חלק ב' [50 נקודות]

יש להציג את כל התוצאות עם שלוש ספרות אחרי נקודה עשרונית אלא אם צוין אחרת!

נתונה טבלת נתונים של לקוחות בחברה סלולארית. מטרת הטבלה לחזות את הסתברות הנטישה (churn) של לקוח.

מס' רשומה	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
דקות שיחה בשנה הקודמת	1029	658	1680	77	1176	1141	490	798	595	2268	294	1407	2534	994	2093
דקות שיחה בשנה האחרונה	3216	3156	2952	0	2912	164	2060	2892	2044	0	0	1204	292	228	3560
נטישה	לא	לא	לא	לא	לא	לא	לא	לא	לא	לא	כן	כן	כן	כן	כן

א. יש לחשב את האנטרופיה המותנית, מדד ה-Gini ומדד ה-twoing של משתנה המטרה "נטישה" עבור שתי נקודות הפיצול הבאות של המשתנה "**דקות שיחה בשנה האחרונה**": 228, 1204. **24 נקודות.**

Threshold	Conditional Entropy	Gini Index	Twoing
228			
1204			

ב. עפ"י החישובים בסעיף הקודם, מה יהיה השינוי בהסתברות הנטישה כתוצאה מירידת דקות השיחה בשנה האחרונה מתחת לכל ערך סף? **4 נקודות.**

Threshold	Churn Probability Change
228	
1204	

ג. יש לנרמל את כל הערכים של המשתנים "דקות שיחה בשנה האחרונה" ו"דקות שיחה בשנה הקודמת" בשיטת z-score normalization תוך חישוב סטיית התקן של אוכלוסיה בגודל טבלת הנתונים. **בעמודה האחרונה** יש לסכם את הערכים המנורמלים של כל משתנה. **10 נקודות**.

מס' רשומה	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	סה"כ
דקות שיחה בשנה הקודמת																
דקות שיחה בשנה האחרונה																

ד. יש לסווג את התצפיות המנורמלות בעזרת המודל Linear SVM בעל הפרמטרים הבאים:

```
Total number of Support Vectors: 15
Bias (offset): -0.346
w[דקות שיחה בשנה הקודמת] = 0.429
w[דקות שיחה בשנה האחרונה] = -0.506
```

שימו לב: המודל מניח שסיווג הנטישה הוא "**1**" בשעה שאי-הנטישה מיוצגת ע"י "**-1**".

בעמודה האחרונה יש לסכם את מספר השגיאות של המודל. **8 נקודות**

מס' רשומה	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	סה"כ שגיאות
סיווג חזוי																
שגיאה (כן /לא)																

ה. יש לחשב את רווח בר-סמך לדיוק המודל Linear SVM ברמת-ביטחון של 95%.

4 נקודות.

Upper Bound of Accuracy	Lower Bound of Accuracy