Miniproject in Privacy preserving Data Mining

Data Privacy Through Optimal k-Anonymization

Shir Ruso (I.D 209328467) Eden Nachum (I.D 313302838)

Advisor: Nadav Voloch

תוכן עניינים

הצגת הבעיה

עסקים וארגונים מחזיקים כיום נתונים אישיים יותר מאי פעם. הנתונים הללו משמשים כדי לשרת את הלקוחות בצורה טובה יותר, וכן לנהל פעילויות עסקיות באופן יעיל יותר. מנגד, ישנם גורמים זדוניים אשר מעוניינים לגשת לנתונים אישיים ולעקוב אחר מידע רגיש. מציאת דרך לשמור על הנתונים ועל התועלת שמופקת מהם תוך צמצום של סיכון לדליפת מידע רגיש הפכה למטרה מרכזית עבור מומחי אבטחה ברחבי העולם.

סיטואציה זו הובילה לעליית k-anonymity, מודל פרטיות שהוצע לראשונה לפני למעלה משני עשורים, ומאז התפתח והפך לטכניקה יעילה של הגנת פרטיות. [4]

אנונימיזציה של מאגר נתונים היא יצירת עותק של מאגר נתונים שממנו הושמטו או הוסרו פרטים מזהים, כדי לאפשר מסירה של מאגר נתונים זה לגורם חיצוני בלי לסכן את שמירת הסודיות של הנתונים שבמאגר המקורי. [1]

ל-K אנונימיות יש את המאפיין שכל רשומה אינה ניתנת להבדלה לפחות עם עוד k-1 רשומות אחרות. ככל שה-K גדול יותר, ההסתברות לזהות אדם באמצעות linking attack הולכת וקטנה. [2]

עבור אדם נתון, נתונים מזהים (שם, מיקוד, מין וכדומה) עשויים להופיע לצד נתונים רגישים (רישומי בריאות, מרשמים, מידע פיננסי, סיסמאות וכדומה). גורמים בעלי כוונות לא טהורות עשויים לשלב בין נתונים מזהים ובין נתונים רגישים כדי לזהות מחדש את אותו אדם ולסכן את פרטיותו. המטרה של k-anonymity היא להבטיח שלא יהיה ניתן לחבר בין שתי קטגוריות אלו.

מאגר הנתונים מכיל רשומות, כאשר כל רשומה מתייחסת לאדם ספציפי באוכלוסייה. להלן מאגר נתונים לפני שעבר תהליך אנונימיזציה:

Name	Age	Gender	State of domicile	Religion
Ramsha	30	Female	Tamil Nadu	Hindu
Yadu	24	Female	Kerala	Hindu
Salima	28	Female	Tamil Nadu	Muslim
Sunny	27	Male	Karnataka	Parsi
Joan	24	Female	Kerala	Christian
Bahuksana	23	Male	Karnataka	Buddhist
Rambha	19	Male	Kerala	Hindu
Kishor	29	Male	Karnataka	Hindu
Johnson	17	Male	Kerala	Christian
John	19	Male	Kerala	Christian

[3]

ישנן 2 טכניקות ידועות על מנת להשיג k-anonymity (עבור k נתון) והן:

1. הכללה- בשיטה זו, הערכים עבור תכונה מסוימת מוחלפים בערכים רחבים יותר. ההכללה מסירה מידע שניתן לזהות מנתונים שונים, ע"י הפחתת הספציפיות של תכונה. לדוגמא, במקום הערך 19 של התכונה "גיל", ניתן להחליף את הערך לאינטרוול $[0-20]$. כמו כן, את הערך 23 ניתן להחליף לאינטרוול $[20-30]$ וכדומה.
2. הדחקה – הסרת ערך של תכונה לחלוטין ממאגר הנתונים. יש צורך להשתמש בהדחקה עבור ערכים שאינם רלוונטיים למטרת איסוף הנתונים. לדוגמא, אם נאספים נתונים במטרה לקבוע באיזה גיל יש סיכוי לאנשים לפתח מחלה, דיכוי נתוני הגיל יהפוך את הנתונים לחסרי תועלת. מצד שני, דיכוי תעודת הזהות שלהם או שמם לא יפחית מתועלת המחקר.

דוגמא לשימוש ב- k-anonymity עבור מאגר הנתונים שהוצג לעיל:

Name	Age	Gender	State of domicile	Religion
*	20 < Age ≤ 30	Female	Tamil Nadu	*
*	20 < Age ≤ 30	Female	Tamil Nadu	*
*	20 < Age ≤ 30	Female	Kerala	*
*	20 < Age ≤ 30	Female	Kerala	*
*	20 < Age ≤ 30	Male	Karnataka	*
*	20 < Age ≤ 30	Male	Karnataka	*
*	20 < Age ≤ 30	Male	Karnataka	*
*	Age ≤ 20	Male	Kerala	*
*	Age ≤ 20	Male	Kerala	*
*	Age ≤ 20	Male	Kerala	*

[3]

נבחין כי גודל כל קבוצה הוא לפחות 2, ולכן מאגר נתונים זה עבר תהליך של k-anonymity עבור k=2.

עיקרי המאמר

מטרת החוקרים היא למצוא את האנונימיזציה האופטימלית ע"י מדדי עלות. החוקרים משתמשים בפונקציית עלות אשר מחשבת עבור כל אנונימיזציה את עלותה. המטרה היא לאתר את האנונימיזציה עם העלות הנמוכה ביותר. אנונימיזציה זו תייצג את האנונימיזציה האופטימלית עבור מאגר הנתונים עם k ספציפי.

החוקרים יצרו קבוצה של מספרים – Σ, כך שכל מספר מייצג ערך קיים בטבלה עבור תכונה מסוימת. לדוגמא הקבוצה $\{1,2,3,4,5,6,7,8,9\}$ מייצגת את התרשים הבא:

AGE GENDER MARITAL STATUS
<[10-29][30-39][40-49]> <[M][F]> <[Married][Widowed][Divorced][Never Married]>
.....**1***.....**2**.....**3**.....**4***.**5**.....**6***.....**7**.....**8**.....**9**.....

[2]

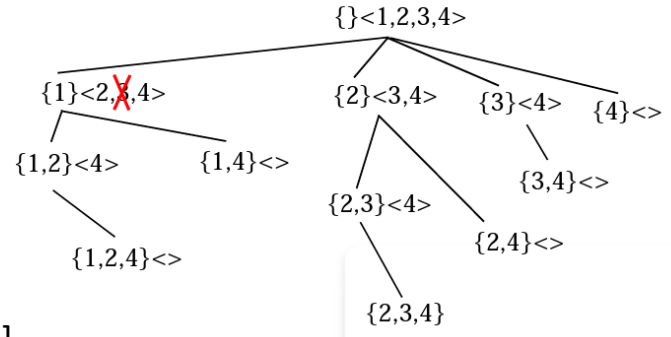
כל האנונימיזציות האפשריות הן למעשה קבוצת החזקה של קבוצת מספרים זו.

בצורה נאיבית, ניתן לעבור סדרתית על קבוצת החזקה ולהשוות עלות של כל קבוצת חזקה מול העלות האופטימלית עד כה ,עד שתמצא הקבוצה עם העלות הנמוכה ביותר. קבוצה זו תייצג את האנונימיזציה האופטימלית.

מאחר וגודל קבוצת החזקה הוא $2^{|\Sigma|}$, במקרים בהם גודלה של Σ מאוד גדול, זמן הריצה של הקלט יהיה מאוד איטי. לכן, החוקרים בחרו בשיטה שנקראת "גיזום". החוקרים פרסו את קבוצת החזקה על גבי עץ. כל קודקוד בעץ מורכב מ-head ו-tail, כאשר head מייצג אנונימיזציה ספציפית וה-tail מכיל ערכים אופציונליים שיכולים להתווסף לאנונימיזציה זו. מטרת שיטת הגיזום היא למחוק ערכים אשר נמצאים ב-tail שהוספתם ל-head לא תוביל לאנונימיזציה בעלת עלות נמוכה יותר מזו שנמצאה עד כה.

בתמונה ניתן לראות כיצד פרוסות כל האנונימיזציות מעל הקבוצה $\{1,2,3,4\}$ על גבי עץ. כמו כן, ניתן לראות את השפעת הגיזום של הערך 3 על גודל העץ.

Figure 2. Set enumeration tree over alphabet {1,2,3,4}.



[2]

נבחין כי כל אנונימיזציה מחלקת את ה-data set לקבוצות שונות אשר כל הרשומות השייכות לקבוצה מסוימת זהות. המטרה היא למצוא אנונימיזציה אשר תחלק את ה-data set לקבוצות זרות אשר גודל כל קבוצה הוא לפחות k. במאמר, החוקרים מקבילים את הקבוצות הללו למחלקות שקילות המושרות ע"י אותה אנונימיזציה.
ייתכן וקיימות מספר אנונימיזציות אשר מקיימות מטרה זו , על כן החוקרים השתמשו במדדי עלות כדי לאתר את האנונימיזציה עם העלות הנמוכה ביותר שלטענתם היא האופטימלית.

תיאור האלגוריתם

הפונקציה הראשית של האלגוריתם היא $\boldsymbol{K-OPTIMIZE}$. תחילה, אנו קוראים לפונקציה זו עם הערכים $K-OPTIMIZE\ (k,\phi,\Sigma_{all},\infty)$.
האלגוריתם בנוי באופן שבו האנונימיזציה הראשונה שנבחנת היא האנונימיזציה הכי כללית , כלומר ה- head מאותחל לקבוצה הריקה וה- tail מכיל את כל שאר הערכים. במהלך ריצת האלגוריתם , נבחנות אנונימיזציות ספציפיות יותר, כלומר אנונימיזציות שבהן הועברו ערכים מה- tail ל- head.
כמו כן, העלות הראשונית שנשלחת לפונקציה היא העלות הגבוה שניתן להשיג (∞) אשר תעודכן במהלך ריצת האלגוריתם.
בסיום האלגוריתם, הערך שחוזר הינו העלות הטובה ביותר שניתן להשיג מבין כל האנונימיזציות האפשריות.
האנונימיזציה האופטימלית הינה האנונימיזציה שעלותה היא למעשה העלות שחזרה מהפונקציה $K-OPTIMIZE$, ו-$best\_anonymization$ יכיל את האנונימיזציה הנ"ל.

$best\_anonymization \leftarrow \phi$; ; $this\ feild\ holds\ the\ current\ best\ anonymization$

$\boldsymbol{K-OPTIMIZE(k, head\ set\ H, tail\ set\ T, best\ cost\ c)}$
; ; $This\ function\ returns\ the\ lowest\ cost\ of\ any\ anonymization\ within\ the\ sub-tree$
; ; $rooted\ at\ H\ that\ has\ a\ cost\ less\ than\ c$.
; ; $Otherwise.\ it\ returns\ c.$
1) $T \leftarrow \boldsymbol{PRUNE-USELESS-VALUES}(H,T)$
2) $c_{optional} \leftarrow \boldsymbol{COMPUTE-COST}(H)$
3) $\boldsymbol{if}\ (c_{optional} < c)$
4) $\quad \boldsymbol{then}\ G \leftarrow H,\ \ c \leftarrow c_{optional}$
5) $T \leftarrow \boldsymbol{PRUNE}\ (H,T,c)$
6) $T \leftarrow \boldsymbol{REORDER-TAIL}\ (H,T)$
7) $\boldsymbol{while}\ T\ is\ non-empty\ \boldsymbol{do}$
8) $\quad v \leftarrow the\ first\ value\ in\ the\ ordered\ set\ T$
9) $\quad H_{new} \leftarrow H \cup \{v\}$
10) $\quad T \leftarrow T - \{v\}$
11) $\quad c \leftarrow K-OPTIMIZE(k, H_{new}, T, c)$
12) $\quad T \leftarrow PRUNE(H,T,c)$
13) $\boldsymbol{return}\ c$

<u>תיאור האלגוריתם לפי שלבים:</u>

<u>שלב 1:</u> $PRUNE - USELESS - VALUES(H,T)$ - פונקציה זו גוזמת ערכים מ- T שמייצגים ספציאליזציות . גזימה זו נעשית מתוך הבנה כי ערכים אלה אינם יכולים לשפר את עלות האנונימיזציה. נבחין, כי הוספת ערך מ-T לאנונימיזציה H יכולה לפצל מחלקת שקילות (שהושרתה ע"י H) ל-2 מחלקות שקילות או לא להשפיע כלל. מטרת הפונקציה היא לסנן ערכים שהוספתם לאנונימיזציה H , יוצרת לפחות מחלקת שקילות אחת שגודלה קטן מ-K. מאחר ומטרת האלגוריתם היא למצוא אנונימיזציה שמשרה מחלקות שקילות בגודל לפחות k, ערכים אלו "חסרי תועלת" ולכן יגזמו.

$E - set\ of\ all\ the\ equivalence\ classes$
$\boldsymbol{PRUNE - USELESS - VALUES(H,T)}$
  $\boldsymbol{for\ each}\ v\ in\ T\ \boldsymbol{do}$
    $E_{new} \leftarrow\ UPDATE - EQUIVALENCE - CLASSES(H,v,E)$
    $\boldsymbol{if}\ at - least\ one\ e\ in\ E_{new}\ is\ less\ than\ size\ k$
      $T \leftarrow T -\ \{v\}$
  $\boldsymbol{return}\ T$

$\boldsymbol{UPDATE - EQUIVALENCE - CLASSES(H,v,E)}$
$;;This\ function\ return\ new\ set\ of\ equivalences\ classes\ considering\ adding\ value\ v\ to\ H.$

<u>שלב 2:</u> $COMPUTE - COST(H)$ - חישוב עלות האנונימיזציה הנוכחית.
הפונקציה מתחשבת בגודל מחלקות השקילות שהאנונימיזציה משרה וכן בגודל ה-data set.

$\boldsymbol{COMPUTE - COST(H)}$
$;;This\ function\ computes\ the\ following\ formula:$
$;;C_{DM}(H) = \sum_{\forall E\ s.t\ |E|\geq k} |E|^2 + \sum_{\forall E\ s.t\ |E|<k} |D||E|$
$;;In\ \ this\ \ expression, the\ \ sets\ E\ \ refer\ to\ \ the\ equivalence\ classes$
$;;\ \ of\ \ tuples\ \ in\ \ D\ induced\ \ by\ \ the\ anonymization\ H.$
$Sum \leftarrow 0$
$\boldsymbol{for\ each}\ e\ in\ E\ \boldsymbol{do}$
  $\boldsymbol{if}\ |e| \geq k$
    $Sum \leftarrow Sum + |e|^2$
  $\boldsymbol{else}$
    $Sum \leftarrow Sum + |e| \cdot |D|$
$\boldsymbol{return}\ Sum$

<u>שלבים 3-4:</u> בדיקה האם העלות שחזרה משלב קודם נמוכה מהעלות הנוכחית. אם כן, סימן שנמצאה אנונימיזציה טובה יותר ולכן נעדכן את העלות האופטימלית הנוכחית להיות העלות שחזרה משלב 2. כמו כן, נעדכן את האנונימיזציה האופטימלית הנוכחית להיות האנונימיזציה שנבדקת בשלב זה באלגוריתם.

שלב 5: $\boldsymbol{PRUNE}\ (H,T,c)$ – האלגוריתם תחילה מנסה לגזום את הקודקוד הנוכחי <H,T> . במידה והאלגוריתם נכשל לגזום את הקודקוד , האלגוריתם מנסה לגזום ערכים מה-tail של הקודקוד.
גזימת הקודקוד הנוכחי היא למעשה גזימת כל הקודקודים מה-tail. המשמעות היא שלא משנה איזה ערך נוסיף ל-H , לא נצליח להשיג אנונימיזציה טובה יותר מאשר האנונימיזציה שמכילה את ערכי H בלבד. לכן, נחסוך קריאות רקורסיביות לילדיו של קודקוד זה.
אם האלגוריתם נכשל בגזימת הקודקוד, משמע הוספת ערכים מה-tail עשויה להביא לאנונימיזציה טובה יותר. במצב זה האלגוריתם עובר על ערכי ה-tail וגוזם את הערכים אשר הוספתם ל-head לא יכולה להוביל לאנונימיזציה בעלת עלות נמוכה יותר מזו שנמצאה עד כה.

שיטת הגיזום מסתמכת על הפונקציה $COMPUTE-LOWER-BOUND(k,H,A)$.

בפונקציה PRUNE בודקים האם הערך שחזר מ $COMPUTE-LOWER-BOUND(k,H,A)$ עבור ה- head וה-all-set יותר גדול מהעלות האופטימלית שנמצאה עד כה. במידה והערך גדול יותר, אזי הקודקוד הנ"ל אינו יכול להוביל לאנונימיזציה טובה יותר.
לפיכך, מתבצעת גזימה בהתאם למקרים שצוינו לעיל.

$\boldsymbol{COMPUTE-LOWER-BOUND(k,H,A)}$
$;;A\ is\ the\ union\ of\ the\ head\ and\ tail\ sets\ of\ a\ given\ node.$
$;;E_{A,t}\ is\ the\ equivalence\ class\ induced\ by\ the\ allset\ A\ that\ contains\ t.$
$sum \leftarrow 0$
$\boldsymbol{for\ each}\ t\ in\ D\ \boldsymbol{do}$
$\boldsymbol{if}\ t\ suppressed\ by\ H$
$\boldsymbol{then}\ sum \leftarrow sum + |D|$
$\boldsymbol{else}$
$sum \leftarrow sum + \max(|E_{A,t}|, k)$
$\boldsymbol{return}\ sum$

$\boldsymbol{PRUNE(k, head\ set\ H, tail\ set\ T, best\ cost\ c)}$
$;;this\ function\ creates\ and\ returns\ a\ new\ tail\ set\ by\ removing\ values\ from\ T\ that\ can\ not$
$;;lead\ to\ anonymization\ with\ cost\ lower\ than\ c.$
$T_{new} \leftarrow \phi$
$T_{copy} \leftarrow T$
$\boldsymbol{if}(\boldsymbol{COMPUTE-LOWER-BOUND}(k,H,H \cup T) \geq c)$
$\boldsymbol{return}\ T_{new}$
$\boldsymbol{for-each}\ v\ in\ T\ \boldsymbol{do}$:
$T_{copy} \leftarrow \mathrm{T_{copy}} - \{\mathrm{v}\}$
$H_{new} \leftarrow H \cup \{v\}$
$\boldsymbol{if}\left(\boldsymbol{COMPUTE-LOWER-BOUND}\left(k, H_{new}, H_{new} \cup T_{copy}\right) < c\right)$
$T_{new} \leftarrow T_{new} \cup \{v\}$
$\boldsymbol{return}\ T_{new}$

שלב 6: $\boldsymbol{REORDER-TAIL}\ (H,T)$ – האלגוריתם מסדר מחדש את ערכי ה-tail באופן שיכול להגדיל את אפשרויות הגזימה.

$\boldsymbol{REORDER-TAIL(H,T)}$
$T_{new} \leftarrow \phi$
$S \leftarrow \phi$
$\boldsymbol{for\ each}\ v\ in\ T\ \boldsymbol{do}$
$counter \leftarrow COUNT-SPLITING(H,v)$
$S \leftarrow S \cup \{(v, counter)\}$
$S \leftarrow sort\ S\ by\ the\ counters\ in\ each\ pair$
$\boldsymbol{for\ each}\ (v, counter) in\ S\ \boldsymbol{do}$
$T_{new} \leftarrow T_{new} \cup \{v\}$ $;;preserve\ ordering$
$\boldsymbol{retrun}\ T_{new}$

$\boldsymbol{COUNT-SPLITING(H,v)}$
$;;the\ function\ counts\ the\ number\ of\ equivalence\ classes\ induced\ by\ H\ that\ are\ split\ by$
$;;specializing\ on\ v.$

שלבים 7-11: האלגוריתם עובר בצורת DFS על העץ המושרש של הקודקוד הנוכחי. האופן שבו סריקת DFS מתבצעת היא ע"י קריאות רקורסיביות ל- $K-OPTIMIZE$ על כל אחד מילדיו הישרים של אותו קודקוד.

שלב 12: מתבצעת קריאה נוספת ל-PRUNE זאת מאחר וייתכן וה- best cost השתנה לאור קריאה לפונקציה $K-OPTIMIZE$ עם אחד הילדים המושרשים של הקודקוד. ייתכן ונוכל לגזום ערכים נוספים לאור שינוי זה שקודם לכן לא נגזמו, ובכך לחסוך קריאות רקורסיביות עבור העצים המושרשים בילדיו הנוספים של אותו קודקוד.

שלב 13: האלגוריתם מחזיר את ה-best cost שמתאימה לאנונימיזציה האופטימלית השמורה במשתנה הגלובלי best anonymization.

תיאור מאגר המידע

השתמשנו במאגר המידע $Stroke\ Prediction\ Dataset$ מאתר $Kaggle$.
מטרת מאגר המידע היא לחזות האם מטופל עשוי לקבל שבץ לפי נתונים כמו מגדר, גיל, מחלות שונות וכדומה. הטבלה מכילה 5110 רשומות וכן 12 תכונות.

Attribute Information:

1) id: unique identifier
2) gender: "Male", "Female" or "Other"
3) age: age of the patient
4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5) heart disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6) ever married: "No" or "Yes"
7) work type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8) Residence type: "Rural" or "Urban"
9) avg glucose level: average glucose level in blood
10) bmi: body mass index
11) smoking status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
12) stroke: 1 if the patient had a stroke or 0 if not
*Note: "Unknown" in smoking status means that the information is unavailable for this patient

קישור למאגר מידע :
https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

תיאור המימוש

בחרנו לממש את הפרויקט בשפת JAVA.

תחזקנו מערך המכיל את כל הערכים האפשריים של כל תכונה. תכונות מספריות המרנו לטווחים. כמו כן, כל ערך הקיים במערך ייוצג ע"י מספר ייחודי. לדוגמא, עבור התכונה "מין", הערכים "Male"ו-"Female" יוצגו כמספרים 0 ו-1 בהתאמה.
באופן דומה, עבור התכונה "גיל", הטווחים $[0-27], [28-55], [56-82]$ יוצגו כמספרים 2-4 בהתאמה.

<u>יצרנו 5 מחלקות:</u>

- **מחלקת $\boldsymbol{Patient}$:** המחלקה מחזיקה את כל המידע הדרוש עבור כל רשומה ב- data set. לכל מטופל יהיו שדות התואמים לתכונות במאגר מידע.
- **מחלקת $\boldsymbol{Tuple}$:** המחלקה מייצגת רשומה בטבלה שהוכללה. כל רשומה מאותחלת כרשומה הכללית ביותר, כך שבמהלך ריצת האלגוריתם, הרשומות עוברות תהליך ספציאליזציה. כל רשומה מאותחלת להיות
$[" * ", "A", "B", "C", "D", "E", "F", "G", "H", "I", "J", "K"]$.
"*" מייצג את התכונה "ID" ולא ישתנה לאורך כל ריצת האלגוריתם.
"A" מייצג את התכונה "Gender", "B" מייצג את התכונה "Age" וכן הלאה.
במהלך הריצה הרשומה עשויה להשתנות . דוגמא אפשרית היא:
$[" * ", 1, "B", 5, "D", "E", "F", "G", "H", "I", 24, "K"]$.
בדוגמא זו, האנונימיזציה הנוכחית מכילה את הערכים 1,5 ו-24 (ייתכן שמכילה ערכים נוספים). כאשר המספר 1 מייצג את הערך "Female", המספר 5 מייצג את הערך "Has Hypertension", וכן המספר 24 מייצג את הערך "Formerly Smoked".
- **מחלקת $\boldsymbol{Equivalence\ Class}$** : המחלקה מייצגת את המחלקות שקילות בתוכנית. כל מחלקת שקילות מכילה רשימה של מטופלים אשר חולקים Tuple זהה (tuple שעבר ספציאליזציה לאור אנונימיזציה כלשהי). המחלקה מכילה פונקציה עיקרית בתוכנית "induceEC". הפונקציה מקבלת ערך שנוסף לאנונימיזציה שהמחלקה מושרית על ידה, ומשרה את המחלקה ע"י האנונימיזציה בתוספת הערך החדש. הפונקציה מחזירה 2 מחלקות שקילות מאחר והוספת הערך יכול לגרום לפיצול מחלקת השקילות לכל היותר ל-2 מחלקות. במידה והערך לא השפיע על מחלקת השקילות, נחזיר 2 מחלקות שקילות שאחת מהן זהה למחלקה המקורית והשנייה ריקה.
- **מחלקת $\boldsymbol{Head}$** : המחלקה מחזיקה אנונימיזציה כלשהי ורשימה של מחלקות שקילות אשר מושרות על ידי אותה אנונימיזציה.

תוצאות הרצה על מאגר המידע

נציג דוגמת הרצה עבור k=10 על 200 רשומות מהטבלה.
האנונימיזציה שהתקבלה היא:

age: [0,27]
work type: children
work type: Never_worked
bmi: [70,98]
stroke: 1
stroke: 0
residence type: Rural
residence type: Urban
avg glucose level: [55,127]
smoking status: formerly smoked
smoking status: never smoked

האנונימיזציה ממחישה אילו ערכים יופיעו ברשומות במאגר הנתונים לאחר הפעלת האלגוריתם על מאגר הנתונים.
לדוגמא עבור מטופל שגילו נע בין 0 ל- 27 , ברשומה המייצגת אותו במאגר הנתונים לאחר הפעלת האלגוריתם יופיע הערך [0,27] תחת התכונה גיל. מצד שני, עבור מטופל שגילו גדול מ-27, ברשומה המייצגת אותו במאגר הנתונים לאחר הפעלת האלגוריתם יופיע הערך B תחת התכונה גיל. (כאשר B מייצג את כל שאר הערכים שהם שני הטווחים [28-55] ו- [56-82]).

האנונימיזציה הנ"ל מחלקת את מאגר המידע למחלקות שקילות בגודל 10 לפחות.
לכל מחלקה קיים tuple משותף לכל הרשומות הנמצאות במחלקת השקילות. אותו tuple ישוכפל במאגר הנתונים הסופי כמספר הרשומות הנמצאות באותה מחלקה.

להלן מחלקות השקילות הנוצרות בסוף ריצת האלגוריתם ע"י האנונימיזציה שנבחרה:

id	gender	age	hypertension	heart disease	ever married	work type	residence type	avg glucose level	bmi	smoking status	stroke
*	A	B	C	D	E	F	Urban	H	I	J	1
60182	Female	49	0	0	1	Private	Urban	171.23	34.4	smokes	1
56112	Male	64	0	1	1	Private	Urban	191.61	37.5	smokes	1
34120	Male	75	1	0	1	Private	Urban	221.29	25.8	smokes	1
54827	Male	69	0	1	1	Self-employed	Urban	195.23	28.3	smokes	1
43717	Male	57	1	0	1	Private	Urban	212.08	44.2	smokes	1
61960	Male	82	0	1	1	Private	Urban	144.9	26.4	smokes	1
7937	Male	60	1	0	1	Govt_job	Urban	213.03	20.2	smokes	1
42117	Male	43	0	0	1	Self-employed	Urban	143.43	45.9	Unknown	1
11762	Female	76	0	0	1	Private	Urban	207.28	34.9	Unknown	1
59437	Female	57	0	0	1	Private	Urban	221.89	37.3	smokes	1
65105	Male	81	0	0	1	Private	Urban	213.22	26.1	Unknown	1
68025	Female	79	0	1	0	Private	Urban	205.33	31	smokes	1
29552	Female	55	1	1	1	Private	Urban	210.4	40	smokes	1
20463	Male	81	1	1	1	Private	Urban	250.89	28.1	smokes	1
39186	Female	57	0	1	1	Private	Urban	216.58	31	Unknown	1
25974	Male	78	0	0	1	Self-employed	Urban	218.46	26.8	Unknown	1

id	gender	age	hypertension	heart disease	ever married	work type	residence type	avg glucose level	bmi	smoking status	stroke
*	A	B	C	D	E	F	Urban	H	I	never smoked	1
5317	Female	79	0	1	1	Private	Urban	214.09	28.2	never smoked	1
13861	Female	52	1	0	1	Self-employed	Urban	233.29	48.9	never smoked	1
68794	Female	79	0	0	1	Self-employed	Urban	228.7	26.6	never smoked	1
39373	Female	82	1	0	1	Self-employed	Urban	196.92	22.2	never smoked	1
58631	Male	73	1	0	1	Self-employed	Urban	194.99	32.8	never smoked	1
5111	Female	54	1	0	1	Govt_job	Urban	180.93	27.7	never smoked	1
17004	Female	70	0	0	1	Private	Urban	221.58	47.5	never smoked	1
36236	Male	80	1	0	1	Private	Urban	240.09	27	never smoked	1
28291	Female	79	0	1	1	Private	Urban	226.98	29.8	never smoked	1
50522	Female	72	0	0	1	Govt_job	Urban	131.41	28.4	never smoked	1
16817	Female	78	1	0	0	Private	Urban	130.54	20.1	never smoked	1
71279	Female	71	0	0	1	Govt_job	Urban	263.32	38.7	never smoked	1
17308	Female	72	1	0	1	Private	Urban	221.79	30	never smoked	1
20426	Female	78	1	0	0	Private	Urban	203.87	45.7	never smoked	1
31179	Male	63	0	0	1	Private	Urban	208.65	30.7	never smoked	1

id	gender	age	hypertension	heart disease	ever married	work type	residence type	avg glucose level	bmi	smoking status	stroke
*	A	B	C	D	E	F	Urban	H	I	formerly smoked	1
9046	Male	67	0	1	1	Private	Urban	228.69	36.6	formerly smoked	1
56669	Male	81	0	0	1	Private	Urban	186.21	29	formerly smoked	1
54401	Male	80	0	1	1	Self-employed	Urban	252.72	30.5	formerly smoked	1
10710	Female	56	0	0	1	Private	Urban	185.17	40.4	formerly smoked	1
37132	Male	82	0	0	1	Govt_job	Urban	200.59	29	formerly smoked	1
55824	Male	76	0	0	1	Private	Urban	140.1	29.9	formerly smoked	1
67981	Male	66	0	0	1	Private	Urban	151.16	27.5	formerly smoked	1
46703	Male	68	0	1	1	Private	Urban	223.83	31.9	formerly smoked	1
66258	Female	71	0	0	1	Self-employed	Urban	195.71	34.1	formerly smoked	1
42899	Male	78	0	0	1	Self-employed	Urban	133.19	23.6	formerly smoked	1
67895	Female	82	1	1	1	Govt_job	Urban	215.94	27.9	formerly smoked	1
24905	Female	65	0	0	1	Private	Urban	205.77	46	formerly smoked	1
64373	Male	59	0	0	1	Private	Urban	200.62	35.8	formerly smoked	1
68627	Male	80	1	1	1	Private	Urban	175.29	31.5	formerly smoked	1
23368	Female	77	1	0	1	Self-employed	Urban	199.84	28	formerly smoked	1

id	gender	age	hypertension	heart disease	ever married	work type	residence type	avg glucose level	bmi	smoking status	stroke
*	A	B	C	D	E	F	Urban	[55 , 127]	I	J	1
60491	Female	78	0	0	1	Private	Urban	58.57	24.2	Unknown	1
12175	Female	54	0	0	1	Private	Urban	104.51	27.3	smokes	1
72366	Male	76	0	0	1	Private	Urban	104.47	20.3	Unknown	1
49130	Male	74	0	0	1	Private	Urban	98.55	25.6	Unknown	1
37726	Female	80	1	0	1	Self-employed	Urban	68.56	26.2	Unknown	1
12363	Male	64	0	1	1	Govt_job	Urban	74.1	28.8	Unknown	1
33175	Female	57	0	0	1	Govt_job	Urban	110.52	28.5	Unknown	1
48405	Male	80	0	1	1	Private	Urban	68.53	24.2	smokes	1
71639	Female	68	0	0	0	Govt_job	Urban	82.1	27.1	Unknown	1
7547	Male	74	0	0	1	Private	Urban	72.96	31.3	smokes	1
72918	Female	53	1	0	1	Private	Urban	62.55	30.3	Unknown	1
70943	Female	80	0	0	1	Private	Urban	73.54	24	Unknown	1
12482	Male	68	0	0	1	Self-employed	Urban	77.82	27.5	smokes	1
67432	Female	60	0	0	1	Private	Urban	97.43	26.4	smokes	1
28378	Male	61	1	1	1	Private	Urban	112.24	37.4	smokes	1
54724	Female	81	0	0	0	Govt_job	Urban	70.3	25.8	smokes	1
31154	Female	39	0	0	1	Self-employed	Urban	97.76	29.6	smokes	1
12917	Female	79	0	0	1	Private	Urban	97.73	21.5	smokes	1
28493	Male	57	0	0	1	Private	Urban	86.3	31.7	Unknown	1

id	gender	age	hypertension	heart disease	ever married	work type	residence type	avg glucose level	bmi	smoking status	stroke
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
10434	Female	69	0	0	0	Private	Urban	94.39	22.8	never smoked	1
27458	Female	60	0	0	0	Private	Urban	89.22	37.8	never smoked	1
14248	Male	48	0	0	0	Govt_job	Urban	84.2	29.7	never smoked	1
62602	Female	49	0	0	1	Private	Urban	60.91	29.9	never smoked	1
1261	Male	54	0	0	1	Private	Urban	71.22	28.5	never smoked	1
35626	Male	81	0	0	1	Self-employed	Urban	99.33	33.7	never smoked	1
47167	Female	77	1	0	1	Self-employed	Urban	124.13	31.4	never smoked	1
17013	Male	78	1	0	0	Private	Urban	113.01	24	never smoked	1
66638	Female	68	1	0	0	Self-employed	Urban	79.79	29.7	never smoked	1
51169	Male	81	0	0	1	Private	Urban	72.81	26.3	never smoked	1
66315	Female	57	0	0	0	Self-employed	Urban	68.02	37.5	never smoked	1
4639	Female	69	0	0	1	Govt_job	Urban	82.81	28	never smoked	1
59125	Female	53	0	0	1	Govt_job	Urban	64.17	41.5	never smoked	1
5563	Female	77	0	0	1	Private	Urban	105.22	31	never smoked	1
8045	Female	74	1	0	1	Private	Urban	70.28	21.8	never smoked	1
37651	Female	69	1	1	0	Self-employed	Urban	72.17	36.8	never smoked	1
62861	Female	78	0	0	1	Private	Urban	67.29	24.6	never smoked	1
32503	Female	80	0	0	1	Self-employed	Urban	76.57	34.1	never smoked	1
62466	Female	80	0	0	1	Private	Urban	64.44	45	never smoked	1
54567	Female	46	0	0	1	Private	Urban	78.18	30.8	never smoked	1
491	Female	74	0	0	1	Self-employed	Urban	74.96	26.6	never smoked	1
35578	Male	78	0	0	0	Self-employed	Urban	90.19	26.9	never smoked	1
33943	Female	39	0	0	1	Private	Urban	83.24	26.3	never smoked	1
66866	Female	48	0	0	1	Private	Urban	74.11	20.5	never smoked	1
2548	Female	81	0	0	1	Self-employed	Urban	95.84	21.5	never smoked	1
2390	Male	78	0	0	1	Self-employed	Urban	116.1	27.1	never smoked	1
69959	Female	80	1	0	0	Private	Urban	66.03	35.4	never smoked	1
68356	Female	73	0	0	1	Self-employed	Urban	70.94	34.4	never smoked	1
1836	Female	51	1	0	1	Private	Urban	88.2	28.4	never smoked	1

id	gender	age	hypertension	heart disease	ever married	work type	residence type	avg glucose level	bmi	smoking status	stroke
*	A	B	C	D	E	F	Urban	[55 , 127]	I	formerly smoked	1
4219	Male	71	0	0	1	Private	Urban	102.87	27.2	formerly smoked	1
47472	Female	58	0	0	1	Private	Urban	107.26	38.6	formerly smoked	1
6118	Male	59	0	0	1	Private	Urban	86.23	30	formerly smoked	1
71673	Female	79	0	0	1	Private	Urban	110.85	24.1	formerly smoked	1
14499	Male	47	0	0	1	Private	Urban	86.94	41.1	formerly smoked	1
8154	Male	57	1	0	1	Govt_job	Urban	78.92	27.7	formerly smoked	1
31720	Female	38	0	0	0	Self-employed	Urban	82.28	24	formerly smoked	1
3512	Female	70	1	0	1	Self-employed	Urban	89.13	34.2	formerly smoked	1
44993	Female	79	1	0	0	Govt_job	Urban	98.02	22.3	formerly smoked	1
66204	Male	59	0	0	1	Private	Urban	111.04	32	formerly smoked	1
16077	Male	63	0	1	1	Self-employed	Urban	116.69	34.5	formerly smoked	1
66071	Male	51	1	0	1	Private	Urban	112.16	42.5	formerly smoked	1
51314	Female	78	0	0	1	Private	Urban	106.74	33	formerly smoked	1

id	gender	age	hypertension	heart disease	ever married	work type	residence type	avg glucose level	bmi	smoking status	stroke
*	A	B	C	D	E	F	Rural	H	I	J	1
70630	Female	71	0	0	1	Govt_job	Rural	193.94	22.4	smokes	1
64778	Male	82	0	1	1	Private	Rural	208.3	32.5	Unknown	1
56841	Male	58	0	1	1	Private	Rural	240.59	31.4	smokes	1
12062	Female	54	0	0	1	Self-employed	Rural	191.82	40.4	smokes	1
13491	Male	80	0	0	1	Private	Rural	259.63	31.7	smokes	1
30683	Female	75	0	0	1	Private	Rural	199.2	26.6	Unknown	1
63453	Female	56	0	0	1	Govt_job	Rural	162.23	27.3	Unknown	1
69112	Male	68	1	1	1	Private	Rural	271.74	31.1	smokes	1
54921	Male	78	1	0	1	Self-employed	Rural	134.8	33.6	Unknown	1
1210	Female	68	0	0	1	Private	Rural	211.06	39.3	Unknown	1

id	gender	age	hypertension	heart disease	ever married	work type	residence type	avg glucose level	bmi	smoking status	stroke
*	A	B	C	D	E	F	Rural	H	I	never smoked	1
1665	Female	79	1	0	1	Self-employed	Rural	174.12	24	never smoked	1
58202	Female	50	1	0	1	Self-employed	Rural	167.41	30.9	never smoked	1
19824	Male	76	1	0	1	Private	Rural	243.58	33.6	never smoked	1
59190	Female	79	0	1	1	Private	Rural	127.29	27.7	never smoked	1
50784	Male	63	0	0	1	Private	Rural	228.56	27.4	never smoked	1
2458	Female	78	0	0	1	Private	Rural	235.63	32.3	never smoked	1
63973	Female	77	0	0	1	Govt_job	Rural	190.32	31.4	never smoked	1
41069	Female	45	0	0	1	Private	Rural	224.1	56.6	never smoked	1
53401	Male	71	1	1	0	Govt_job	Rural	216.94	30.9	never smoked	1
44033	Male	56	1	0	1	Private	Rural	249.31	35.8	never smoked	1
53440	Female	73	1	0	1	Private	Rural	190.14	36.5	never smoked	1
69551	Male	69	1	0	0	Private	Rural	182.99	36.5	never smoked	1
20387	Female	68	1	0	1	Self-employed	Rural	206.09	26.7	never smoked	1
58978	Female	70	0	1	1	Private	Rural	239.07	26.1	never smoked	1
24669	Female	77	0	1	1	Private	Rural	231.56	36.9	never smoked	1
14431	Male	72	1	0	1	Self-employed	Rural	185.49	37.1	never smoked	1
31421	Male	73	0	1	1	Govt_job	Rural	219.73	28.6	never smoked	1
32729	Female	81	0	0	1	Private	Rural	184.4	27.5	never smoked	1

id	gender	age	hypertension	heart disease	ever married	work type	residence type	avg glucose level	bmi	smoking status	stroke
*	A	B	C	D	E	F	Rural	H	I	formerly smoked	1
47269	Male	74	0	0	1	Private	Rural	219.72	33.7	formerly smoked	1
25831	Male	63	0	1	1	Private	Rural	196.71	36.5	formerly smoked	1
2326	Female	67	1	0	1	Private	Rural	179.12	28.1	formerly smoked	1
45277	Female	74	0	0	1	Private	Rural	231.61	34.6	formerly smoked	1
56546	Male	79	0	1	1	Private	Rural	129.98	22.6	formerly smoked	1
29281	Male	76	1	0	1	Self-employed	Rural	194.37	27	formerly smoked	1
41081	Male	63	0	0	1	Private	Rural	137.3	31.7	formerly smoked	1
58267	Male	70	1	0	1	Private	Rural	242.52	45.5	formerly smoked	1
12689	Female	63	0	0	1	Govt_job	Rural	205.35	42.2	formerly smoked	1
36857	Male	77	0	0	1	Self-employed	Rural	162.14	32.6	formerly smoked	1

id	gender	age	hypertension	heart disease	ever married	work type	residence type	avg glucose level	bmi	smoking status	stroke
*	A	B	C	D	E	F	Rural	[55 , 127]	I	J	1
12095	Female	61	0	1	1	Govt_job	Rural	120.46	36.8	smokes	1
33879	Male	42	0	0	1	Private	Rural	83.41	25.4	Unknown	1
47306	Male	58	0	0	0	Private	Rural	92.62	32	Unknown	1
36338	Female	39	1	0	1	Private	Rural	58.09	39.2	smokes	1
65842	Female	67	1	0	1	Self-employed	Rural	61.94	25.3	smokes	1
57419	Male	59	0	0	1	Private	Rural	96.16	44.1	Unknown	1
32399	Male	54	0	0	1	Private	Rural	96.97	29.1	smokes	1
3253	Male	61	0	1	1	Private	Rural	111.81	27.3	smokes	1
4712	Female	81	0	1	1	Self-employed	Rural	78.7	19.4	Unknown	1
60744	Male	61	1	0	1	Self-employed	Rural	76.11	27.3	smokes	1
45965	Female	59	0	0	1	Private	Rural	116.44	23.8	smokes	1
43054	Female	50	0	0	1	Private	Rural	102.16	31.4	smokes	1
39912	Female	32	0	0	1	Private	Rural	76.13	29.9	smokes	1
36255	Male	59	0	0	1	Self-employed	Rural	118.03	35.5	smokes	1
23410	Female	72	0	0	1	Private	Rural	97.92	26.9	smokes	1
35684	Male	69	0	0	1	Private	Rural	93.81	28.5	Unknown	1
62019	Male	54	0	0	1	Govt_job	Rural	87.85	31.1	smokes	1
33454	Female	63	0	0	1	Govt_job	Rural	106.58	23.9	Unknown	1
8899	Male	49	0	0	0	Private	Rural	104.86	31.9	smokes	1

id	gender	age	hypertension	heart disease	ever married	work type	residence type	avg glucose level	bmi	smoking status	stroke
*	A	B	C	D	E	F	Rural	[55 , 127]	I	never smoked	1
31112	Male	80	0	1	1	Private	Rural	105.92	32.5	never smoked	1
53882	Male	74	1	1	1	Private	Rural	70.09	27.4	never smoked	1
12109	Female	81	1	0	1	Private	Rural	80.43	29.7	never smoked	1
70822	Male	80	0	0	1	Self-employed	Rural	104.12	23.5	never smoked	1
38829	Female	82	0	0	1	Private	Rural	59.32	33.2	never smoked	1
55927	Female	80	1	0	1	Private	Rural	74.9	22.2	never smoked	1
7371	Female	80	1	0	1	Self-employed	Rural	72.67	28.9	never smoked	1
19773	Female	52	0	0	1	Private	Rural	96.59	26.4	never smoked	1
26727	Female	79	0	0	0	Private	Rural	88.92	22.9	never smoked	1
54385	Male	45	0	0	1	Private	Rural	64.14	29.4	never smoked	1
30456	Female	79	0	0	1	Private	Rural	93.05	24.2	never smoked	1
20439	Male	82	0	1	1	Govt_job	Rural	103.68	25	never smoked	1
72081	Female	57	1	0	1	Govt_job	Rural	67.41	32.9	never smoked	1
56939	Female	55	0	0	1	Self-employed	Rural	92.98	25.6	never smoked	1
34567	Female	81	1	0	1	Self-employed	Rural	74.02	25	never smoked	1
210	Male	81	0	0	1	Self-employed	Rural	91.54	31.4	never smoked	1
8580	Female	77	0	0	1	Self-employed	Rural	90	32	never smoked	1
28484	Female	78	0	0	1	Self-employed	Rural	109.47	30.8	never smoked	1
37060	Female	81	0	0	1	Private	Rural	80.13	23.4	never smoked	1
68023	Male	79	0	0	1	Private	Rural	72.73	28.4	never smoked	1
10552	Female	81	0	0	1	Self-employed	Rural	81.95	16.9	never smoked	1

id	gender	age	hypertension	heart disease	ever married	work type	residence type	avg glucose level	bmi	smoking status	stroke
*	A	B	C	D	E	F	Rural	[55 , 127]	I	formerly smoked	1
38047	Female	65	0	0	1	Private	Rural	100.98	28.2	formerly smoked	1
712	Female	82	1	1	0	Private	Rural	84.03	26.5	formerly smoked	1
24977	Female	72	1	0	1	Private	Rural	74.63	23.1	formerly smoked	1
4651	Male	78	0	0	1	Private	Rural	78.03	23.9	formerly smoked	1
19557	Female	45	0	0	1	Private	Rural	93.72	30.2	formerly smoked	1
27169	Female	66	1	0	1	Govt_job	Rural	116.55	31.1	formerly smoked	1
66159	Female	80	0	1	1	Self-employed	Rural	66.72	21.7	formerly smoked	1
71796	Female	70	0	1	1	Private	Rural	59.35	32.3	formerly smoked	1
35512	Female	70	0	0	1	Self-employed	Rural	76.34	24.4	formerly smoked	1
42072	Female	50	1	0	1	Private	Rural	73.18	30.3	formerly smoked	1
68798	Female	58	0	0	1	Private	Rural	59.86	28	formerly smoked	1
2182	Female	80	1	0	1	Self-employed	Rural	91.02	32.9	formerly smoked	1
36841	Male	78	1	0	1	Self-employed	Rural	56.11	25.5	formerly smoked	1
30184	Male	82	0	0	1	Private	Rural	86.62	29.5	formerly smoked	1
62439	Female	51	0	0	1	Govt_job	Rural	103.43	27.3	formerly smoked	1

הרשומה הראשונה המודגשת היא זאת שתופיע במאגר הנתונים הסופי במקום הרשומות שנמצאות תחתיה (הרשומות המקוריות במאגר הנתונים).

נבחין, כי האלגוריתם אכן יוצר קבוצות של רשומות זהות אשר גודל כל קבוצה הוא לפחות k , ובכך אנו מאפשרים שמירה על פרטיות המטופלים. כמו כן, אנו שומרים על התכונה שכל רשומה אינה ניתנת להבדלה עם לפחות k-1 רשומות אחרות.

הצגנו אלגוריתם המזהה אנונימיזציות אופטימליות תחת מודל גמיש של הכללת ערכי תכונות. כמו כן, תקפנו את הבעיה דרך חיפוש על פני קבוצת חזקה אשר מכילה את כל הערכים הקיימים במאגר המידע. חיפוש זה התבצע באמצעות אסטרטגיית חיפוש עצים שמטרתה לחקור את האנונימיזציות החל מהכלליות ביותר ליותר ספציפיות. בנוסף, האלגוריתם משלב גיזום צמתים וכן סידור מחדש של ערכי זנב.

להלן מאגר הנתונים לאחר החלת האנונימיזציה:

id	gender	age	hypertension	heart disease	ever married	work type	residence type	avg glucose level	bmi	smoking status	stroke
*	A	B	C	D	E	F	Urban	H	I	J	1
*	A	B	C	D	E	F	Urban	H	I	J	1
*	A	B	C	D	E	F	Urban	H	I	J	1
*	A	B	C	D	E	F	Urban	H	I	J	1
*	A	B	C	D	E	F	Urban	H	I	J	1
*	A	B	C	D	E	F	Urban	H	I	J	1
*	A	B	C	D	E	F	Urban	H	I	J	1
*	A	B	C	D	E	F	Urban	H	I	J	1
*	A	B	C	D	E	F	Urban	H	I	J	1
*	A	B	C	D	E	F	Urban	H	I	J	1
*	A	B	C	D	E	F	Urban	H	I	J	1
*	A	B	C	D	E	F	Urban	H	I	J	1
*	A	B	C	D	E	F	Urban	H	I	J	1
*	A	B	C	D	E	F	Urban	H	I	J	1
*	A	B	C	D	E	F	Urban	H	I	J	1
*	A	B	C	D	E	F	Urban	H	I	J	1
*	A	B	C	D	E	F	Urban	H	I	never smoked	1
*	A	B	C	D	E	F	Urban	H	I	never smoked	1
*	A	B	C	D	E	F	Urban	H	I	never smoked	1
*	A	B	C	D	E	F	Urban	H	I	never smoked	1
*	A	B	C	D	E	F	Urban	H	I	never smoked	1
*	A	B	C	D	E	F	Urban	H	I	never smoked	1
*	A	B	C	D	E	F	Urban	H	I	never smoked	1
*	A	B	C	D	E	F	Urban	H	I	never smoked	1
*	A	B	C	D	E	F	Urban	H	I	never smoked	1
*	A	B	C	D	E	F	Urban	H	I	never smoked	1
*	A	B	C	D	E	F	Urban	H	I	never smoked	1
*	A	B	C	D	E	F	Urban	H	I	never smoked	1
*	A	B	C	D	E	F	Urban	H	I	never smoked	1
*	A	B	C	D	E	F	Urban	H	I	never smoked	1
*	A	B	C	D	E	F	Urban	H	I	never smoked	1
*	A	B	C	D	E	F	Urban	H	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	H	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	H	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	H	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	H	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	H	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	H	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	H	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	H	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	H	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	H	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	H	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	H	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	H	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	H	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	J	1

*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Urban	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	H	I	J	1
*	A	B	C	D	E	F	Rural	H	I	J	1
*	A	B	C	D	E	F	Rural	H	I	J	1
*	A	B	C	D	E	F	Rural	H	I	J	1
*	A	B	C	D	E	F	Rural	H	I	J	1
*	A	B	C	D	E	F	Rural	H	I	J	1
*	A	B	C	D	E	F	Rural	H	I	J	1
*	A	B	C	D	E	F	Rural	H	I	J	1
*	A	B	C	D	E	F	Rural	H	I	J	1
*	A	B	C	D	E	F	Rural	H	I	J	1
*	A	B	C	D	E	F	Rural	H	I	never smoked	1
*	A	B	C	D	E	F	Rural	H	I	never smoked	1
*	A	B	C	D	E	F	Rural	H	I	never smoked	1
*	A	B	C	D	E	F	Rural	H	I	never smoked	1
*	A	B	C	D	E	F	Rural	H	I	never smoked	1
*	A	B	C	D	E	F	Rural	H	I	never smoked	1
*	A	B	C	D	E	F	Rural	H	I	never smoked	1
*	A	B	C	D	E	F	Rural	H	I	never smoked	1
*	A	B	C	D	E	F	Rural	H	I	never smoked	1
*	A	B	C	D	E	F	Rural	H	I	never smoked	1
*	A	B	C	D	E	F	Rural	H	I	never smoked	1
*	A	B	C	D	E	F	Rural	H	I	never smoked	1
*	A	B	C	D	E	F	Rural	H	I	never smoked	1
*	A	B	C	D	E	F	Rural	H	I	never smoked	1
*	A	B	C	D	E	F	Rural	H	I	never smoked	1
*	A	B	C	D	E	F	Rural	H	I	never smoked	1
*	A	B	C	D	E	F	Rural	H	I	never smoked	1
*	A	B	C	D	E	F	Rural	H	I	never smoked	1

*	A	B	C	D	E	F	Rural	H	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	H	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	H	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	H	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	H	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	H	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	H	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	H	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	H	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	H	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	J	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	never smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	formerly smoked	1
*	A	B	C	D	E	F	Rural	[55 , 127]	I	formerly smoked	1

מקורות מידע:

1. **ויקיפדיה:**
https://he.wikipedia.org/wiki/%D7%94%D7%AA%D7%9E%D7%9E%D7%94 .
2. **המאמר שאנו מממשות:**
"Data Privacy Through Optimal K-Anonymization" אשר נכתב על ידי Roberto J.Bayardo ו- Rakesh Agrawal.
3. **ויקיפדיה:**
https://en.wikipedia.org/wiki/K-anonymity .
4. **K-Anonymity: Everything You Need to Know**
https://www.immuta.com/articles/k-anonymity-everything-you-need-to-know-2021-guide/