# מטלת מנחה (ממ"ן) 12

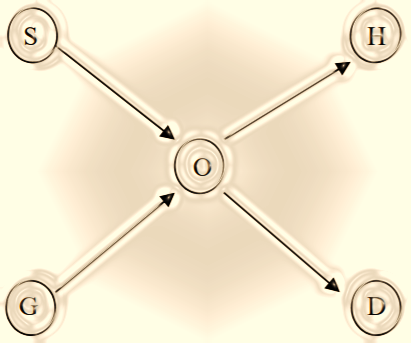
**שאלה 1 – רשת בייסיאנית – חישוב הסתברויות**

מבוא - תיאור הרשת וההסתברויות הנתונות

יהיו המשתנים המקריים הבאים –

*  - השמנה בטנית
*  - אכילה מופרזת של סוכרים
*  - תורשה
*  - יתר לחץ דם
*  - מחלת סוכרת

והרשת הבייסיאנית שלהלן –



הרשת מתארת את העובדות הבאות –

* השמנה בטנית () מתפתחת כתוצאה מאכילה מופרזת של סוכרים (( ו/או כתוצאה מתורשה ().
* השמנה בטנית () גורמת ליתר לחץ דם () ו/או למחלת סוכרת ().

נניח שמתקיים –



וכן נגדיר את ההסתברויות המותנות הבאות –



סעיף א – חישוב 

נחשב את ההסתברות של נבדק ללקות בהשמנה בטנית, דהיינו את. הסתברות זו נתונה ע"י –



נציב במשוואה זו את ערכי ההסתברויות הנתונות לעיל, כאשר במקרה של הסתברות מאורע משלים  נציב במקומו את ההסתברות השקולה , ונקבל –



**מכאן שההסתברות של נבדק ללקות בהשמנה בטנית הינה **

סעיף ב – חישוב 

נחשב את ההסתברות שנבדק בעל גורם תורשתי () ילקה בהשמנה בטנית ((, דהיינו את. הסתברות זו נתונה ע"י –



ההסתברות במכנה כבר נתונה מראש: .

נותר לחשב את ההסתברות שבמונה. ראשית נבחין כי מהרשת נובע כי השמנה בטנית (( מושפעת לא רק מהגורם התורשתי (), אלא גם מאכילה/אי אכילה מופרזת של סוכרים (). נרשום בטבלה את ההתפלגות המותנית של  ביחס ל- ו- כפי שנובע מהנתונים לעיל –

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

מעניינים אותנו השורות שבהן הנבדק הינו בעל גורם תורשתי, דהיינו השורות שבהן : שורות 1 ו-3.

נחשב את ההסתברות של איחוד מאורעות זרים אלו ונקבל –



נציב את התוצאה שהתקבלה במקום המונה בביטוי המקורי ונקבל –



**מכאן שההסתברות של נבדק בגל גורם תורשתי ללקות בהשמנה בטנית הינה **

**שאלה 2 – אשכול היררכי אגלומרטיבי**

נבצע אשכול היררכי אגלומרטיבי (Bottom-Up) לנתוני האימון המטוייבים ממטלה 11 שלהלן –

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| מספר רשומה | גיל נבדק | משקל | תזונה | ספורט | מצב סוציואקונומי |
| 1 | 58 | גבוה | רגילה | ל | נמוך |
| 2 | 60 | גבוה | רגילה | כ | גבוה |
| 3 | 32 | גבוה | צמחונית | ל | נמוך |
| 4 | 58 | תקין | רגילה | כ | גבוה |
| 5 | 64 | נמוך | צמחונית | כ | גבוה |
| 6 | 39 | נמוך | רגילה | ל | נמוך |
| 7 | 70 | נמוך | צמחונית | כ | גבוה |
| 8 | 27 | תקין | צמחונית | ל | נמוך |
| 9 | 45 | נמוך | רגילה | ל | גבוה |
| 10 | 64 | תקין | רגילה | כ | גבוה |
| 11 | 48 | תקין | רגילה | כ | גבוה |
| 12 | 62 | תקין | צמחונית | כ | גבוה |
| 13 | 50 | גבוה | רגילה | ל | גבוה |
| 14 | 24 | תקין | צמחונית | ל | נמוך |

בניית ה-Dissimilarity matrix

לצורך האשכול עלינו לבנות את טבלת השונויות/מרחקים (Dissimilarity matrix). אופן מדידת המרחק בין 2 רשומות עבור תכונה כלשהי תלוי בסוג/טיפוס התכונה. להלן טיפוסי התכונות שלעיל –

|  |  |  |
| --- | --- | --- |
| # | תכונה | טיפוס |
| 1 | גיל נבדק | נומרי (רציף) |
| 2 | משקל | אורדינלי (גבוה, תקין, נמוך) |
| 3 | תזונה | נומינלי (רגילה, צמחונית) |
| 4 | ספורט | בינארי סימטרי [[1]](#footnote-1)(כ = כן, ל = לא) |
| 5 | מצב סוציואקונומי | נומינלי (גבוה, נמוך) |

עלינו לבצע טרנספורמציה על הנתונים שלעיל על מנת שנוכל להשוות בניהם ולמדוד מרחקים. נתחיל עם התכונה האורדינלית של משקל. ראשית נקצה את הערך 1 לנמוך, 2 לתקין ו-3 לגבוה, ואז ננרמל את הערכים לטווח שבין 0 ל-1 ע"י שימוש בנוסחה (2.21) [מספר הלימוד](http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf) (עמוד 74) –



נקבל את הנרמול שלהלן –

|  |  |  |  |
| --- | --- | --- | --- |
| **משקל** | **ערך נומינלי** | **ערך מנורמל** | **נוסחה** |
| נמוך | 1 | 0 | (1-1)/2 |
| תקין | 2 | 0.5 | (2-1)/2 |
| גבוה | 3 | 1 | (3-1)/2 |

הפירוט נמצא בקובץ אקסל שלהלן –



כעת נטפל בתכונה הנומרית של הגיל. נבצע נרמול לטווח ערכים שבין 0 ל-1 בשיטת ה-Min-max normalization תוך שימוש במשוואה (3.8) מעמוד 114 בספר הלימוד, כאשר המקסימום החדש הוא 1 והמינימום החדש הוא 0. נקבל –



נפעיל נוסחה זו של נרמול על כל הערכים ונציב בטבלה, יחד עם הערכים המנורמלים של המשקל. נקבל את הטבלה המעודכנת שלהלן (ערכי הגיל עוגלו ל-2 ספרות עשרוניות) –

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| מספר רשומה | גיל נבדק | משקל | תזונה | ספורט | מצב סוציואקונומי |
| 1 | 0.74 | 1 | רגילה | ל | נמוך |
| 2 | 0.78 | 1 | רגילה | כ | גבוה |
| 3 | 0.17 | 1 | צמחונית | ל | נמוך |
| 4 | 0.74 | 0.5 | רגילה | כ | גבוה |
| 5 | 0.87 | 0 | צמחונית | כ | גבוה |
| 6 | 0.33 | 0 | רגילה | ל | נמוך |
| 7 | 1 | 0 | צמחונית | כ | גבוה |
| 8 | 0.07 | 0.5 | צמחונית | ל | נמוך |
| 9 | 0.46 | 0 | רגילה | ל | גבוה |
| 10 | 0.87 | 0.5 | רגילה | כ | גבוה |
| 11 | 0.52 | 0.5 | רגילה | כ | גבוה |
| 12 | 0.83 | 0.5 | צמחונית | כ | גבוה |
| 13 | 0.57 | 1 | רגילה | ל | גבוה |
| 14 | 0 | 0 | צמחונית | ל | נמוך |

כעת נוכל לחשב בנקל את המרחקים בין הווקטורים השונים לפי נוסחה (2.22) שבתחתית עמוד 75 בספר הלימוד, המספקת דרך למדוד מרחק בין 2 וקטורים של תכונות מטיפוסים שונים. לפי סימון הנוסחה, לכל  יתקיים , שכן אין ערכים חסרים ואין משתנים בינאריים א-סימטריים ששניהם שליליים. לפיכך, כל שעלינו לעשות להשוות בין 2 רשומות הוא לחשב את המרחק בין 5 התכונות שלהן ולחלק ב-5 (מספר התכונות הכולל): עבור תכונה נומרית נחשב את המרחק ע"י חישוב הערך המוחלט שבין מופעי 2 הרשומות מחולק בערך בהפרש שבין הערך המקסימאלי למינימאלי בתכונה זו מבין סך הרשומות, ועבור תכונה בינארית/נומינלית המרחק יהיה 0 אם הערכים זהים ו-1 אם יש הבדל. נבחין כי לאחר הנרמול תכונת המשקל האורדינלית היא כעת נומרית.

למען נוחות בחישוב, כדי שנוכל לחשב בקלות באקסל, נמיר כל תכונה נומינלית עם 2 ערכים לתכונה בינארית עם ערכים 0 ו-1. כעת, ומאחר שהערכים המקסימליים של כלל התכונות הם 1 והמינימליים הם 0, כל שנדרש לחשב את המרחקים הוא לסכום את הערכים המוחלטים שבין 2 רשומות ולחלק במספר התכונות הכולל.

בעמוד הבא נתונה מטריצת המרחקים המלאה (Dissimilarity matrix) הכוללת את ערכי כל החישובים.

החישובים עצמם מפורטים בקובץ אקסל שלהלן –



להלן מטריצת המרחקים של רשומות נתוני האימון מסט הנתונים המטוייב שלעיל. לטובת בהירות אנחנו משמיטים את הערכים שמעל האלכסון הראשי של המטריצה, שכן ממילא המטריצה סימטרית, כלומר לכל  ממילא מתקיים  או בכתיב מטריצות – , לפיכך נוכל להסתפק במטריצה יותר קומפקטית –

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14 | 13 | 12 | 11 | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  | 0 | 1 |
|  |  |  |  |  |  |  |  |  |  |  |  | 0 | 0.408 | 2 |
|  |  |  |  |  |  |  |  |  |  |  | 0 | 0.722 | 0.314 | 3 |
|  |  |  |  |  |  |  |  |  |  | 0 | 0.814 | 0.108 | 0.5 | 4 |
|  |  |  |  |  |  |  |  |  | 0 | 0.326 | 0.74 | 0.418 | 0.826 | 5 |
|  |  |  |  |  |  |  |  | 0 | 0.708 | 0.582 | 0.432 | 0.69 | 0.282 | 6 |
|  |  |  |  |  |  |  | 0 | 0.734 | 0.026 | 0.352 | 0.766 | 0.444 | 0.852 | 7 |
|  |  |  |  |  |  | 0 | 0.686 | 0.352 | 0.66 | 0.734 | 0.12 | 0.842 | 0.434 | 8 |
|  |  |  |  |  | 0 | 0.578 | 0.508 | 0.226 | 0.482 | 0.356 | 0.658 | 0.464 | 0.456 | 9 |
|  |  |  |  | 0 | 0.382 | 0.76 | 0.326 | 0.608 | 0.3 | 0.026 | 0.84 | 0.118 | 0.526 | 10 |
|  |  |  | 0 | 0.07 | 0.312 | 0.69 | 0.396 | 0.538 | 0.37 | 0.044 | 0.77 | 0.152 | 0.544 | 11 |
|  |  | 0 | 0.262 | 0.208 | 0.574 | 0.552 | 0.134 | 0.8 | 0.108 | 0.218 | 0.632 | 0.31 | 0.718 | 12 |
|  | 0 | 0.552 | 0.31 | 0.36 | 0.222 | 0.6 | 0.686 | 0.448 | 0.66 | 0.334 | 0.48 | 0.242 | 0.234 | 13 |
| 0 | 0.714 | 0.666 | 0.804 | 0.874 | 0.492 | 0.114 | 0.6 | 0.266 | 0.574 | 0.848 | 0.234 | 0.956 | 0.548 | 14 |

ביצוע האשכול האגלומרטיבי

כעת משחשבנו את המרחקים שבין הפריטים (ווקטורים), נוכל לבנות את האשכול ההיררכי האגלורמטיבי. לצורך חישוב **המרחק שבין הקלאסטרים** נשתמש ב-**Minimum distance**. אנחנו מתחילים עם קלאסטר אחד לכל פריט, סה"כ 14. **תנאי העצירה** יהיה איחוד כלל הקלאסטרים לקלאסטר אחד.

**שלב ראשון**: נבדוק מהו המרחק המינימאלי בין כל 2 קלאסטרים מה-14 שלעיל. בעזרת הפעלת פונקציית מינימום על טווח הערכים של מטריצת המרחקים באקסל שלעיל נמצא שהמרחק המינימאלי הוא 0.026 . יש מרחק כזה גם בין קלאסטר מספר 4 לבין קלאסטר מספר 10, וגם בין קלאסטר מספר 5 לקלאסטר מספר 7. נבחר באופן שרירותי לטפל ב-4 ו-10 קודם. נאחד אותם לקאלסטר אחד: במטריצת המרחקים נאחדם לשורה אחת, ונעדכן את המרחקים ממנה לכל יתר השורות ולהפך להיות המרחק המינימאלי מבין 2 השורות המקוריות –

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14 | 13 | 12 | 11 | 9 | 8 | 7 | 6 | 5 | 4,10 | 3 | 2 | 1 |  |
|  |  |  |  |  |  |  |  |  |  |  |  | 0 | 1 |
|  |  |  |  |  |  |  |  |  |  |  | 0 | 0.408 | 2 |
|  |  |  |  |  |  |  |  |  |  | 0 | 0.722 | 0.314 | 3 |
|  |  |  |  |  |  |  |  |  | ~~0~~ | ~~0.814~~ | ~~0.108~~ | ~~0.5~~ | ~~4~~ |
|  |  |  |  |  |  |  |  | 0 | 0.3  ~~0.326~~ | 0.74 | 0.418 | 0.826 | 5 |
|  |  |  |  |  |  |  | 0 | 0.708 | 0.582 | 0.432 | 0.69 | 0.282 | 6 |
|  |  |  |  |  |  | 0 | 0.734 | 0.026 | 0.326  ~~0.352~~ | 0.766 | 0.444 | 0.852 | 7 |
|  |  |  |  |  | 0 | 0.686 | 0.352 | 0.66 | 0.734 | 0.12 | 0.842 | 0.434 | 8 |
|  |  |  |  | 0 | 0.578 | 0.508 | 0.226 | 0.482 | 0.356 | 0.658 | 0.464 | 0.456 | 9 |
|  |  |  |  | 0.356  ~~0.382~~ | 0.734  ~~0.76~~ | 0.326 | 0.582  ~~0.608~~ | 0.3 | 0  ~~0.026~~ | 0.814  ~~0.84~~ | 0.108  ~~0.118~~ | 0.5  ~~0.526~~ | 4.10 |
|  |  |  | 0 | 0.312 | 0.69 | 0.396 | 0.538 | 0.37 | 0.044 | 0.77 | 0.152 | 0.544 | 11 |
|  |  | 0 | 0.262 | 0.574 | 0.552 | 0.134 | 0.8 | 0.108 | 0.208 | 0.632 | 0.31 | 0.718 | 12 |
|  | 0 | 0.552 | 0.31 | 0.222 | 0.6 | 0.686 | 0.448 | 0.66 | 0.334 | 0.48 | 0.242 | 0.234 | 13 |
| 0 | 0.714 | 0.666 | 0.804 | 0.492 | 0.114 | 0.6 | 0.266 | 0.574 | 0.848 | 0.234 | 0.956 | 0.548 | 14 |

**שלב שני**: נחזור על פעולת האיחוד עם המרחק המינימאלי של 0.026 הפעם כאמור עם קלאסטרים 5 ו-7 -

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14 | 13 | 12 | 11 | 9 | 8 | 6 | 5, 7 | 4, 10 | 3 | 2 | 1 |  |
|  |  |  |  |  |  |  |  |  |  |  | 0 | **1** |
|  |  |  |  |  |  |  |  |  |  | 0 | 0.408 | **2** |
|  |  |  |  |  |  |  |  |  | 0 | 0.722 | 0.314 | **3** |
|  |  |  |  |  |  |  | ~~0~~ | ~~0.3~~ | ~~0.74~~ | ~~0.418~~ | ~~0.826~~ | **~~5~~** |
|  |  |  |  |  |  | 0 | 0.708 | 0.582 | 0.432 | 0.69 | 0.282 | **6** |
|  |  |  |  |  |  | 0.734 | 0  ~~0.026~~ | 0.3  ~~0.326~~ | 0.74  ~~0.766~~ | 0.418  ~~0.444~~ | 0.826  ~~0.852~~ | **5, 7** |
|  |  |  |  |  | 0 | 0.352 | 0.66 | 0.734 | 0.12 | 0.842 | 0.434 | **8** |
|  |  |  |  | 0 | 0.578 | 0.226 | 0.482 | 0.356 | 0.658 | 0.464 | 0.456 | **9** |
|  |  |  |  | 0.356 | 0.734 | 0.582 | 0.3 | 0 | 0.814 | 0.108 | 0.5 | **4.10** |
|  |  |  | 0 | 0.312 | 0.69 | 0.538 | 0.37 | 0.044 | 0.77 | 0.152 | 0.544 | **11** |
|  |  | 0 | 0.262 | 0.574 | 0.552 | 0.8 | 0.108 | 0.208 | 0.632 | 0.31 | 0.718 | **12** |
|  | 0 | 0.552 | 0.31 | 0.222 | 0.6 | 0.448 | 0.66 | 0.334 | 0.48 | 0.242 | 0.234 | **13** |
| 0 | 0.714 | 0.666 | 0.804 | 0.492 | 0.114 | 0.266 | 0.574 | 0.804 | 0.234 | 0.956 | 0.548 | **14** |

**שלב שלישי**: המרחק המינימאלי הנוכחי בין קלאסטרים הוא 0.044 בין קלאסטרים (4,10) ו-11. נאחד –

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14 | 13 | 12 | 9 | 8 | 6 | 5, 7 | 4,10,11 | 3 | 2 | 1 |  |
|  |  |  |  |  |  |  |  |  |  | 0 | **1** |
|  |  |  |  |  |  |  |  |  | 0 | 0.408 | **2** |
|  |  |  |  |  |  |  |  | 0 | 0.722 | 0.314 | **3** |
|  |  |  |  |  | 0 | 0.708 | 0.582 | 0.432 | 0.69 | 0.282 | **6** |
|  |  |  |  |  | 0.734 | 0  ~~0.026~~ | 0.3  ~~0.326~~ | 0.74  ~~0.766~~ | 0.418  ~~0.444~~ | 0.826  ~~0.852~~ | **5, 7** |
|  |  |  |  | 0 | 0.352 | 0.66 | 0.734 | 0.12 | 0.842 | 0.434 | **8** |
|  |  |  | 0 | 0.578 | 0.226 | 0.482 | 0.356 | 0.658 | 0.464 | 0.456 | **9** |
|  |  |  | 0.312  ~~0.356~~ | 0.69  ~~0.734~~ | 0.538  ~~0.582~~ | 0.3 | 0 | 0.77  ~~0.814~~ | 0.108 | 0.5 | **4.10,11** |
|  |  | 0 | 0.574 | 0.552 | 0.8 | 0.108 | 0.208 | 0.632 | 0.31 | 0.718 | **12** |
|  | 0 | 0.552 | 0.222 | 0.6 | 0.448 | 0.66 | 0.31  ~~0.334~~ | 0.48 | 0.242 | 0.234 | **13** |
| 0 | 0.714 | 0.666 | 0.492 | 0.114 | 0.266 | 0.574 | 0.804 | 0.234 | 0.956 | 0.548 | **14** |

**שלב רביעי**: המרחק המינימאלי הבא בין קלאסטרים הוא 0.108 בין קלאסטרים (4,10,11) ו-2. נאחד –

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14 | 13 | 12 | 9 | 8 | 6 | 5, 7 | 3 | 2,4,10,11 | 1 |  |
|  |  |  |  |  |  |  |  |  | 0 | **1** |
|  |  |  |  |  |  |  | 0 | 0.722 | 0.314 | **3** |
|  |  |  |  |  | 0 | 0.708 | 0.432 | 0.582  ~~0.69~~ | 0.282 | **6** |
|  |  |  |  |  | 0.734 | 0 | 0.74 | 0.3 | 0.826 | **5, 7** |
|  |  |  |  | 0 | 0.352 | 0.66 | 0.12 | 0.734  ~~0.842~~ | 0.434 | **8** |
|  |  |  | 0 | 0.578 | 0.226 | 0.482 | 0.658 | 0.356  ~~0.464~~ | 0.456 | **9** |
|  |  |  | 0.312 | 0.69 | 0.538 | 0.3 | 0.77 | 0  ~~0.108~~ | 0.408  ~~0.5~~ | **2,4,10,11** |
|  |  | 0 | 0.574 | 0.552 | 0.8 | 0.108 | 0.632 | 0.31 | 0.718 | **12** |
|  | 0 | 0.552 | 0.222 | 0.6 | 0.448 | 0.66 | 0.48 | 0.242 | 0.234 | **13** |
| 0 | 0.714 | 0.666 | 0.492 | 0.114 | 0.266 | 0.574 | 0.234 | 0.956 | 0.548 | **14** |

**שלב חמישי**: המרחק המינימאלי הבא אף הוא 0.108 , הפעם בין קלאסטרים (5,7) ו-12. נאחד –

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14 | 13 | 9 | 8 | 6 | 5, 7,12 | 3 | 2,4,10,11 | 1 |  |
|  |  |  |  |  |  |  |  | 0 | **1** |
|  |  |  |  |  |  | 0 | 0.722 | 0.314 | **3** |
|  |  |  |  | 0 | 0.708 | 0.432 | 0.582 | 0.282 | **6** |
|  |  |  | 0 | 0.352 | 0.66 | 0.12 | 0.734  ~~0.842~~ | 0.434 | **8** |
|  |  | 0 | 0.578 | 0.226 | 0.482 | 0.658 | 0.356  ~~0.464~~ | 0.456 | **9** |
|  |  | 0.312 | 0.69 | 0.538 | 0.3 | 0.77 | 0  ~~0.108~~ | 0.408  ~~0.5~~ | **2, 4,10,11** |
|  |  | 0.574 | 0.552 | 0.734  ~~0.8~~ | 0  ~~0.108~~ | 0.632 | 0.3  ~~0.31~~ | 0.718 | **5,7,12** |
|  | 0 | 0.222 | 0.6 | 0.448 | 0.66 | 0.48 | 0.242 | 0.234 | **13** |
| 0 | 0.714 | 0.492 | 0.114 | 0.266 | 0.574 | 0.234 | 0.956 | 0.548 | **14** |

**שלב שישי**: המרחק המינימאלי הבא הוא 0.114 , בין קלאסטרים 8 ו-14. נאחד –

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 13 | 9 | 8, 14 | 6 | 5, 7, 12 | 3 | 2,4,10,11 | 1 |  |
|  |  |  |  |  |  |  | 0 | **1** |
|  |  |  |  |  | 0 | 0.722 | 0.314 | **3** |
|  |  |  | 0 | 0.708 | 0.432 | 0.582 | 0.282 | **6** |
|  | 0 | 0.578 | 0.226 | 0.482 | 0.658 | 0.356 | 0.456 | **9** |
|  | 0.312 | 0.69 | 0.538 | 0.3 | 0.77 | 0 | 0.408 | **2, 4,10,11** |
|  | 0.574 | 0.552 | 0.734 | 0 | 0.552 | 0.3 | 0.718 | **5,7,12** |
| 0 | 0.222 | 0.6 | 0.448 | 0.66 | 0.48 | 0.242 | 0.234 | **13** |
| 0.714 | 0.492 | 0  ~~0.114~~ | 0.266 | 0.574 | 0.12  ~~0.234~~ | 0.734  ~~0.956~~ | 0.434  ~~0.548~~ | **8, 14** |

**שלב שביעי**: המרחק המינימאלי הבא גם הוא 0.12 , בין קלאסטרים 3 ו-(8,14). נאחד –

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 13 | 9 | 6 | 5, 7, 12 | 3, 8 ,14 | 2,4,10,11 | 1 |  |
|  |  |  |  |  |  | 0 | **1** |
|  |  | 0 | 0.708 | 0.432 | 0.582 | 0.282 | **6** |
|  | 0 | 0.226 | 0.482 | 0.578 | 0.356 | 0.456 | **9** |
|  | 0.312 | 0.538 | 0.3 | 0.69 | 0 | 0.408 | **2, 4,10,11** |
|  | 0.574 | 0.734 | ~~0~~ | 0.552 | 0.3 | 0.718 | **5,7,12** |
| 0 | 0.222 | 0.448 | 0.66 | 0.48 | 0.242 | 0.234 | **13** |
| 0.714 | 0.492 | 0.266 | 0.574 | 0  ~~0.12~~ | 0.722  ~~0.734~~ | 0.314  ~~0.434~~ | **3, 8, 14** |

**שלב שמיני**: המרחק המינימאלי הבא הוא 0.208 , בין קלאסטרים (2,4,10,11) ו-(5,7,12). נאחד –

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 13 | 9 | 6 | 3, 8 ,14 | 2,4,5,7,10,11,12 | 1 |  |
|  |  |  |  |  | 0 | **1** |
|  |  | 0 | 0.432 | 0.582 | 0.282 | **6** |
|  | 0 | 0.226 | 0.578 | 0.356 | 0.456 | **9** |
|  | 0.312 | 0.538 | 0.552  ~~0.69~~ | 0 | 0.408 | 2,4,5,7,10,11,12 |
| 0 | 0.222 | 0.448 | 0.48 | 0.242 | 0.234 | **13** |
| 0.714 | 0.492 | 0.266 | 0 | 0.574 | 0.314 | **3, 8, 14** |

**שלב תשיעי**: המרחק המינימאלי הבא הוא 0.222 , בין קלאסטרים 9 ו-13. נאחד –

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 9,13 | 6 | 3, 8 ,14 | 2,4,5,7,10,11,12 | 1 |  |
|  |  |  |  | 0 | **1** |
|  | 0 | 0.432 | 0.582 | 0.282 | **6** |
| 0.312 | 0.538 | 0.552 | 0 | 0.408 | 2,4,5,7,10,11,12 |
| 0  ~~0.222~~ | 0.226  ~~0.448~~ | 0.48 | 0.242 | 0.234 | **9, 13** |
| 0.492 | 0.266 | 0 | 0.574 | 0.314 | **3, 8, 14** |

**שלב עשירי**: המרחק המינימאלי הבא הוא 0.226 , בין קלאסטרים 6 ו-(9,13). נאחד –

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 6,9,13 | 3, 8 ,14 | 2,4,5,7,10,11,12 | 1 |  |
|  |  |  | 0 | **1** |
| 0.312 | 0.552 | 0 | 0.408 | 2,4,5,7,10,11,12 |
| 0  ~~0.226~~ | 0.432  ~~0.48~~ | 0.242 | 0.234 | **6, 9, 13** |
| 0.266 | 0 | 0.574 | 0.314 | **3, 8, 14** |

**שלב 11**: המרחק המינימאלי הבא הוא 0.234 , בין קלאסטרים 1 ו-(6,9,13). נאחד –

|  |  |  |  |
| --- | --- | --- | --- |
| 3, 8 ,14 | 2,4,5,7,10,11,12 | 1, 6,9,13 |  |
| 0.552 | 0 | 0.312 | 2,4,5,7,10,11,12 |
| 0.432  ~~0.48~~ | 0.242 | 0 | **1, 6, 9, 13** |
| 0 | 0.574 | 0.266  ~~0.314~~ | **3, 8, 14** |

**שלב 12**: המרחק המינימאלי הבא הוא 0.242 , בין קלאסטרים(1,6,9,13) ו-(2,4,5,7,10,11,12). נאחד –

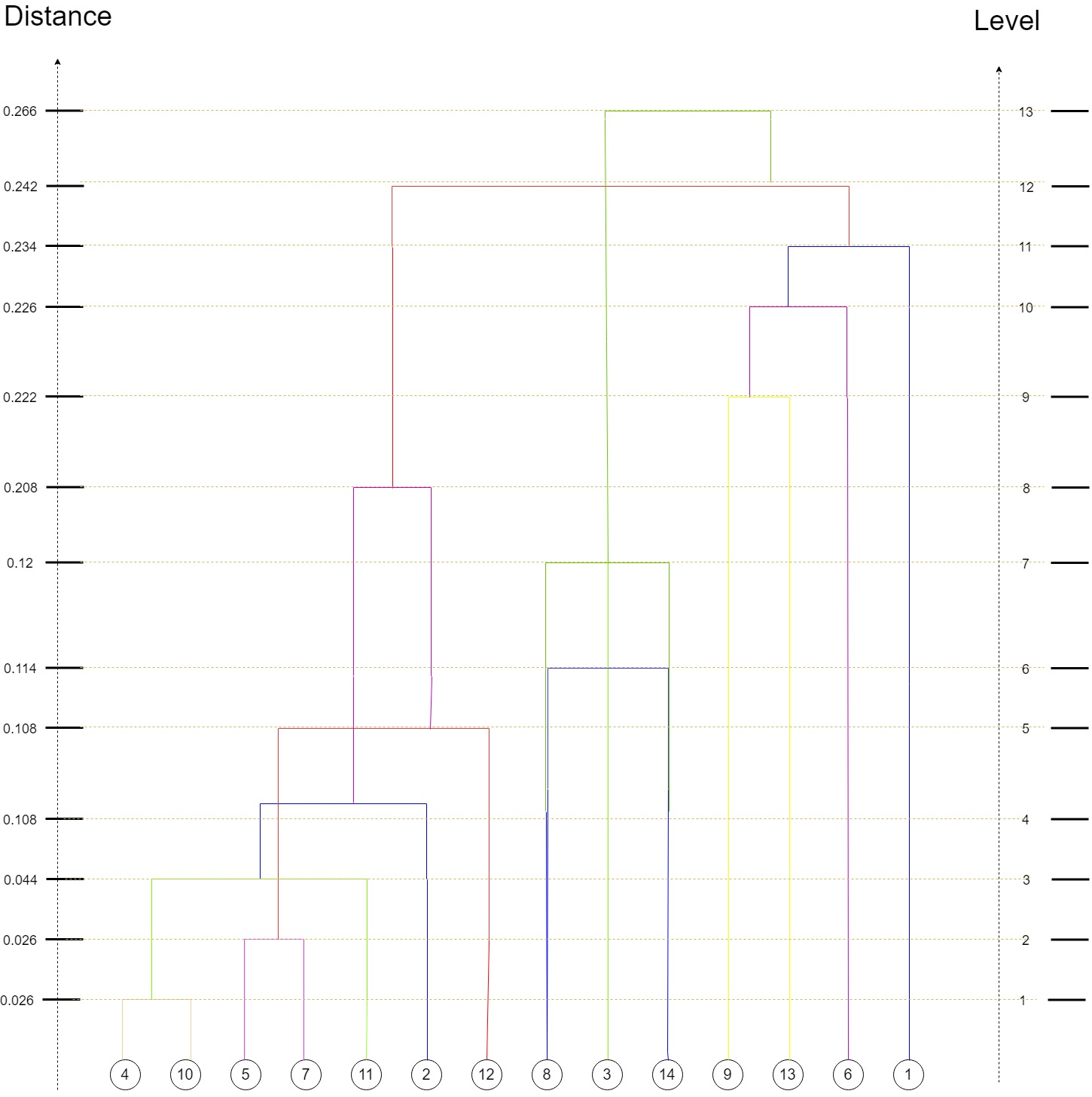
|  |  |  |
| --- | --- | --- |
| 3, 8 ,14 | 1,2,4,5,6,7,9,10,11,12,13 |  |
| 0.432  ~~0.552~~ | 0  0.~~242~~ | 1,2,4,5,6,7,9,10,11,12,13 |
| 0 | 0.266  ~~0.574~~ | **3, 8, 14** |

**שלב 13**: המרחק המינימאלי הבא והאחרון בין קלאסטרים הוא 0.266 , בין 2 הקלאסטרים האחרונים שנותרו. נאחדם ונקבל –

|  |  |
| --- | --- |
| 1,2,3,4,5,6,7,8,9,10,11,12,13,14 |  |
| 0 | 1,2,3,4,5,6,7,8,9,10,11,12,13,14 |

בניית הדנדוגרם

להלן הדנדוגרם עבור הבנייה שלעיל –



ניתן לנסות להשיג רזולוציה טובה יותר ב-PDF –



או לחלופין לטעון את הקובץ הבא לאתר ב-URL : <https://www.draw.io/>



**שאלה 3 – חוקי הקשר – אלגוריתם FP-growth**

מבוא - תיאור מסד הנתונים ואילוצי חוקי ההקשר

תהי נתונה טבלת הטרנסאקציות הבאה –

|  |  |  |
| --- | --- | --- |
| # | מספר הזמנה | פריטים |
| 1 | 1001 | {i1, i4, i5} |
| 2 | 1024 | {i1, i2, i3, i5} |
| 3 | 1012 | {i1, i2, i4, i5} |
| 4 | 1031 | {i1, i3, i4, i5} |
| 5 | 1015 | { i2, i3, i5} |
| 6 | 1022 | { i2, i4, i5} |
| 7 | 1029 | {i3, i4} |
| 8 | 1040 | {i1, i2, i3} |
| 9 | 1033 | {i1, i4, i5} |
| 10 | 1038 | {i1, i2, i5} |

כמו כן, נניח שקיימות הדרישות הבאות –

* Min\_Support = 60%
* Min\_Confidence = 80%

סעיף א – מציאת הקבוצות התדירות עם FP-growth

נמצא את כל הקבוצות התדירות ברשימת הטרנסאקציות לעיל, תוך שימוש באלגוריתם FP-growth –

**שלב ראשון - מניית התדירויות של כל פריט אינדיווידואלי**

נסרוק את מסד הנתונים בפעם ראשונה ונמנה את מספר השכיחויות של כל פריט בודד –

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | מספר פריט | מספר מופעים אבסולוטי | שכיחות יחסית | אחוז תמיכה |
| 1 | i1 | 7 | 7/10 | 70% |
| 2 | i2 | 6 | 6/10 | 60% |
| 3 | i3 | 5 | 5/10 | 50% |
| 4 | i4 | 6 | 6/10 | 60% |
| 5 | i5 | 8 | 8/10 | 80% |

**שלב שני – בדיקת תנאי סף שכיחות (תמיכה מינימאלית) ומיון הרשימה הנותרת**

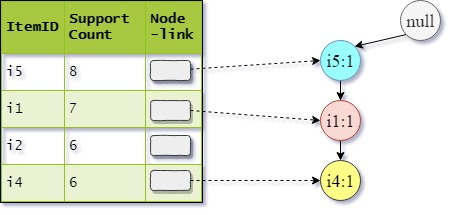
נבדוק האם כל הפריטים עומדים בדרישת ה-Min\_Support = 60%: פריט i3 אינו שכיח דיו (50% בלבד), ומתכונת האנטי-מונוטוניות של קבוצות שכיחות, נובע כי כל קבוצה שתכיל פריט זה גם היא לא תהיה שכיחה, לכן נוכל להתעלם מפריט זה במהלך חישוב הקבוצות השכיחות ולחסוך בביצועים. נמיין את הפריטים שנותרו בסדר יורד לפי השכיחות היחסית שלהם ומיון משני לפי סדר לקסיקוגרפי –

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | מספר פריט | מספר מופעים אבסולוטי | שכיחות יחסית | אחוז תמיכה |
| 1 | i5 | 8 | 8/10 | 80% |
| 2 | i1 | 7 | 7/10 | 70% |
| 3 | i2 | 6 | 6/10 | 60% |
| 4 | i4 | 6 | 6/10 | 60% |

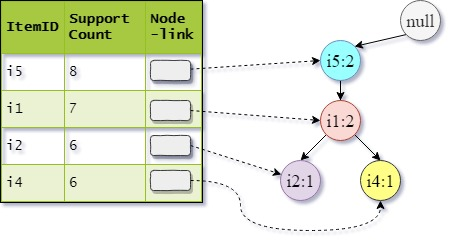
**שלב שלישי – בניית עץ השכיחויות של התבניות השכיחות**

ראשית נוסיף צומת שורש שישמש כזקיף (sentinel) עם התגית "null" (או NIL) ונתחיל מסד הנתונים בפעם השנייה (ואחרונה) טרנסאקציה אחר טרנסאקציה, כאשר בכל טרנסאקציה, נסרוק את הפריטים לפי סדר המיון לעיל, ונוסיף אותם לעץ כצמתים/נעדכן את המונה שלהם ע"פ הצורך.

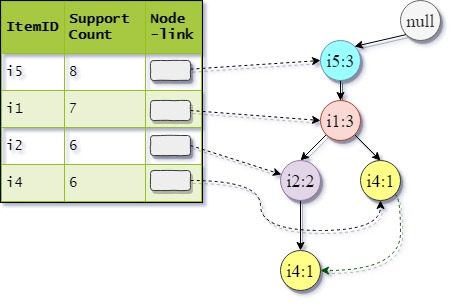
1. **טרנסאקציה 1001:** {i1, i4, i5}. נקבל את העץ הבא (המספר מימין לפריט בכל צומת מציין את מונה השכיחויות הנוכחי תחת ענף/מסלול זה, והטבלה מכילה את הקישור (הצבעה/הפנייה)) לצמתי העץ:



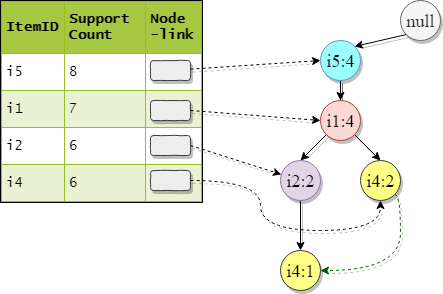
1. **טרנסאקציה 1024:** {i1, i2, ~~i3~~, i5}. נקבל את העץ הבא (אנחנו כאמור מתעלמים מ-i3) –



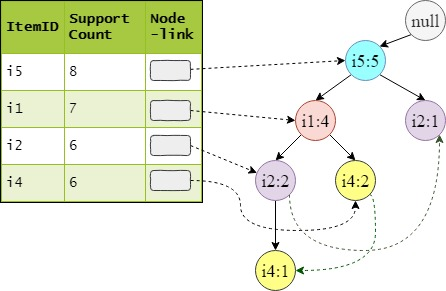
1. **טרנסאקציה 1012:** {i1, i2, i4, i5}. נקבל את העץ הבא –



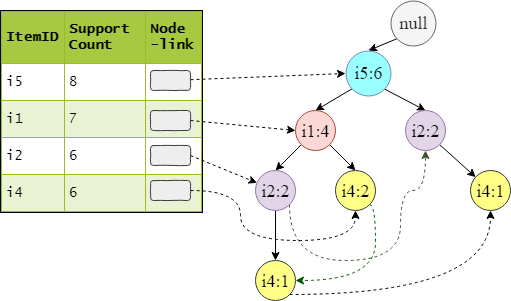
1. **טרנסאקציה 1031:** {i1, ~~i3~~, i4, i5}. נקבל את העץ הבא (עדכון מונים בלבד) –



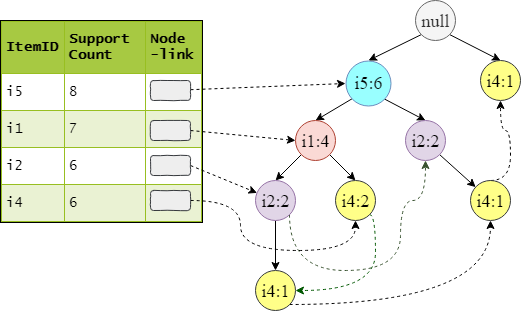
1. **טרנסאקציה 1015:** { i2, ~~i3~~, i5}. נקבל את העץ הבא (נוסף צומת חדש של i2 כבן ישיר של i5) –



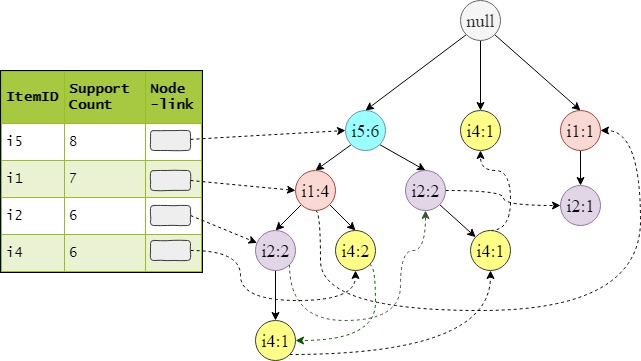
1. **טרנסאקציה 1022:** { i2, i4, i5}. נקבל את העץ הבא (ל- i2נוסף בן i4)–



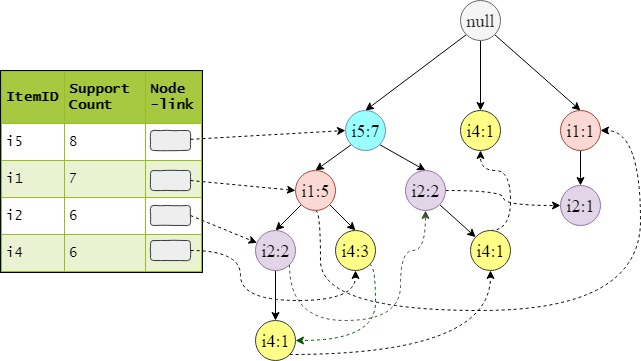
1. **טרנסאקציה 1029:** {~~i3~~, i4}. נקבל את העץ הבא (נוסף לשורש העץ בן ימני של i4) –



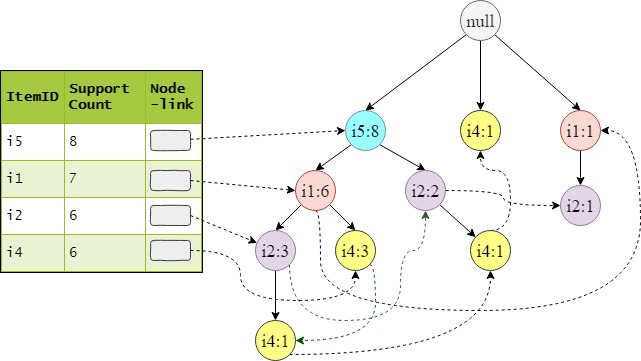
1. **טרנסאקציה 1040:** {i1, i2, ~~i3~~}. נקבל את העץ הבא (נוסף לשורש בן נוסף: צומת עבור פריט i1) –



1. **טרנסאקציה 1033:** {i1, i4, i5}. נקבל את העץ הבא (עדכון מונים בלבד) –



1. **טרנסאקציה 1038:** {i1, i2, i5}, עדכון מונים בלבד, ולבסוף נקבל את עץ ה-FP הסופי שלהלן –



ניתן גם לטעון את קובץ הקלט הבא לאתר <https://www.draw.io/> לקבלת העץ כאובייקט (ולא כתמונה):



**שלב רביעי – מציאת הקבוצות השכיחות מתוך העץ**

כעת משסיימנו את בניית עץ ה-FP נוכל לגזור/לשחזר ממנו את הקבוצות השכיחות (תדירות). התהליך ייעשה מלמטה למעלה[[2]](#footnote-2) ("Bottom-Up"): נתחיל בפריט הכי פחות שכיח, נתייחס לתת-גרף המושרה מעץ ה-FP על ידי כל המסלולים המובילים מהשורש לצמתים של פריט זה, ונחבר את ערכי המונים לכל תת-מסלול, תוך כדי מיזוג מונים של תתי-מסלולים זהים מענפים שונים בעץ. נחזור על תהליך לכל אחד מהפריטים. להלן התבניות המתקבלות מתהליך זה, כאפשר הפריטים מסודרים לפי סדר עולה של שכיחות (תמיכה), בשורת כל פריט מצוינים כל התבניות הנגזרות ממסלולים בעץ המסתיימים בפריט זה, ומצד לתבנית מספר המופעים הכולל שלה בעץ –

|  |  |
| --- | --- |
| פריט | תבניות שכיחות הנגזרות מהעץ |
| i4 | {i4: **6**},{i2, i4: **2**},{i1, i2, i4: **1**},{i5, i1, i2, i4: **1**},{i5, i2, i4:**2**},{i1, i4: **4**},{i5, i1, i4: **4**} |
| i2 | {i2: **6**},{i1, i2: **4**},{i5, i1, i2: **3**},{i5, i2: **5**} |
| i1 | {i1: **7**},{i5, i1: **6**} |
| i5 | {i5: **8**} |

**שלב חמישי – מציאת חוקי ההקשר החזקים מתוך הקבוצות השכיחות**

כעת נבדוק איזה מבין התבניות השכיחות שנמצאו רלוונטיות להפקת חוקי הקשר חזקים, דהיינו חוקי הקשר העומדים בתנאי הסף המינימאליים הן של התמיכה (60%), והן של הביטחון (80%). נתחיל בתמיכה (Min\_Support): 60% מ-10 (מספר הרשומות בבסיס הנתונים) הינו 6. לפיכך, רק תבניות שמופיעות 6 או יותר פעמים עשויות להיות רלוונטיות. מלבד התבניות הטריוויאליות בנות פריט בודד אותן מצאנו בסריקה הראשונה של ה-DB (ולכן מלכתחילה עומדות בתנאי סף של התמיכה המינימאלית), התבנית היחידה של 2 פריטים או יותר שעומדת בתנאי סף המינימאלי של התמיכה הינה התבנית {i5, i1} המופיעה 6 פעמים. נבדוק עמידה בממד הביטחון עבור 2 החוקים שניתן לייצר מתבנית זו –

* חוק ההקשר i5} 🡪 {i1}} בעל מידת ביטחון של 6/8 = 75% ולכן אינו עומד בדרישת הסף של 80%.
* חוק ההקשר i1} 🡪 {i5}} בעל מידת ביטחון של 6/7 = 85.7% ולכן עומד בדרישת הסף של 80%.

לסיכום, ההקשר החזק היחיד שנמצא מתוך מסד הנתונים לעיל תחת דרישות הסף שהוגדרו, דהיינו החוק היחיד שעומד הן בדרישות התמיכה והן בדרישות הביטחון הינו החוק שלהלן –



1. הבחירה במשתנה זה כבינארי סימטרי לעומת אי-סימטרי נראית טבעית, שכן שני המקרים נפוצים ומקובלים באותה המידה בסט הנתונים. [↑](#footnote-ref-1)
2. אמנם השגרה כרקורסיבית עצמה כפי שמובאת בספר הלימוד (עמוד 260 סעיף 2) מקבלת כפרמטר של הקריאה הראשונית את שורש העץ, כלומר השגרה מתחילה דווקא מלמעלה ("Top-down"), אולם כל עוד העץ אינו מכיל מסלול יחיד הרקורסיה צוללת עמוק יותר ומתחילה את העבודה בפועל רק כאשר מגיעים למסלול יחיד לעלה בקרקעית, לכן לצורכי סימולציה טבעי ונח יותר להתחיל מלמטה ולהתעלם ממחסנית הקריאות. [↑](#footnote-ref-2)