

## מבחן בקורס "מבוא לכריית נתונים"

מועד ב'

סמסטר ב', תשע"ח

מרצה: ניבה חזון

חומר עזר: דף נוסחאות (דף אחד משני הצדדים) ומחשבון מדעי  
ללא יכולות תכנות.

השאלון מכיל 8 דפים (כולל דף זה).

יש לענות בתוך השאלון בלבד (דפים נוספים לא יבדקו)!

בהצלחה ☺



## חלק א' (40 נקודות, 4 נקודות לכל שאלה)

יש לבחור תשובה אחת בלבד לכל שאלה.

1. אדם ביצע בדיקה לנוכחות מחלה מסוימת. הבדיקה יצאה חיובית למרות שהאדם בפועל אינו חולה. זוהי דוגמה ל:

א. False negative

ב. False positive

ג. True positive

ד. True negative

2. המרחק בין שני אשכולות בשיטת complete link מחושב עפ"י

א. המרחק בין מרכזי האשכולות

ב. המרחק הארוך ביותר בין שני אובייקטים בשני האשכולות

ג. המרחק הממוצע בין שני האשכולות

ד. המרחק הקצר ביותר בין שני אובייקטים בשני האשכולות

3. חלוקת ערכים לאינטרוולים לפי רוחב שווה

א. מקטינה את שונות הנתונים

ב. יוצרת אינטרוולים בעלי שכיחות שווה

ג. יוצרת אינטרוולים בעלי טווח שווה

ד. תשובות א' ו-ג' נכונות

4. עפ"י תורת האינפורמציה, אי וודאות האירוע שווה למקסימום אם

א. הסתברות אחת התוצאות שווה ל-1

ב. הסתברות אחת התוצאות שווה ל-0

ג. התוצאות מתפלגות התפלגות נורמאלית

ד. התוצאות מתפלגות התפלגות אחידה

5. מהי ההנחה הבסיסית של אלגוריתם Naïve Bayes?

א. קיימת תלות זהה בין כל משתנה למשתנה ולכן ניתן לכפול ביניהם

ב. לא קיימת תלות בין אף אחד מהמשתנים ולכן ניתן לכפול ביניהם

ג. קיימות תלויות שונות בין משתנים שונים ולכן לא ניתן לכפול ביניהם

ד. לא קיימת תלות בין אף אחד מהמשתנים ולכן לא ניתן לכפול ביניהם

6. אלגוריתם מסוג eager לעומת אלגוריתם מסוג lazy:

א. eager שומר את כל נתוני האימון ו-lazy שומר חלק מנתוני האימון

ב. eager שומר חלק מנתוני האימון ו-lazy שומר את כל נתוני האימון

ג. eager שומר את נתוני האימון ו-lazy שומר מודל

ד. eager שומר מודל ו-lazy שומר את נתוני האימון

7. בית חולים מעוניין להעריך את משך האשפוז של חולה מסוים. מדובר במשימה של:

א. ניתוח אשכולות (Clustering)

ב. סיווג (Classification)

ג. חיזוי (Prediction)

ד. אף תשובה אינה נכונה



8. כדי לדעת אילו פריטים נרכשים בדר"כ ביחד, המדד שצריך לחשב הוא:

- א. confidence
- ב. mutual information
- ג. conditional probability
- ד. support

9. בבעיית סיווג בינארי, דיוק האימון של מודל בעל אנטרופיה השווה ל-0 הוא:

- א. 50%
- ב. 0%
- ג. 100%
- ד. לא ניתן לדעת

10. תצפית חריגה במודל אשכולות (clustering):

- א. קרובה למרכז של אשכול מסוים
- ב. רחוקה ממרכזי כל האשכולות
- ג. רחוקה ממרחק אשכול מסוים
- ד. קרובה למרכז כל האשכולות



## חלק ב' (60 נקודות)

בכל הסעיפים בחלק זה יש להראות את כל החישובים הרלוונטיים במקומות המיועדים לכך בלבד (דפים נוספים לא יבדקו).

רופאים מ - Cleveland Clinic Foundation מעוניינים לחזות נוכחות של מחלת לב בחולים. לשם כך, נאסף מידע מ 10 חולים המכיל 5 משתנים מועמדים ומשתנה מטרה אחד - prediction. הנתונים בטבלה:

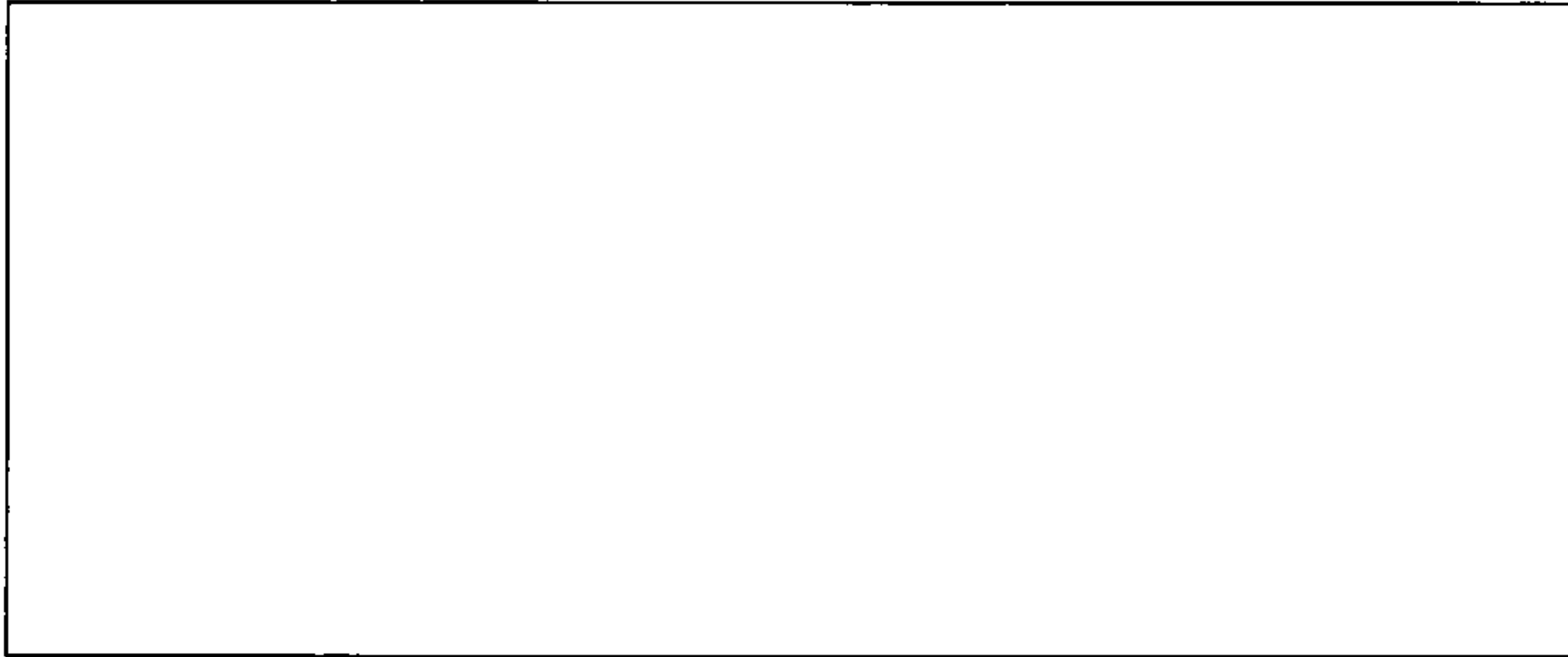
record	gender	bs	cp	fbs	slope	prediction
1	female	69	typical angina	TRUE	downsloping	absence
2	male	14	asymptomatic	FALSE	upsloping	presence
3	male	98	asymptomatic	FALSE	downsloping	presence
4	male	54	typical angina	TRUE	downsloping	absence
5	female	36	typical angina	FALSE	upsloping	absence
6	female	4	anginal pain	TRUE	upsloping	absence
7	female	77	asymptomatic	FALSE	downsloping	presence
8	female	44	anginal pain	FALSE	upsloping	absence
9	male	23	asymptomatic	TRUE	upsloping	presence
10	female	21	anginal pain	FALSE	downsloping	absence

א. (8 נק') יש לבצע נרמול בשיטת z-score לערכי המשתנה "bs".

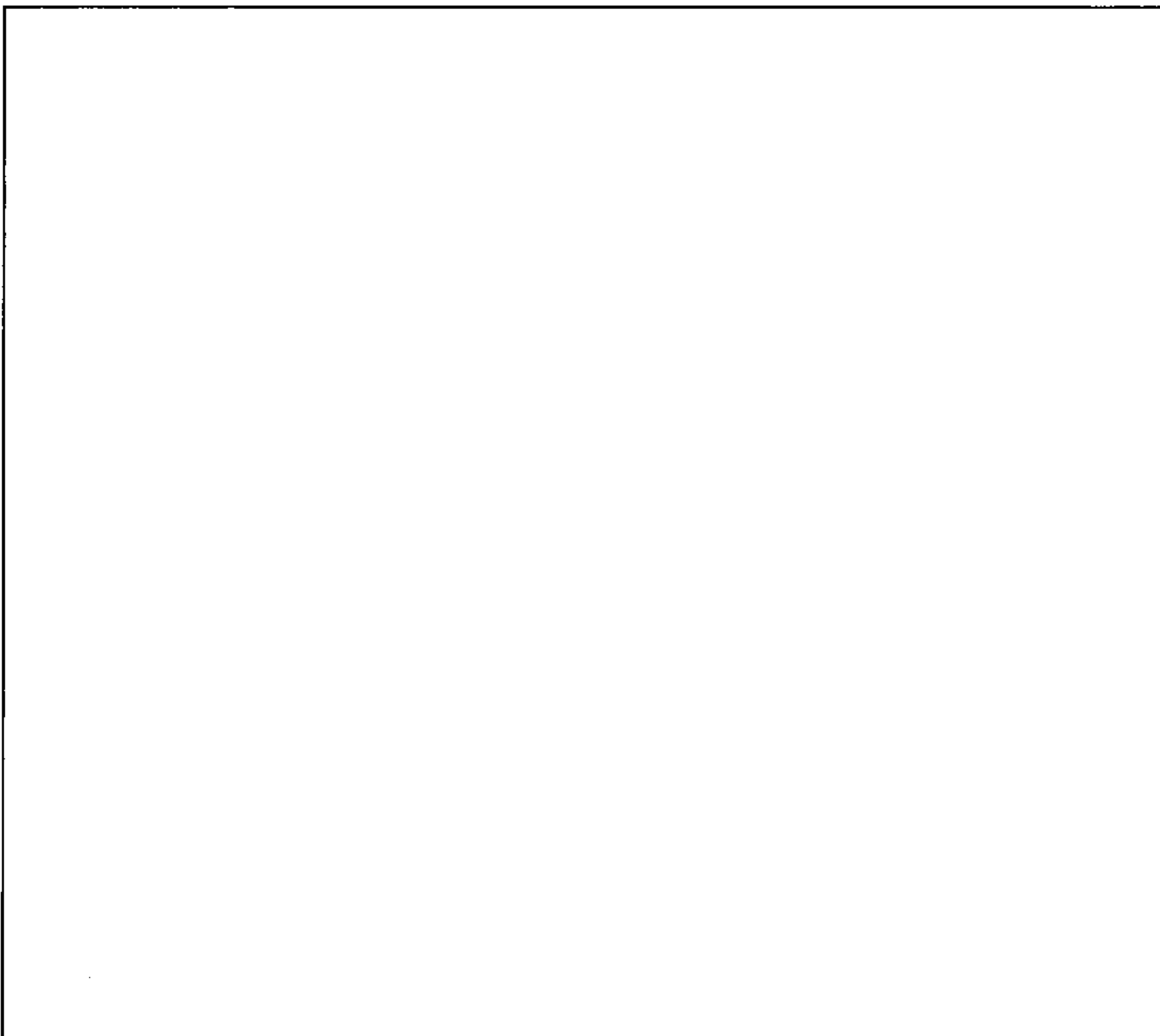




ב. (6 נק') יש לבצע דיסקרטיזציה לשני טווחים בשיטת עומק שווה לערכי המשתנה "bs" המנורמל.



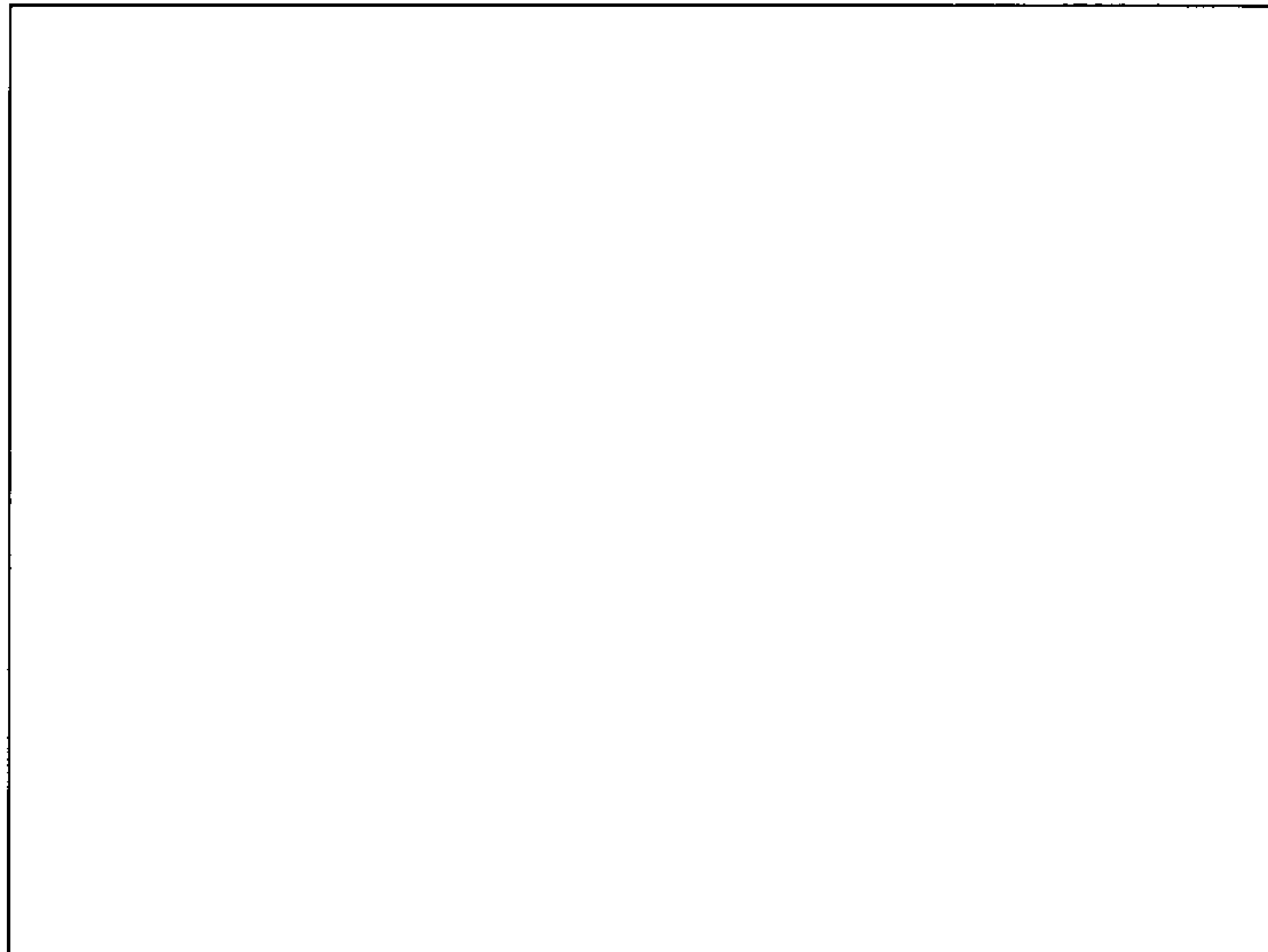
ג. (10 נק') יש לחשב את מדד ה-Gini (Gini index) של משתנה המטרה "prediction" עבור כל אחת מהתכונות "slope" ו- "fbs" ולהחליט לפי איזו תכונה עדיף לפצל את קדקוד השורש.





ד. (12 נק') יש לבצע מבחן חי בריבוע ( $\chi^2$ ) למשתנה "cp" ולהחליט האם כדאי לפצל את קדקוד השורש לפי משתנה זה.

דרגות חופש (DF)	1	2	3	4	5
$\chi^2$	3.841	5.991	7.815	9.488	11.070



ה. (12 נק') יש להשתמש באלגוריתם KNN (K nearest neighbors) על מנת לסווג את התצפית הבאה (K=1):

record	gender	bs	cp	fbs	slope	prediction
1	female	50	typical angina	False	downsloping	

- על מנת לחשב מרחק בין תצפיות יש להשתמש במרחק מנהטן כאשר המרחקים מוגדרים באופן הבא:
- עבור המשתנים gender, fbs, slope המרחק בין ערכים שונים הוא 1.
  - עבור המשתנה "bs" יש להשתמש במשתנה לאחר הדיסקרטיזציה (מסעיף ב') – המרחק בין אינטרוולים שונים הוא 1.
  - עבור המשתנה "cp" יש להשתמש בטבלת המרחקים הבאה:



cp	anginal pain	typical angina	asymptomatic
anginal pain	0		
typical angina	1	0	
asymptomatic	2	1	0

1. (12 נק') יש לבצע אלגוריתם חלוקה לאשכולות בצורה הירארכית (Hierarchical Clustering) לפי שיטת AGNES (Agglomerative Nesting) על תצפיות 1-4 מנתוני האימון. יש לצייר בכל שלב את הדנדוגרמה המתאימה ולהשתמש ב-single link. (יש למלא את טבלת המרחקים המצורפת, אין צורך למלא תאים אפורים).

record	1	2	3
2			
3			
4			



