model tarafından doğru tahmin edilmiş.

F-Measure, Precision ve Recall arasında bir denge kurarak modelin genel performansını değerlendirmeye yardımcı olur. Modelin genel performansını değerlendirir.

8. Confusion Matrix Tablosu



		Predicted		Σ
		no	yes	
Actual	no	3711	289	4000
	yes	344	177	521
Σ		4055	466	4521

Output: ☒ Predictions ☐ Probabilities

☒ Apply Automatically

Select Correct Select Misclassified Clear Selection

$$\text{Precision(P)} = \text{TP} / (\text{TP} + \text{FP})$$

$$= 177 / (177 + 289)$$

$$= 0,38$$

$$\text{Recall(r)} = \text{TP} / (\text{TP} + \text{FN})$$

$$= 177 / (177 + 344)$$

$$= 0,34$$

$$\text{F-measure(F)} = 2rp / (r + p)$$

$$= 2 * 0,34 * 0,38 / (0,34 + 0,38)$$

$$= 0,359$$

		NO	YES
NO		TN	FP
YES		FN	TP

8.1. Performans Karşılaştırması

Test and Score - Orange					
Evaluation results for target yes					
Model	AUC	CA	F1	Prec	Recall
Tree	0.589	0.883	0.420	0.491	0.367
CN2 Rule Induction	0.811	0.860	0.359	0.380	0.340

Accuracy, F, Precision, Recall değerlerine baktığımızda Karar Ağacı modelinin daha iyi sınıflandırdığını söyleyebiliriz.

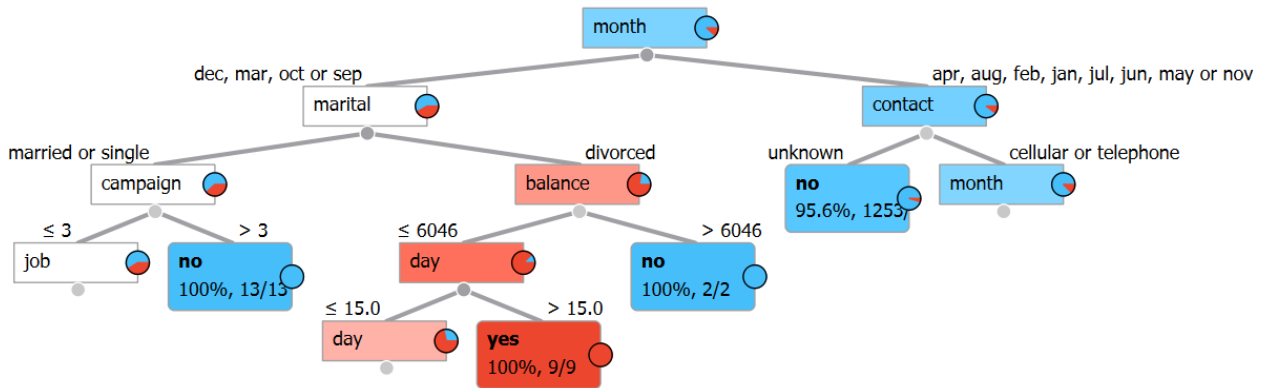
9. Gain Ratio ve Gini Değerleri

		#	Gain ratio	Gini
1	C month	12	0.010	0.011
2	N pdays		0.020	0.007
3	N previous		0.018	0.006
4	C contact	3	0.014	0.004
5	C job	12	0.003	0.003
6	C housing	2	0.008	0.002
7	N balance		0.003	0.002
8	N campaign		0.002	0.001
9	C loan	2	0.007	0.001
10	C marital	3	0.002	0.001
11	C education	4	0.001	0.001
12	N age		0.001	0.001
13	T day		0.001	0.000

En düşük gini'ye sahip olduğu için day özneliği en iyi özneliktir. Gain Ratio değeri düşüktür.

10. Indirect Method

10.1. Karar ağaçlarından yararlanarak kural oluşturma



R1: (month=apr,aug,feb,jan,jul,jun,may, nov) and (contact=unknown)→no

R2: (month=dec,mar,oct,sep) and (marital=divorced) and (balance>6046) →no

R3: (month=dec,mar,oct,sep) and (marital=divorced) and (balance<6046) and (day>15)→yes

R4: (month=dec,mar,oct,sep) and (marital=married or single) and (campaign>3)→ no

R: A→Y

$$\text{Coverage}(r) = |A|/|D|$$

$$\text{Accuracy}(r) = |A \cap y|/|A|$$

$$\text{Coverage}(R1) = 1253/4521=0,277$$

$$\text{Accuracy}(R1) = 1253/1311=0,95$$

$$\text{Coverage}(R2) = 2/4521= 0,0004$$

$$\text{Accuracy}(R2) = 2/2=1$$

$$\text{Coverage}(R3) = 9/4521= 0,002$$

$$\text{Accuracy}(R3) = 9/9=1$$

$$\text{Coverage}(R4) = 13/4521=0,0028$$

$$\text{Accuracy}(R4) = 13/13=1$$

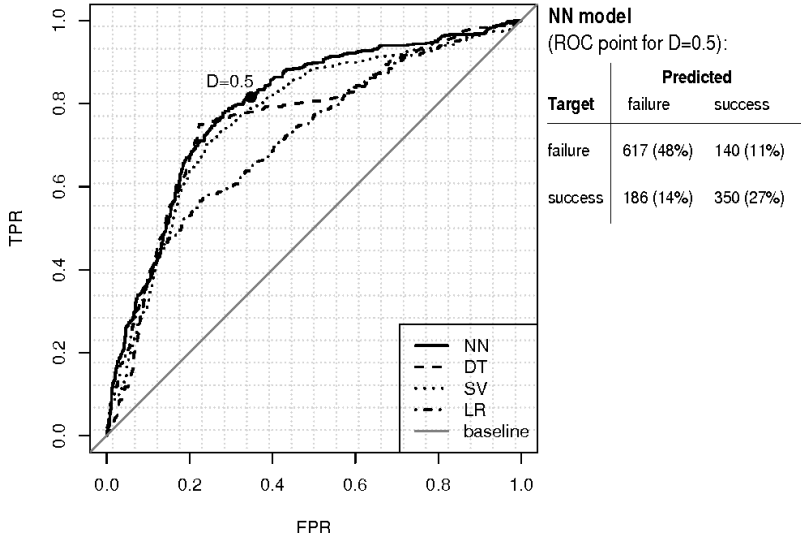
$$\text{Accuracy}(R2) = \text{Accuracy}(R3) = \text{Accuracy}(R4) > \text{Accuracy}(R1)$$

$$\text{Coverage}(R1) > \text{Coverage}(R4) > \text{Coverage}(R3) > \text{Coverage}(R2)$$

Doğruluk oranı (Accuracy) en düşük olan: R1 kuralı. Fakat kapsam bakımından bakacak olursak en yüksek kapsama sahip R1 kuralıdır. **Bu yüzden R1 kuralı daha güvenilirdir.**

11. Veri Seti Üzerinde Yapılan Akademik Çalışma

Makale: [A data-driven approach to predict the success of bank telemarketing](#)



Sinir ağları sınıflandırması ile yapılmış Confusion Matrix ve NN, DT, SV, LR modellerine göre çizilmiş ROC eğrileri.

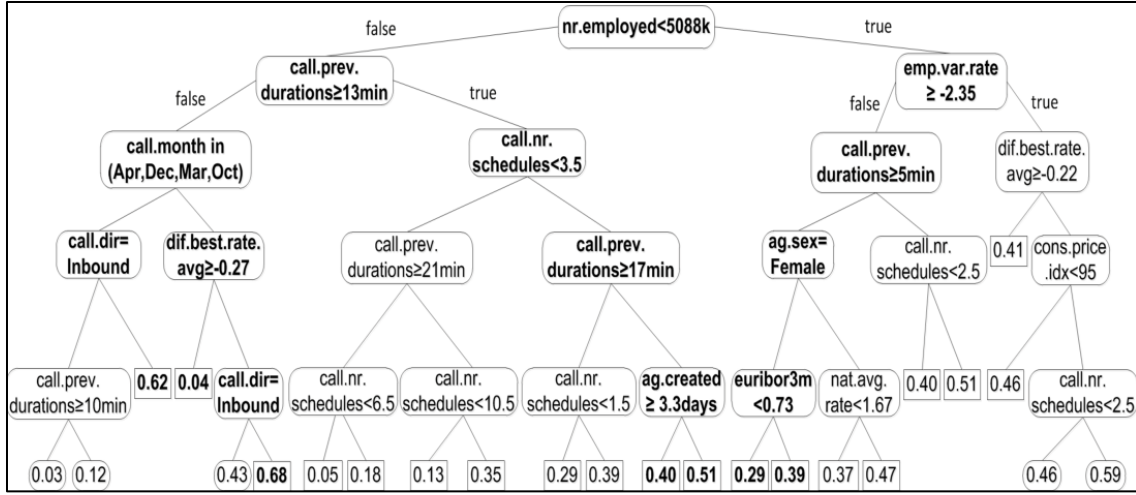
Sinir ağları ile modellenmiş sınıflandırmanın ROC eğrisi bizim kural tabanlı sınıflandırmaya çok benzemektedir.

Comparison of DM models for the modeling phase (best achieved test value)

Metric	LR	DT	SVM ($\tilde{\gamma} = 2^{-7.8}, C = 3$)	NN ($\tilde{H} = 6, N_r = 7$)
AUC	0.900	0.833	0.891	0.929*
ALIFT	0.849	0.756	0.844	0.878*

* - Statistically significant under a pairwise comparison with SVM, LR and DT.

LR, DT, SVM, NN sınıflandırmalarının ROC eğrisi altında kalan alanları verilmiştir. En iyi model Sinir Ağları. Bizim Kural Tabanlı Sınıflandırmamızda (CN2) AUC değeri 0.811 bulundu.



Sinir ağlarından çıkarılan karar ağacı.

12.Kaynakça

- Moro,S., Rita,P., and Cortez,P.. (2012). Bank Marketing. UCI Machine Learning Repository. <https://doi.org/10.24432/C5K306>.
- Introduction to Data Mining, 2nd Edition
- Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decis. Support Syst.*, 62, 22-31.