מדעי הנתונים ובינה עסקית

תרגיל מס' 1

הוראות כלליות:

א. הגשת העבודה היא בזוגות.

ב. יש להגיש את העבודה דרך המודל (שותף אחד מכל זוג צריך להגיש).

ג. יש להגיש את העבודה עד לתאריך 9.4.2020.

ד. יש לכתוב תשובות סופיות במקום המיועד ובסוף המסמך לצרף את **כל** החישובים המתאימים.

ה. יש להקפיד על **3** ספרות אחרי הנקודה העשרונית, **אי הקפדה תגרור הורדת נקודות**.

ו. תשובה סופית, גם אם היא נכונה, ללא הצגת דרך החישוב המלאה בנספחים, **לא תזכה בניקוד**.

שאלה 1 (30 נק')

כאשר:1931-1968הנתונים הבאים מתארים את שיעורי הפשיעה בקנדה בשנים

Tfr	Total fertility rate per 1000 women.
Partic	Women's labor-force participation rate per 1000.
Degrees	Women's post-secondary degree rate per 10,000.
Fconvict	Female indictable-offense conviction rate per 100,000.
Ftheft	Female theft conviction rate per 100,000.
Mconvict	Male indictable-offense conviction rate per 100,000.
Mtheft	Male theft conviction rate per 100,000.

year	tfr	partic	degrees	fconvict	ftheft	mconvict	mtheft
1935	2755	238	13.2	79.4	20.4	765.7	247.1
1936	2696	240	13.2	91	22.1	816.5	254.9
1937	2646	241	12.2	100.4	22.4	821.8	272.4
1938	2701	242	12.6	108.9	21.8	956.8	285.8
1939	2654	244	12.3	123.6	21.1	1035.7	292.2
1940	2766	245	12	157.3	21.4	951.6	256
1941	2832	246	11.7	154.3	25.3	850.9	205.8
1942	2964	268	11.2	143.9	27.1	769.7	188
1943	3041	333	11.5	147.5	29	811.2	205.8
1944	3010	335	11.1	97.3	24.2	865.4	207.9

1945	3018	331	12.5	76.6	24.7	866.8	197.8
1946	3374	253	15	72.8	20.5	968	195.9
1947	3545	243	17.6	68.9	20.7	894	198.9
1948	3441	241	21.2	66.7	22.8	830.8	198.6
1949	3456	242	22.7	55	18.5	811.2	216.1
1950	3455	237	21.4	55	19.3	775.4	212
1951	3503	242	20.7	55	20	739.8	208
1952	3641	240	19.5	54.7	19.4	740.3	199.3
1953	3721	233	36.6	47.9	16.1	725.1	176.4

א. יש לנרמל את המשתנה **partic** באמצעות Min-Max normalization כאשר New min = 1, New max = 10.

1. מהו המינימום המקורי של המשתנה? ___233_____
2. מהו המקסימום המקורי של המשתנה? _____335_____
3. מהו הערך המנורמל של המשתנה בשנת 1952? ___1.618_____
4. יש לצרף טבלה המכילה את המשתנה המנורמל בנספחים.

ב. יש לבצע החלקה למשתנה mconvict באמצעות Weighted moving average כאשר:

t	Weight
-3	180.
-2	220.
-1	60.

1. מהו הערך של המשתנה mconvict המוחלק בשנת 1952_____760.484_____
2. מהו הערך של המשתנה mconvict המוחלק בשנת 1946? ____856.484_____
3. נא לצרף את המשתנה המוחלק בנספחים.
4. נא לצרף גרף עם המשתנה המקורי והמשתנה לאחר ההחלקה בנספחים.

שאלה 2 (35 נק')

מעוניינים לחזות נוכחות של מחלת לב בחולים. לשם כך, Cleveland Clinic Foundation רופאים מ-משתנים מועמדים ומשתנה 3) מכיל hw1_dataset המצורף (CSVנאסף מידע על 100 חולים. קובץ ה-prediction מטרה אחד).

הגדרות המשתנים:

Attribute	Domain	Type
gender	1=male, 0=female	Nominal
fbs	0=false, 1=true	Nominal
slope	1=upsloping, 2=flat, 3=downsloping	Nominal
Prediction	0,1,2,3,4	Nominal

א. מהי האנטרופיה של המשתנה gender? ____0.869_____
ב. מהי האנטרופיה של המשתנה fbs? (10 נק') ____0.577_____
ג. מהי האנטרופיה המותנית של משתנה המטרה בהינתן המשתנה slope ____1.577_____
ד. מהי האנטרופיה המקסימלית של המשתנה slope? ____1.585_____
ה. יש לחשב את ה-mutual information של משתנה המטרה ו-gender בשתי דרכים:
1. לפי ההפרש בין האנטרופיות הרלוונטיות _____
2. לפי הנוסחה של mutual information I(X;Y) _____

שאלה 3 (35 נק')

יש להשתמש באותו קובץ נתונים משאלה 2.

א. יש לבנות מודל ID3 (עד רמה 2, קודקוד השורש הוא רמה 0). יש להראות חישוב ידני של לפחות משתנה אחד בכל קדקוד, לשאר המשתנים ניתן לכתוב רק את התוצאה הסופית. יש להציג את העץ שהתקבל.
ב. מהו ה - majority rule accuracy? _____
ג. מהו ה - training accuracy של המודל? _____
ד. 5 חולים חדשים הגיעו לבית החולים. ע"פ המודל שבניתם, סווגו את הרשומות הבאות:

Gender	fbs	Slope	Prediction
0	1	1	**0**
1	0	2	**1**
1	1	3	**0**
0	0	3	**3 (random choice between 3 and 0)**
0	0	2	**0**

בהצלחה!

נספחים

1א.

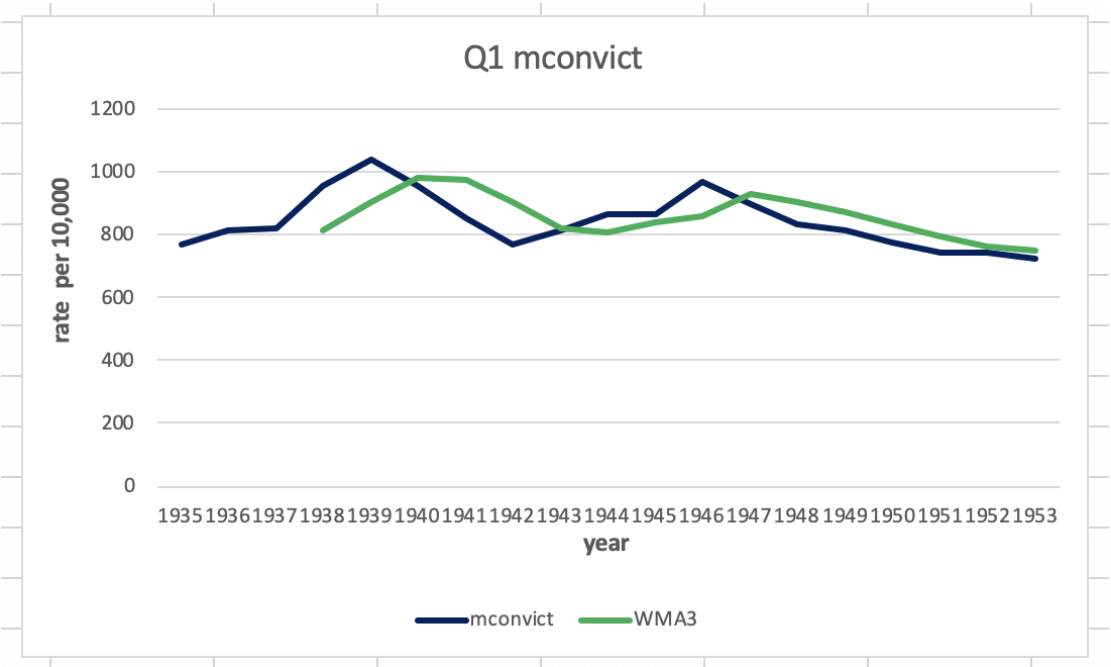
year	normalized partic
1935	1.441
1936	1.618
1937	1.706
1938	1.794
1939	1.971
1940	2.059
1941	2.147
1942	4.088
1943	9.824
1944	10.000
1945	9.647
1946	2.765
1947	1.882
1948	1.706
1949	1.794
1950	1.353
1951	1.794
1952	1.618
1953	1.000

233				1	1
335				2	
10		new max		3	
1		new min			

1ב.

year	mconvict	WMA3
1935	765.7	
1936	816.5	
1937	821.8	
1938	956.8	810.536
1939	1035.7	901.846
1940	951.6	979.84
1941	850.9	971.038
1942	769.7	906.318
1943	811.2	820.306
1944	865.4	809.216
1945	866.8	836.25
1946	968	856.484
1947	894	927.268
1948	830.8	905.384
1949	811.2	869.4
1950	775.4	830.416
1951	739.8	793.248
1952	740.3	760.484
1953	725.1	746.508

Weight	t
0.18	-3
0.22	-2
0.6	-1


Q1 mconvict
1200
1000
800
600
400
200
0
rate per 10,000
1935 1936 1937 1938 1939 1940 1941 1942 1943 1944 1945 1946 1947 1948 1949 1950 1951 1952 1953
year
mconvict
WMA3

2א.

א.

	total	probability
male	71	0.710
female	29	0.290
total	100	1
H(X)	0.869	

ב.

	total	probability
TRUE	13	0.130
FALSE	87	0.870
total	100	1
H(X)	0.557	

predict\slope	1	2	3	total
0	39	12	6	57
1	7	10	2	19
2	2	7	1	10
3	2	6	2	10
4	0	4	0	4
total	50	39	11	100

P(predict, slope)	1	2	3	total
0	0.390	0.120	0.060	0.570
1	0.070	0.100	0.020	0.190
2	0.020	0.070	0.010	0.100
3	0.020	0.060	0.020	0.100
4	0.000	0.040	0.000	0.040
total	0.500	0.390	0.110	1.000

P(predict|slope)	1	2	3
0	0.780	0.308	0.545
1	0.140	0.256	0.182
2	0.040	0.179	0.091
3	0.040	0.154	0.182
4	0.000	0.103	0.000
total	1.000	1.000	1.000

ג.

p(slope, prediction) * log(p(prediction|slope))	1	2	3	total
0	0.140	0.204	0.052	0.396
1	0.199	0.196	0.049	0.444
2	0.093	0.173	0.035	0.301
3	0.093	0.162	0.049	0.304
4	0.000	0.131	0.000	0.131
total	0.524	0.867	0.185	1.577

H(prediction|slope)= -∑[p(slope,prediction) * log (p(prediction|slope))]

ד.

max H(slope) 1.585

האנטרופיה המקסימלית היא:

log(number of total possible values)

ה.

prediction\gender	0	1	total
0	23	34	57
1	1	18	19
2	1	9	10
3	3	7	10
4	1	3	4
total	29	71	100

p(prediction,gender)	0	1	total
0	0.230	0.340	0.570
1	0.010	0.180	0.190
2	0.010	0.090	0.100
3	0.030	0.070	0.100
4	0.010	0.030	0.040
total	0.290	0.710	1.000

P(prediction|gender)	0	1
0	0.793	0.479
1	0.034	0.254
2	0.034	0.127
3	0.103	0.099
4	0.034	0.042
total	1.000	1.000

p(gender, prediction) * log(p(prediction|gender))	0	1	total
0	0.077	0.361	0.438
1	0.049	0.356	0.405
2	0.049	0.268	0.317
3	0.098	0.234	0.332
4	0.049	0.137	0.186
total	0.321	1.357	1.677

ה1.

prediction	total	probability	
0	57.000	0.570	0.462
1	19.000	0.190	0.455
2	10.000	0.100	0.332
3	10.000	0.100	0.332
4	4.000	0.040	0.186
total	100.000	1.000	1.768

H(prediction)	1.768

I(gender; prediction) = H(prediction) - H(prediction|gender)

0.090

ה2.

$$\sum_{x,y} p(x,y) \bullet \log \frac{p(y/x)}{p(y)}$$

p(gender, prediction)/ log((p(predication|gender)/p(prediction)	0	1	total
0	0.110	-0.085	0.024
1	-0.025	0.075	0.050
2	-0.015	0.031	0.015
3	0.001	-0.001	0.000
4	-0.002	0.002	0.000
total	0.069	0.021	0.090

3.א

root	0	1	2	3	4	sum
	57.000	19.000	10.000	10.000	4.000	100
Pi	0.570	0.190	0.100	0.100	0.040	
log(Pi)	-0.811	-2.396	-3.322	-3.322	-4.644	
Pi*log(Pi)	-0.462	-0.455	-0.332	-0.332	-0.186	
$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$						
Info(root)	1.768					

חישוב אנטרופיה של השורש:

$$Info(prediction) = -0.57 * log_2(0.57) - 0.19 * log_2(0.19) - 0.1 * log_2(0.1) - 0.1 * log_2(0.1) - 0.04 * log_2(0.04) = \mathbf{1.768}$$

חישוב information gain של פיצול לפי gender:

$$\mathrm{IG(prediction, Gender)} = \mathrm{Info(prediction)} - \mathrm{Info_{gender}(prediction)}$$

נקבע וקטור לערכי הרשומות במשתנה המטרה:

$$(\#0, \#1, \#2, \#3, \#4)$$

Gender/prediction	0	1	2	3	4	sum
0	23.000	1.000	1.000	3.000	1.000	29.000
1	34.000	18.000	9.000	7.000	3.000	71.000
Pi	0.793	0.034	0.034	0.103	0.034	
	0.479	0.254	0.127	0.099	0.042	

$$\mathrm{Info(Prediction_0)} = -\mathrm{P}_{0|0} * \log_2(P_{0|0}) - \mathrm{P}_{1|0} * \log_2(P_{1|0}) - \mathrm{P}_{2|0} * \log_2(P_{2|0}) - \mathrm{P}_{3|0} * \log_2(P_{3|0}) - \mathrm{P}_{4|0} * \log_2(P_{4|0}) = -0.793 * \log_2(0.793) - 0.034 * \log_2(0.034) - 0.034 * \log_2(0.034) - 0.103 * \log_2(0.103) - 0.034 * \log_2(0.034) = \mathbf{1.106}$$

$$\mathrm{Info(Prediction_1)} = -\mathrm{P}_{0|1} * \log_2(P_{0|1}) - \mathrm{P}_{1|1} * \log_2(P_{1|1}) - \mathrm{P}_{2|1} * \log_2(P_{2|1}) - \mathrm{P}_{3|1} * \log_2(P_{3|1}) - \mathrm{P}_{4|1} * \log_2(P_{4|1}) = -0.479 * \log_2(0.479) - 0.254 * \log_2(0.254) - 0.127 * \log_2(0.127) - 0.099 * \log_2(0.099) - 0.042 * \log_2(0.042) = \mathbf{1.911}$$

$$\text{Info}_{\text{gender}}(\text{prediction}) = \sum\left(\frac{|prediction_i|}{|prediction|} \cdot Info(prediction_i)\right)$$
$$= 0.29 * 1.106 + 0.71 * 1.911 = 0.321 + 1.357 = \mathbf{1.678}$$

$$\text{IG}(\text{prediction}, \text{Gender}) = \text{Info}(\text{prediction}) - \text{Info}_{\text{gender}}(\text{prediction}) =$$
$$1.768 - 1.678 = \mathbf{0.09}$$

חישוב information gain של פיצול לפי fbs:

$$\text{IG}(\text{prediction}, \text{fbs}) = \text{Info}(\text{prediction}) - \text{Info}_{\text{fbs}}(\text{prediction}) =$$
$$1.768 - 1.745 = \mathbf{0.023}$$

חישוב information gain של פיצול לפי slope:

$$\text{IG}(\text{prediction}, \text{fbs}) = \text{Info}(\text{prediction}) - \text{Info}_{\text{slope}}(\text{prediction}) =$$
$$1.768 - 1.577 = \mathbf{0.191}$$

נבחר לפצל לפי המשתנה slope מכיוון שהוא מספק את רווח הידע הגדול ביותר (הקטנת אי הוודאות הגדולה ביותר).

חישוב אנטרופיה של הקדקוד בו slope=1:
האנטרופיה של קדקוד זה חושבה בשלב הקודם: $info(prediction_{slope=1}) = 1.048$
חישוב information gain של פיצול לפי gender:

$$\text{IG}\left(prediction_{slope=1}, \text{Gender}\right) = info(prediction_{slope=1}) - \text{Info}_{\text{gender}}(prediction_{slope=1})$$

$$\text{Info}\left(\text{Prediction}_{slope=1,gender=0}\right) = -\text{P}_{0|0} * \log_2(P_{0|0}) - \text{P}_{1|0} * \log_2(P_{1|0}) - \text{P}_{2|0} * \log_2(P_{2|0}) - \text{P}_{3|0} * \log_2(P_{3|0}) - \text{P}_{4|0} * \log_2(P_{4|0}) =$$
$$-0.944 * \log_2(0.944) - 0.056 * \log_2(0.056) - 0 - 0 - 0 = \mathbf{0.31}$$

$$\text{Info}\left(\text{Prediction}_{slope=1,gender=1}\right) = -\text{P}_{0|1} * \log_2(P_{0|1}) - \text{P}_{1|1} * \log_2(P_{1|1}) - \text{P}_{2|1} * \log_2(P_{2|1}) - \text{P}_{3|1} * \log_2(P_{3|1}) - \text{P}_{4|1} * \log_2(P_{4|1}) =$$
$$-0.688 * \log_2(0.688) - 0.188 * \log_2(0.188) - 0.063 * \log_2(0.063) - 0.063 * \log_2(0.063) - 0 = \mathbf{1.324}$$

$$\text{Info}_{\text{gender}}\left(prediction_{slope=1}\right) = \sum\left(\frac{\left|prediction_{slope=1,i}\right|}{\left|prediction_{slope=1}\right|} \cdot Info\left(prediction_{slope=1,i}\right)\right)$$
$$= \frac{18}{50} * 0.31 + \frac{32}{50} * 1.324 = \mathbf{0.959}$$

$$\mathrm{IG}(prediction_{slope=1}, \mathrm{Gender}) = \mathrm{Info}(prediction_{slope=1}) - \mathrm{Info}_{\mathrm{gender}}(prediction_{slope=1}) =$$
$$1.048 - 0.959 = \mathbf{0.089}$$

חישוב information gain של פיצול לפי fbs:

$$\mathrm{IG}(prediction_{slope=1},, \mathrm{fbs}) = \mathrm{Info}(prediction_{slope=1},) - \mathrm{Info}_{\mathrm{fbs}}(prediction_{slope=1},) =$$
$$1.048 - 0.99 = \mathbf{0.058}$$

נבחר לפצל לפי המשתנה fbs מכיוון שהוא מספק את רווח הידע הגדול ביותר (הקטנת אי הוודאות הגדולה ביותר).

חישוב אנטרופיה של הקדקוד בו slope=2:

האנטרופיה של קדקוד זה חושבה בשלב הקודם: $info(prediction_{slope=2}) = 2.224$

חישוב information gain של פיצול לפי gender:

$$\mathrm{IG}(prediction_{slope=2}, \mathrm{Gender}) = info(prediction_{slope=2}) - \mathrm{Info}_{\mathrm{gender}}(prediction_{slope=2})$$

$$\mathrm{Info}(\mathrm{Prediction}_{slope=2,gender=0}) = -\mathrm{P}_{0|0} * \log_2(P_{0|0}) - \mathrm{P}_{1|0} * \log_2(P_{1|0}) - \mathrm{P}_{2|0} * \log_2(P_{2|0}) - \mathrm{P}_{3|0} * \log_2(P_{3|0}) - \mathrm{P}_{4|0} * \log_2(P_{4|0}) =$$
$$-0.571 * \log_2(0.571) - 0 - 0.143 * \log_2(0.143) - 0.143 * \log_2(0.143) - 0.143 * \log_2(0.143) = \mathbf{1.664}$$

$$\mathrm{Info}(\mathrm{Prediction}_{slope=2,gender=1}) = -\mathrm{P}_{0|1} * \log_2(P_{0|1}) - \mathrm{P}_{1|1} * \log_2(P_{1|1}) - \mathrm{P}_{2|1} * \log_2(P_{2|1}) - \mathrm{P}_{3|1} * \log_2(P_{3|1}) - \mathrm{P}_{4|1} * \log_2(P_{4|1}) =$$
$$-0.25 * \log_2(0.25) - 0.313 * \log_2(0.313) - 0.188 * \log_2(0.188) - 0.156 * \log_2(0.156) - 0.094 * \log_2(0.094) = \mathbf{2.216}$$

$$\mathrm{Info}_{\mathrm{gender}}(prediction_{slope=2}) = \sum\left(\frac{|prediction_{slope=2,i}|}{|prediction_{slope=2}|} \cdot Info(prediction_{slope=2,i})\right)$$
$$= \frac{7}{39} * 1.664 + \frac{32}{39} * 2.216 = \mathbf{2.117}$$

$$\mathrm{IG}(prediction_{slope=1}, \mathrm{Gender}) = \mathrm{Info}(prediction_{slope=1}) - \mathrm{Info}_{\mathrm{gender}}(prediction_{slope=1}) =$$
$$2.224 - 2.117 = \mathbf{0.107}$$

חישוב information gain של פיצול לפי fbs:

$$\mathrm{IG}(prediction_{slope=2}, \mathrm{fbs}) = \mathrm{Info}(prediction_{slope=2},) - \mathrm{Info}_{\mathrm{fbs}}(prediction_{slope=2},) =$$
$$2.224 - 2.122 = \mathbf{0.102}$$

נבחר לפצל לפי המשתנה gender מכיוון שהוא מספק את רווח הידע הגדול ביותר (הקטנת אי הוודאות הגדולה ביותר).

חישוב אנטרופיה של הקדקוד בו slope=3:

האנטרופיה של קדקוד זה חושבה בשלב הקודם: $info(prediction_{slope=3}) = 1.686$

חישוב information gain של פיצול לפי gender:

$$\mathrm{IG}(prediction_{slope=3}, \mathrm{Gender}) = info(prediction_{slope=3}) - \mathrm{Info}_{\mathrm{gender}}(prediction_{slope=3})$$

$$\text{Info}\left(\text{Prediction}_{slope=3,gender=0}\right) = -\text{P}_{0|0} * \log_2(P_{0|0}) - \text{P}_{1|0} * \log_2(P_{1|0}) - \text{P}_{2|0} * \log_2(P_{2|0}) - \text{P}_{3|0} * \log_2(P_{3|0}) - \text{P}_{4|0} * \log_2(P_{4|0}) =$$
$$-0.5 * \log_2(0.5) - 0 - 0 - 0.5 * \log_2(0.5) - 0 = \mathbf{1}$$

$$\text{Info}\left(\text{Prediction}_{slope=3,gender=1}\right) = -\text{P}_{0|1} * \log_2(P_{0|1}) - \text{P}_{1|1} * \log_2(P_{1|1}) - \text{P}_{2|1} * \log_2(P_{2|1}) - \text{P}_{3|1} * \log_2(P_{3|1}) - \text{P}_{4|1} * \log_2(P_{4|1}) =$$
$$-0.571 * \log_2(0.571) - 0.286 * \log_2(0.286) - 0.143 * \log_2(0.143) - 0 - 0 = \mathbf{1.379}$$

$$\text{Info}_{\text{gender}}\left(prediction_{slope=3}\right) = \sum\left(\frac{\left|prediction_{slope=3,i}\right|}{\left|prediction_{slope=3}\right|} \cdot Info\left(prediction_{slope=3,i}\right)\right)$$
$$= \frac{4}{11} * 1 + \frac{7}{11} * 1.379 = \mathbf{1.241}$$

$$\text{IG}\left(prediction_{slope=1}, \text{Gender}\right) = \text{Info}\left(prediction_{slope=1}\right) - \text{Info}_{\text{gender}}\left(prediction_{slope=1}\right) =$$
$$1.686 - 1.241 = \mathbf{0.445}$$

<u>חישוב information gain של פיצול לפי fbs:</u>

$$\text{IG}\left(prediction_{slope=2}, \text{fbs}\right) = \text{Info}\left(prediction_{slope=2},\right) - \text{Info}_{\text{fbs}}\left(prediction_{slope=2},\right) =$$
$$1.686 - 1.523 = \mathbf{0.163}$$

נבחר לפצל לפי המשתנה gender מכיוון שהוא מספק את רווח הידע הגדול ביותר (הקטנת אי הוודאות הגדולה ביותר).