

## Tema 2 - IA

Cotelin Maria-Diana  
342C4

Analizați rezultatele pe baza tabelelor și graficelor obținute. Ce indică aceste grafice despre puterea de predicție a fiecărui atribut în parte? Cum vă puteți folosi de informația aflată în cadrul unui task de predicție?

### Atribute Numerice:

#### **Administrative:**

- Coefficient: 0.138917
- p-value: 3.519760e-54

#### **Administrative\_Duration:**

- Coefficient: 0.093587
- p-value: 2.146514e-25

#### **Informational:**

- Coefficient: 0.095200
- p-value: 3.174034e-26

#### **Informational\_Duration:**

- Coefficient: 0.070345
- p-value: 5.282871e-15

#### **ProductRelated:**

- Coefficient: 0.158538
- p-value: 3.241187e-70

#### **ProductRelated\_Duration:**

- Coefficient: 0.152373
- p-value: 6.115338e-65

#### **BounceRates:**

- Coefficient: -0.150673
- p-value: 1.594198e-63

#### **ExitRates:**

- Coefficient: -0.207071
- p-value: 1.662654e-119

#### **PageValues:**

- Coefficient: 0.492569
- p-value: 0.000000e+00

#### **SpecialDay:**

- Coefficient: -0.082305
- p-value: 5.498934e-20

### Observatii:

Atributele numerice prezintă corelații semnificative cu variabila țintă (Revenue), evidențiate prin p-value-uri mici.

Coeficienții pozitivi indică o corelație pozitivă, în timp ce coeficienții negativi indică o corelație negativă.

PageValues are cel mai puternic corelate, sugerând că această caracteristică este foarte importantă în predicția venitului.

### Atribute Categorice:

#### **Month:**

- Chi-squared: 384.934762
- p-value: 2.238786e-77

#### **OperatingSystems:**

- Chi-squared: 75.027056
- p-value: 1.416094e-13

#### **Browser:**

- Chi-squared: 27.715299
- p-value: 6.087543e-03

#### **Region:**

- Chi-squared: 9.252751
- p-value: 3.214250e-01

#### **TrafficType:**

- Chi-squared: 373.145565
- p-value: 1.652735e-67

#### **VisitorType:**

- Chi-squared: 135.251923
- p-value: 4.269904e-30

#### **Weekend:**

- Chi-squared: 10.390978
- p-value: 1.266325e-03

### Observatii:

Atributele categorice prezintă, de asemenea, corelații semnificative cu variabila țintă, conform p-value-urilor mici. Cu cât valoarea Chi-squared este mai mare, cu atât există o asociere mai puternică între variabile.

Valoarea p este probabilitatea de a obține un rezultat observat sau mai extrem, atunci când ipoteza nulă este adevărată. În contextul testului de independență Chi-squared, ipoteza nulă presupune că nu există nicio asociere între variabilele categorice (sunt independente). Un p-value mic (mai mic decât nivelul de semnificație ales, de obicei 0.05) indică respingerea ipotezei nule și susține existența unei asocieri semnificative între variabile.

## Evaluarea comparativa a rezultatelor

### A. Precision

Precision reprezintă proporția de instanțe pozitive prezise corect din totalul instanțelor prezise pozitiv. Un scor mai mare indică o capacitate mai mare a modelului de a evita clasificarea greșită a instanțelor negative.  $TP/(TP + FP)$

### B. Recall

Recall, cunoscut și sub numele de Sensitivity sau True Positive Rate, reprezintă proporția de instanțe pozitive prezise corect din totalul instanțelor pozitive reale. Scorurile mai mari indică o capacitate mai mare a modelului de a identifica instanțele pozitive.  $TP/(TP+FN)$

### C. F1-Score

F1-score este o metrică care ține cont atât de precision, cât și de recall. Este utilă atunci când se dorește o balanță între cele două metrici. Scorurile mai mari indică un echilibru mai bun între precision și recall.

### D. Accuracy

Accuracy reprezintă proporția totală de instanțe corect clasificate. Scorurile mai mari indică o precizie generală mai bună a modelului.  $(TP + TN)/ Tot$

## 1. LR Implementare Manuala

	Precision		Recall		F1-score		Accuracy	
Scaler	mean	var	mean	var	mean	var	mean	var
MinMax	0.485	0.137	0.007	0.000	0.013	0.000	0.848	0.000
Robust	0.737	0.017	0.373	0.029	0.457	0.001	0.870	0.002
<b>Standard</b>	<b>0.735</b>	<b>0.001</b>	<b>0.370</b>	<b>0.000</b>	<b>0.505</b>	<b>0.000</b>	<b>0.885</b>	<b>0.000</b>

### Observatii:

- **StandardScaler** pare să ofere cele mai bune rezultate în general, având scoruri ridicate pentru toate metricile.
- **RobustScaler** se comportă bine în ceea ce privește recall-ul, dar poate să aibă un impact asupra precision-ului.
- **MinMaxScaler** are scoruri scăzute pentru precision, recall și F1-score, indicând că ar putea avea dificultăți în a gestiona corect ambele clase.

## 2. Implementare scikit-learn

	Precision		Recall		F1-score		Accuracy	
Scaler	mean	var	mean	var	mean	var	mean	var
MinMax	0.775	0.001	0.281	0.001	0.412	0.001	0.875	0.0
<b>Robust</b>	<b>0.759</b>	<b>0.001</b>	<b>0.386</b>	<b>0.000</b>	<b>0.511</b>	<b>0.000</b>	<b>0.885</b>	<b>0.0</b>
<b>Standard</b>	<b>0.759</b>	<b>0.001</b>	<b>0.385</b>	<b>0.000</b>	<b>0.511</b>	<b>0.000</b>	<b>0.885</b>	<b>0.0</b>

Observatii:

- **Precizia** (Precision) rămâne destul de ridicată pentru toate scalările, indicând capacitatea modelului de a evita clasificarea greșită a instanțelor negative.
- **Recall**-ul a crescut față de valorile anterioare pentru toate scalările, indicând o capacitate mai bună de a identifica instanțele pozitive.
- **F1-score** și **Accuracy** au crescut și ele, sugerând o performanță generală îmbunătățită a modelului.

### 3. Decision Tree (scikit-learn)

	Depth	Precision		Recall		F1-score		Accuracy	
Scaler		mean	var	mean	var	mean	var	mean	var
MinMax	3	0.668	0.001	0.612	0.005	0.636	0.002	0.894	0.0
	4	0.703	0.002	0.586	0.005	0.635	0.002	0.898	0.0
	5	0.717	0.002	0.579	0.002	0.638	0.001	0.900	0.0
	6	0.690	0.002	0.582	0.002	0.630	0.001	0.896	0.0
None	3	0.658	0.001	0.630	0.001	0.644	0.001	0.894	0.0
	4	0.703	0.001	0.568	0.004	0.625	0.001	0.897	0.0
	5	0.696	0.003	0.575	0.002	0.627	0.0	0.896	0.0
	6	0.698	0.002	0.572	0.001	0.627	0.001	0.897	0.0
Robust	3	<b>0.652</b>	<b>0.000</b>	<b>0.646</b>	<b>0.000</b>	<b>0.649</b>	<b>0.0</b>	<b>0.893</b>	<b>0.0</b>
	4	0.698	0.000	0.588	0.001	0.638	0.0	0.898	0.0
	5	0.909	0.000	0.576	0.000	0.630	0.0	0.897	0.0
	6	0.698	0.001	0.567	0.001	0.630	0.0	0.891	0.0

Standard	3	0.654	0.000	0.616	0.002	0.633	0.0	0.897	0.0
	4	0.700	0.001	0.580	0.002	0.632	0.0	0.896	0.0
	5	0.705	0.001	0.557	0.002	0.621	0.0	0.896	0.0
	6	0.698	0.001	0.570	0.001	0.626	0.0	0.896	0.0

#### Observatii:

- Scalerele au un impact semnificativ pe performanța modelului în cazul arborelui de decizie. Rezultatul cel mai bun apare pentru Robust Scaler pentru F1 score.
- Performanța generală a modelului, măsurată prin Accuracy, rămâne în general ridicată pentru toate scalerele și adâncimile.
- Pentru Adancimea 3 F1-score pare sa fie cel mai bun, desi acuratetea si precizia sunt mai bune pentru adancimea 5.

#### Concluzii:

Cerinta mentiona evidentierea configuratiei cu cea mai buna metrica pentru F1, in cazul de fata Arbori de decizie. Desi precizia este mai buna la LR, Arborii de decizie au valori mai mari la RCALL, prin urmare si F1-score (care este dependent de ambele) este mai bun. De asemenea, in urma analizei de la punctul 2 am dedus ca pentru acest model caracteristicile sunt puternic corelate, iar acest lucru are impact negativ asupra regresiei logistice.