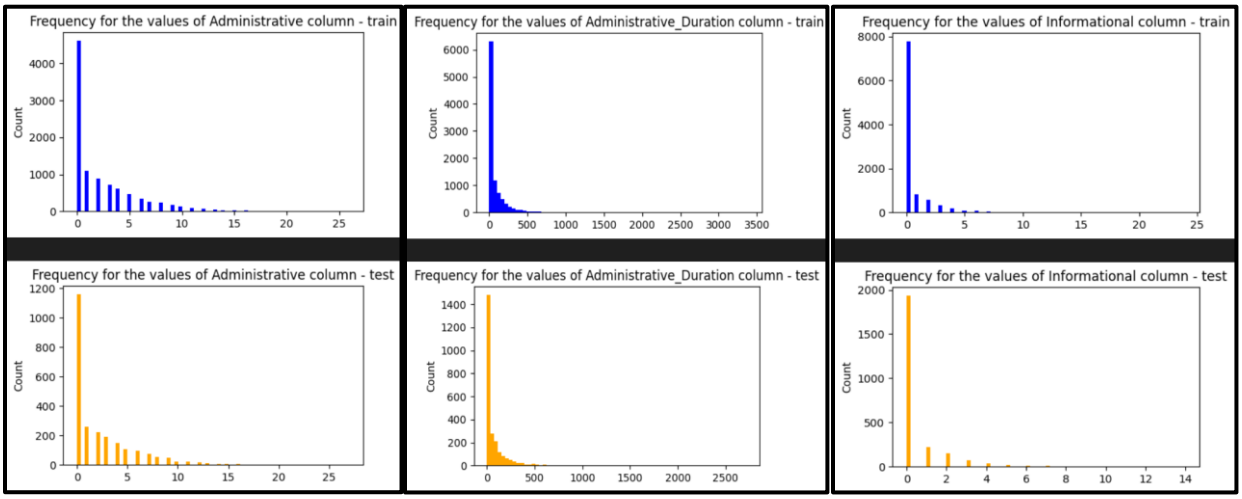
Tema 2 – Inteligenta Artificiala

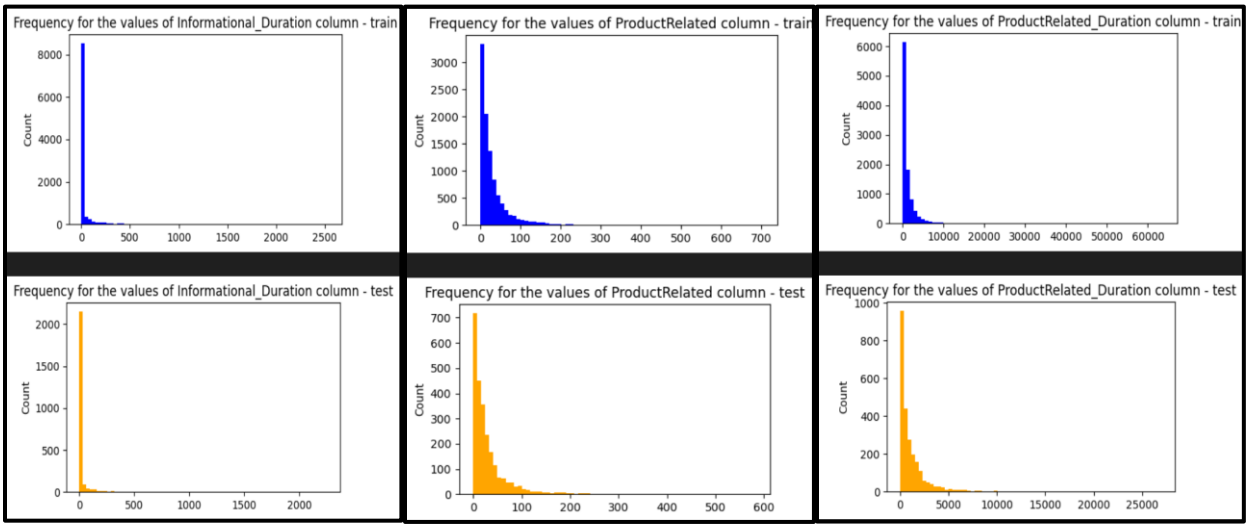
344C3 – Avramescu Cosmin-Alexandru

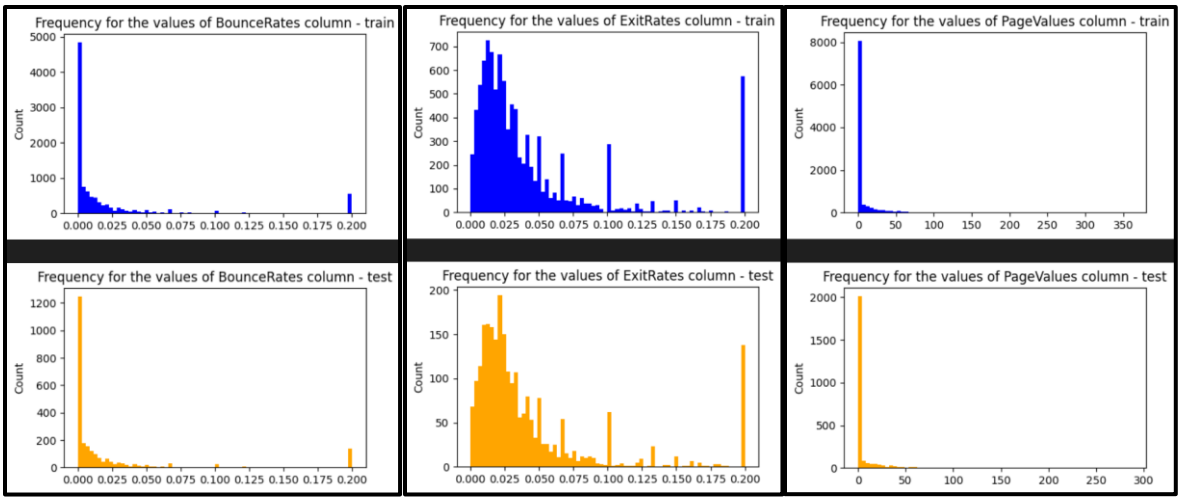
3.1 Explorarea Datelor

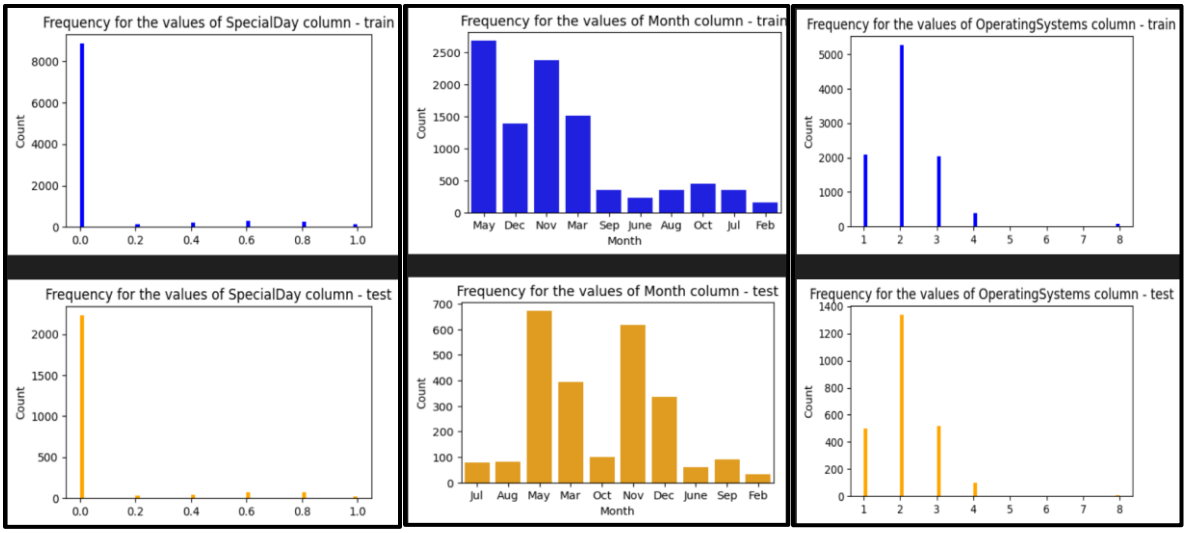
1. Analiza echilibrului de clase

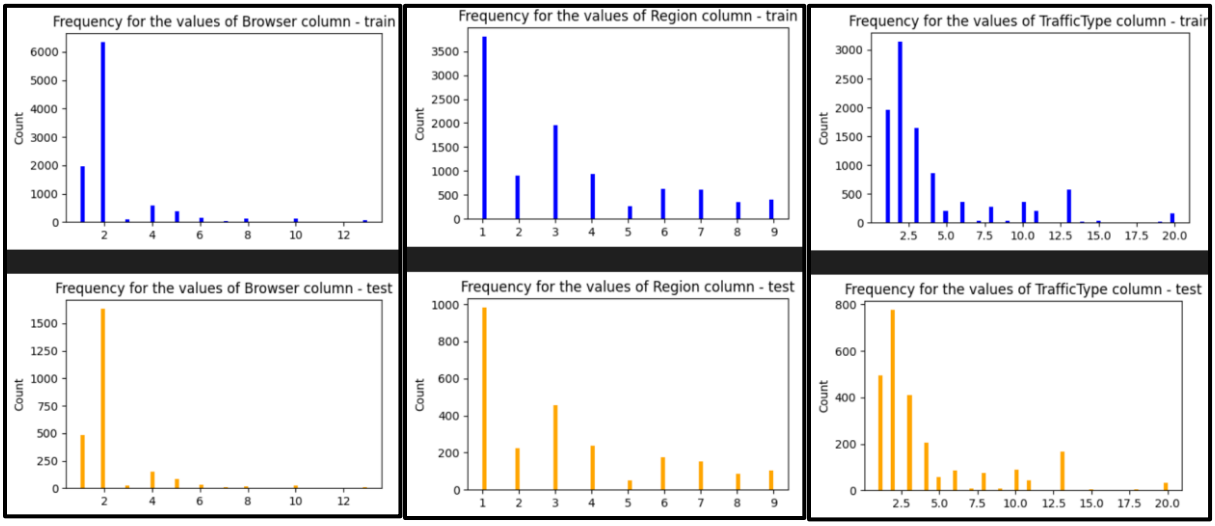
Am imartit setul de date in 80% train (albastru) si 20% test (portocaliu) cu metoda train_test_split(). Dupa cum se poate vedea din toate aceste grafice (pe verticala), impartirea aleatorie a datelor, duce la pastrarea proportiilor intre train si test deoarece formele barelor sunt similare, dar la un ordin de marime mai mic (dupa cum se observa, de exemplu, in primul grafic, 4000 vs 1200 este ordinal de marime). Asadar, datele sunt impartite bine intre train si test ca frecventa de aparitie a claselor.

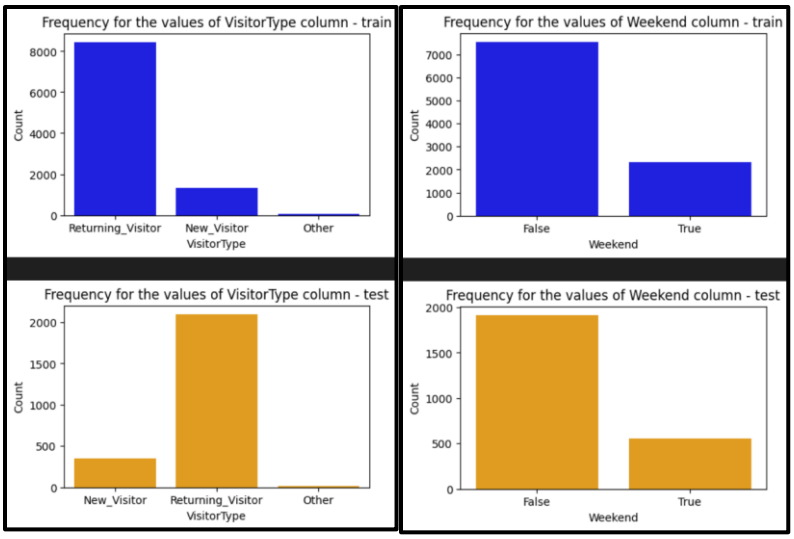





Frequency for the values of BounceRates column - train
Count
0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175 0.200
Frequency for the values of ExitRates column - train
Count
0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175 0.200
Frequency for the values of PageValues column - train
Count
0 50 100 150 200 250 300 350
Frequency for the values of BounceRates column - test
Count
0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175 0.200
Frequency for the values of ExitRates column - test
Count
0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175 0.200
Frequency for the values of PageValues column - test
Count
0 50 100 150 200 250 300


Frequency for the values of SpecialDay column - train
Count
0.0 0.2 0.4 0.6 0.8 1.0
Frequency for the values of Month column - train
Count
May Dec Nov Mar Sep June Aug Oct Jul Feb
Month
Frequency for the values of OperatingSystems column - train
Count
1 2 3 4 5 6 7 8
Frequency for the values of SpecialDay column - test
Count
0.0 0.2 0.4 0.6 0.8 1.0
Frequency for the values of Month column - test
Count
Jul Aug May Mar Oct Nov Dec June Sep Feb
Month
Frequency for the values of OperatingSystems column - test
Count
1 2 3 4 5 6 7 8

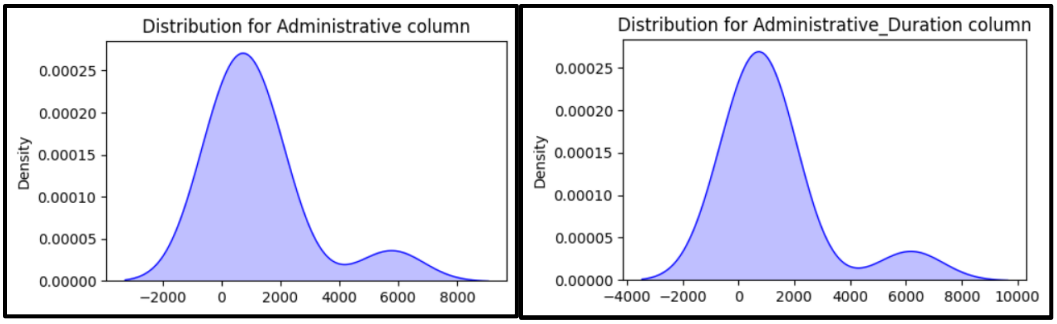

Frequency for the values of Browser column - train
Count
2 4 6 8 10 12
Frequency for the values of Region column - train
Count
1 2 3 4 5 6 7 8 9
Frequency for the values of TrafficType column - train
Count
2.5 5.0 7.5 10.0 12.5 15.0 17.5 20.0
Frequency for the values of Browser column - test
Count
2 4 6 8 10 12
Frequency for the values of Region column - test
Count
1 2 3 4 5 6 7 8 9
Frequency for the values of TrafficType column - test
Count
2.5 5.0 7.5 10.0 12.5 15.0 17.5 20.0

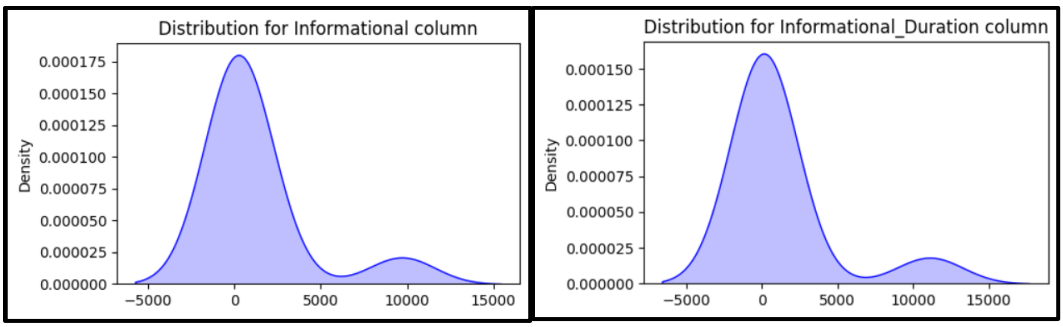


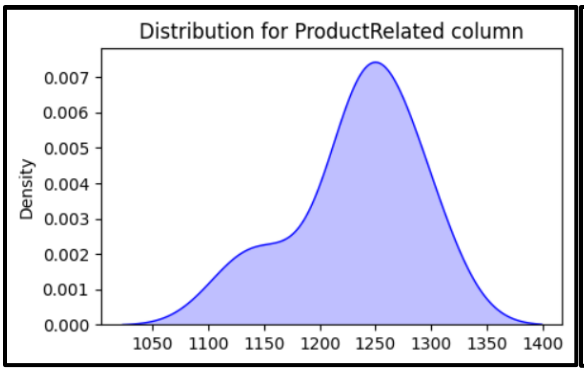
2.A. Analiza distributiei valorilor

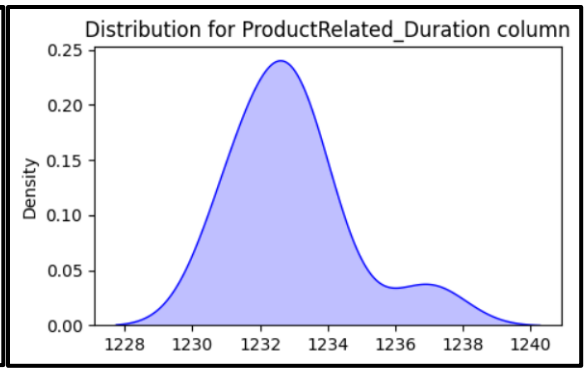
- Atribute numerice: distributia valorilor in percentile cu granularitate de 10%

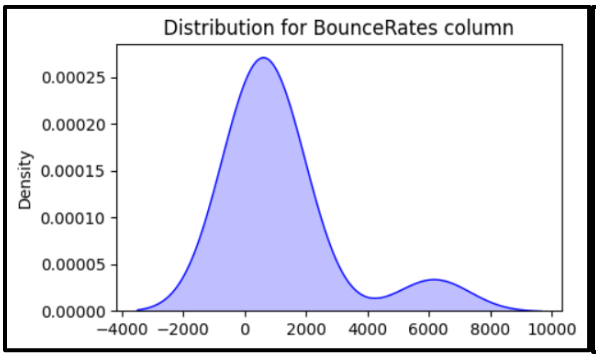
Pentru grafice, am folosit KDE plot deoarece este a varianta mai smooth de a afisa distributia datelor. Dupa cum apare si in fisierul jupyter, fiecare grafic are valori pentru intervalele [(0, 10), (10, 20), (20, 30), (30, 40), (40, 50), (50, 60), (60, 70), (70, 80), (80, 90), (90, 100)], iar rezultatele (ca frecventa de valori pentru fiecare dintre aceste intervale) sunt (pentru Administrative Column) - [0, 0, 0, 0, 5768, 1354, 1114, 915, 1772, 1406]. Ceea ce inseamna ca avem 5768 dintre valori in intervalul 40-50%, insa avem 6561 de valori in total in intervalul 50-100%, iar acestea se vad in varful graficului in jurul valorii ~ 1100 pe axa Ox.

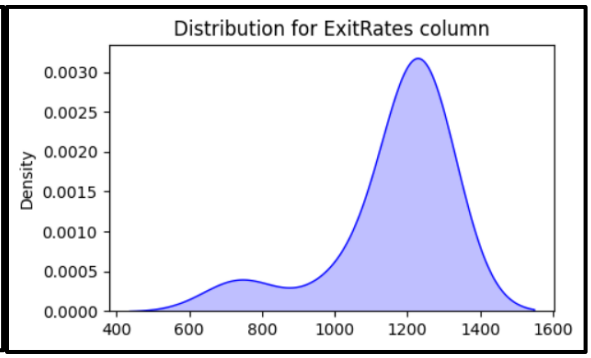


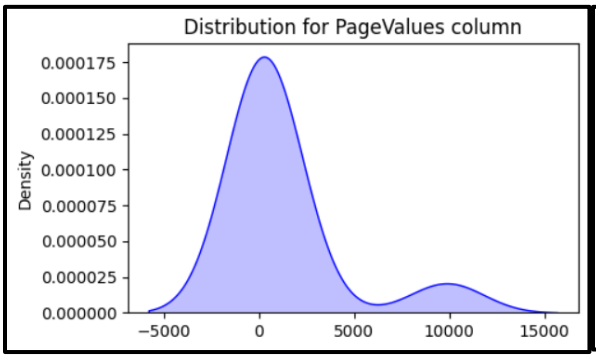


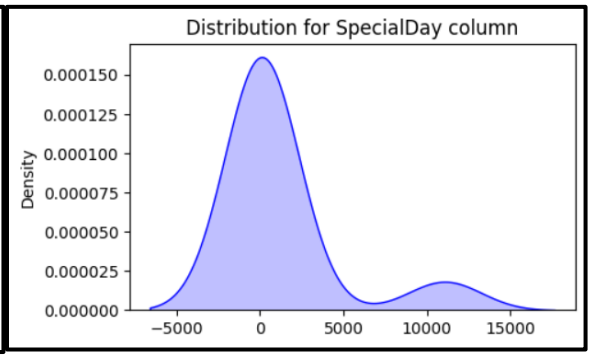

Distribution for ProductRelated column
Density
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
1050 1100 1150 1200 1250 1300 1350 1400

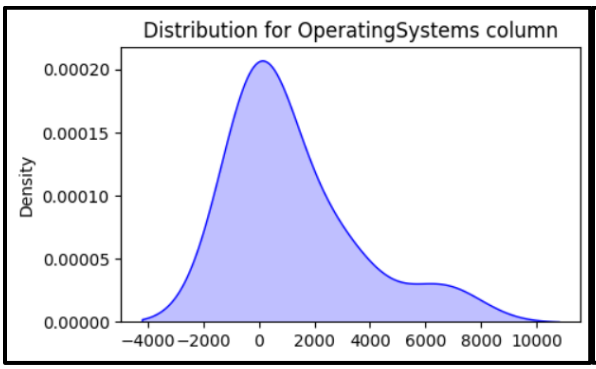

Distribution for ProductRelated_Duration column
Density
0.25
0.20
0.15
0.10
0.05
0.00
1228 1230 1232 1234 1236 1238 1240

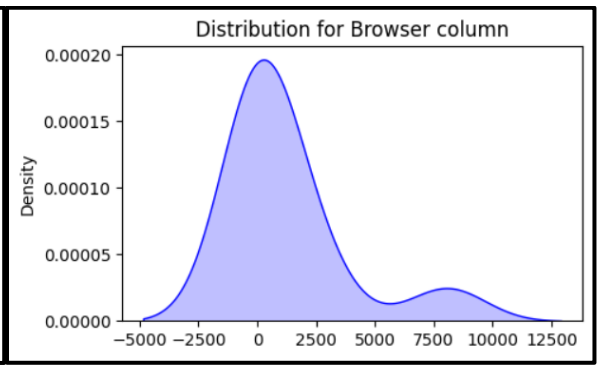

Distribution for BounceRates column
Density
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
−4000 −2000 0 2000 4000 6000 8000 10000

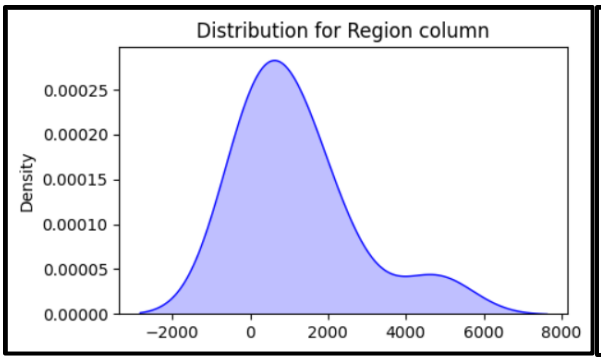

Distribution for ExitRates column
Density
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
400 600 800 1000 1200 1400 1600

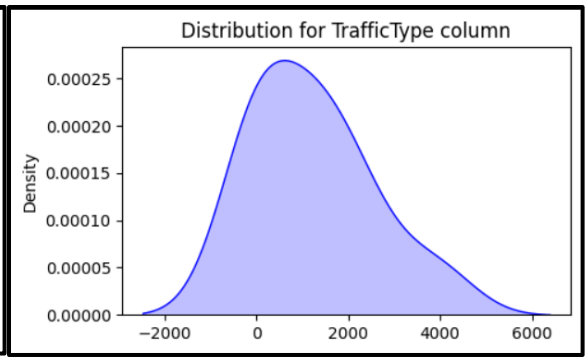

Distribution for PageValues column
Density
0.000175
0.000150
0.000125
0.000100
0.000075
0.000050
0.000025
0.000000
−5000 0 5000 10000 15000


Distribution for SpecialDay column
Density
0.000150
0.000125
0.000100
0.000075
0.000050
0.000025
0.000000
−5000 0 5000 10000 15000


Distribution for OperatingSystems column
Density
0.00020
0.00015
0.00010
0.00005
0.00000
−4000 −2000 0 2000 4000 6000 8000 10000

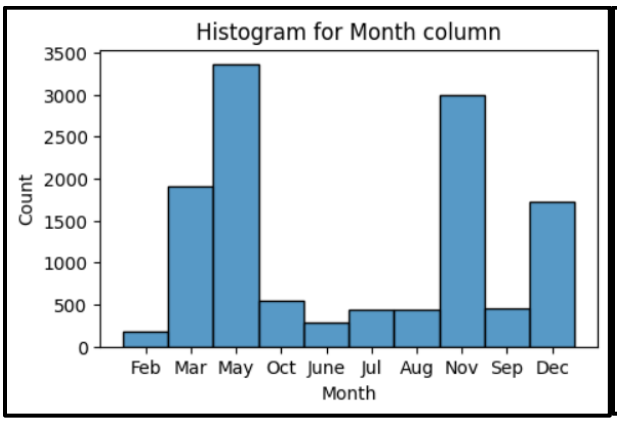

Distribution for Browser column
Density
0.00020
0.00015
0.00010
0.00005
0.00000
−5000 −2500 0 2500 5000 7500 10000 12500

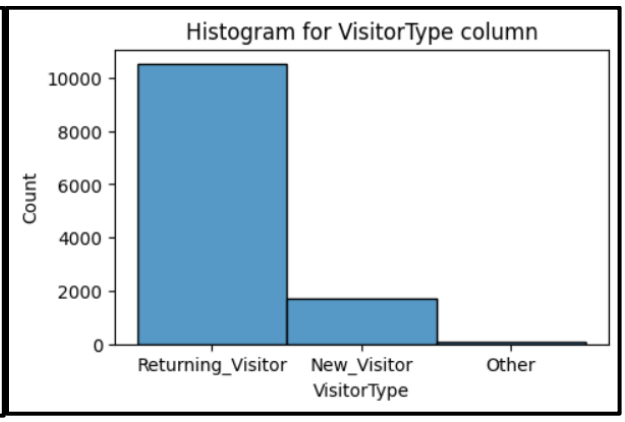


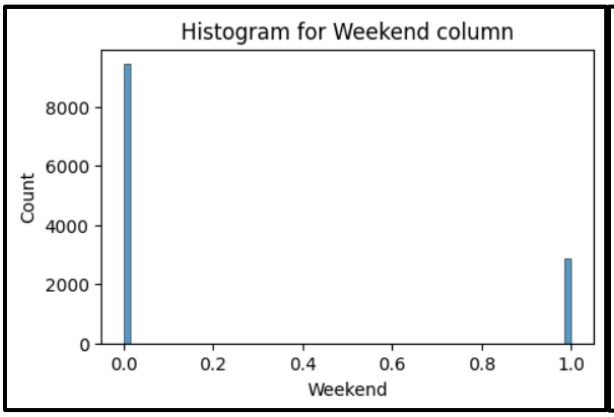


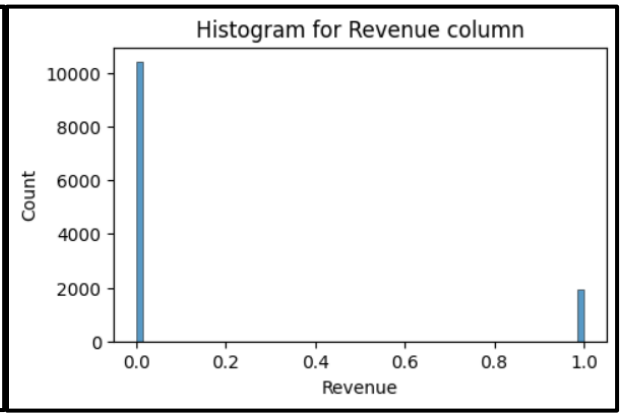
- Atribute categorice: grafic al histogramei

Se pot observa toate valorile posibile pentru fiecare coloanal cu valori categorice (valori care nu sunt numerice, fie sunt object - string, fie sunt bool).







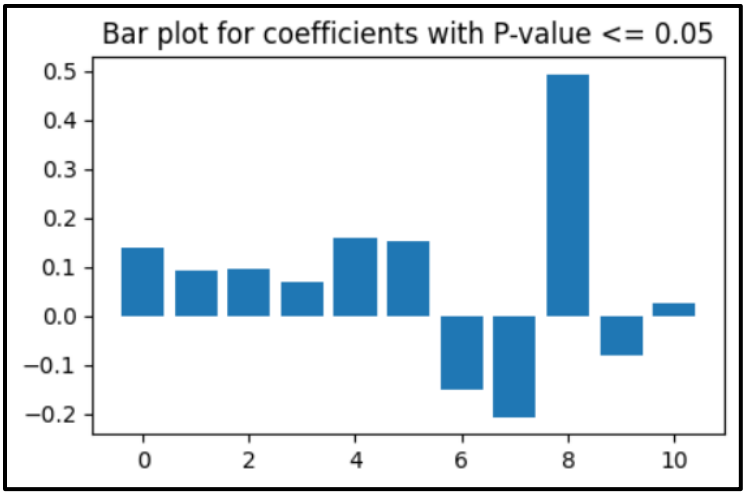


2.B. Analiza gradului de corelare cu variabila tinta Revenue

- Atribute numerice: coeficient Point-Biserial-Correlation

Coeficientii pozitivi arata o corelatie pozitiva (daca valoare coloanei creste, valoare coloanei referinta creste si ea), in timp ce un coeficient negativ arata o corelatie negativa (daca valoarea coloanei creste, valoarea coloanei referinta scade). Valorile p sub 0.05 arata ca acele corelatii nu sunt din intamplare, ceea ce inseamna ca sunt corelatii puternice. De exemplu, OperatingSystem, Region si TrafficType au valori p peste 0.05 si are sens pentru ca ele nu sunt un determinant real pentru variabila de Revenue, ele nu au corelatie reala cu variabila de Revenue.

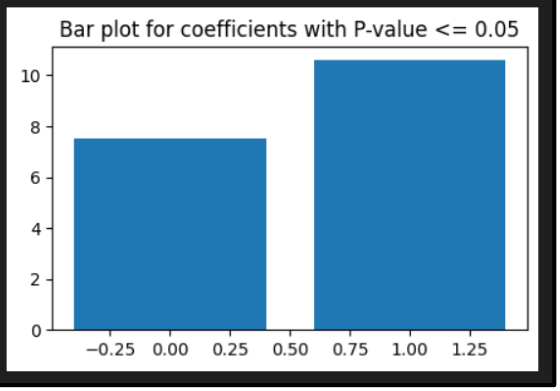
Column name	Coefficient	P-value
Administrative	0.138917	3.51976e-54
Administrative_Duration	0.0935867	2.14651e-25
Informational	0.0952003	3.17403e-26
Informational_Duration	0.0703445	5.28287e-15
ProductRelated	0.158538	3.24119e-70
ProductRelated_Duration	0.152373	6.11534e-65
BounceRates	-0.150673	1.5942e-63
ExitRates	-0.207071	1.66265e-119
PageValues	0.492569	0
SpecialDay	-0.0823046	5.49893e-20
OperatingSystems	-0.0146676	0.103394
Browser	0.0239843	0.00773689
Region	-0.0115951	0.197943
TrafficType	-0.00511297	0.570243



- Atribute categorice: test Pearson Chi-Squared

Rezultatul testului este pe coloanal Coefficient, iar valorile P sub 0.05 arata ca este o corelatie puternica intre acea coloana si coloana Revenue. Ceea ce are sens, pentru ca variabilele Month si Weekend determina daca userul face sau nu cumparaturi pe site (la weekend este mult mai mic p value pentru ca sunt mult mai multe sanse sa fie vanzari in weekend decat in timpul saptamanii), insa si variabila Month conteaza (de exemplu de Craciun sau alte sarbatori cresc sansele de vanzari).

Column name	Coefficient	P-value
Month	7.53615	0.0230965
VisitorType	1.61638	0.445664
Weekend	10.5818	0.00503714



3.2. Regresie Logistica

- Evaluarea comparativa a rezultatelor:

Scaler - Type	Mean - precision	Mean - recall	Mean - F1	Variance - precision	Variance - recall	Variance - F1
MinMaxScaler - manual	0.664586	0.439782	0.515753	0.0109281	0.0112237	0.0031256
MinMaxScaler - library	0.75	0.282051	0.409938	0	3.08149e-33	3.08149e-33
StandardScaler - manual	0.615153	0.510231	0.483106	0.0384813	0.0484522	0.00617049
StandardScaler - library	0.723757	0.373219	0.492481	0	0	0
RobustScaler - manual	0.481333	0.548822	0.420722	0.0734054	0.0868164	0.0259949
RobustScaler - library	0.723757	0.373219	0.492481	0	0	0

Pentru fiecare intrare din tabel, am rulat de 10 ori pe 10 impartiri aleatorii ale setului de date si apoi am calculate mean si variance pentru aceste 10 rulari. Am rulat atat varianta de Logistic Regression manual cat si Logistic Regression din biblioteca. Precision determina cate dintre rezultatele prezise ca pozitive de model (revenue true) sunt corecte (pot aparea fals pozitive). Recall determina cate rezultate pozitive au fost identificare

corect de model (pot aparea fals negative). F1 combina aceste metrici si este ideal sa avem un model care are un F1-score cat mai mare. Mean calculeaza media, iar variance arata cat de aproape sunt valorile de mean. Cum variance este aproape de 0 sau 0 la F1-score, stim ca in cele 10 rulari se obtin rezultate similar aproape de mean. Asadar, la Logistic Regression, **cel mai bun model este cel manual cu MinMaxScaler**. Cel mai bun model este combinatia de Logistic manual cu MinMaxScaler deoarece MinMaxScaler se potriveste feature-urilor care au range-uri diferite (avem multe feature-uri, fiecare cu range-uri de valori foarte diferite intre ele). Interesant este ca doar cu acest scaler este mai bun modelul manual, deoarece in rest este mai bun modelul din biblioteca.

3.3. Arbori de Decizie

- Evaluarea comparativa a rezultatelor:

Scaler - MaxDepth	Mean - precision	Mean - recall	Mean - F1	Variance - precision	Variance - recall	Variance - F1
MinMaxScaler - 3	0.897766	0.900243	0.898888	4.93038e-32	4.93038e-32	1.2326e-32
MinMaxScaler - 4	0.899413	0.904704	0.901265	0	1.2326e-32	0
MinMaxScaler - 5	0.900566	0.904704	0.902206	1.2326e-32	1.2326e-32	0
MinMaxScaler - 6	0.88995	0.897736	0.892183	7.11892e-08	5.43611e-08	6.336e-08
StandardScaler - 3	0.897766	0.900243	0.898888	4.93038e-32	4.93038e-32	1.2326e-32
StandardScaler - 4	0.899413	0.904704	0.901265	0	1.2326e-32	0
StandardScaler - 5	0.900566	0.904704	0.902206	1.2326e-32	1.2326e-32	0
StandardScaler - 6	0.890033	0.89781	0.89226	1.16199e-07	8.96959e-08	1.02602e-07
RobustScaler - 3	0.897766	0.900243	0.898888	4.93038e-32	4.93038e-32	1.2326e-32
RobustScaler - 4	0.899413	0.904704	0.901265	0	1.2326e-32	0
RobustScaler - 5	0.900566	0.904704	0.902206	1.2326e-32	1.2326e-32	0
RobustScaler - 6	0.890046	0.89781	0.892275	1.16205e-07	8.96959e-08	1.026e-07
NoneType - 3	0.897766	0.900243	0.898888	4.93038e-32	4.93038e-32	1.2326e-32
NoneType - 4	0.899413	0.904704	0.901265	0	1.2326e-32	0
NoneType - 5	0.900566	0.904704	0.902206	1.2326e-32	1.2326e-32	0
NoneType - 6	0.890103	0.897884	0.892323	7.50398e-08	5.43611e-08	6.75263e-08

Pentru fiecare intrare din tabel, am rulat de 10 ori pe 10 impartiri aleatorii ale setului de date si apoi am calculate mean si variance pentru aceste 10 rulari. Am rulat varianta de Decision Tree din biblioteca, cu coeficient gini si max depth variabil de la 3 la 6. Precision determina cate dintre rezultatele prezise ca pozitive de model (revenue true) sunt corecte (pot aparea fals pozitive). Recall determina cate rezultate pozitive au fost identificare corect de model (pot aparea fals negative). F1 combina aceste metrici si este ideal sa avem un model care are un F1-score cat mai mare. Mean calculeaza media, iar variance arata cat de aproape sunt valorile de mean. Cum variance este aproape de 0 sau 0 la F1-score, stim ca in cele 10 rulari se obtin rezultate similar aproape de mean. Asadar, la Decision Tree, cel

mai bun model este cel cu max depth de 5, indiferent de scaler deoarece arborii de decizie sunt invariabil la scaling. Un depth mai mare poate duce la overfitting, iar un depth mai mic poate duce la underfitting, asa ca in cazul nostrum valoare de 5 pentru max depth este cea care da cele mai bune rezultata (fata de valorile de la 3 la 6).