

Toán ứng dụng và thống kê

Ngày 30 tháng 7 năm 2021

Differential Privacy with k-Anonymity

Giáo viên lý thuyết: Nguyễn Đình Thúc



Mục lục

1 Thông tin cá nhân	
2 Giới thiệu bài toán	1
2.1 Vấn đề trong thực tế	1
2.2 Định nghĩa Differential Privacy	1
2.3 Ứng dụng trong thực tế	3
3 Một số thuật toán Differential Privacy	3
3.1 Nhóm các thuật toán thêm nhiễu vào kết quả truy vấn	4
3.1.1 Thuật toán Flip a coin	4
3.1.2 Phân phối Laplace	4
3.1.3 Cơ chế RAPPOR	4
3.1.4 Thuật toán RNM	5
3.1.5 Phương pháp PATE	5
3.2 Nhóm các thuật toán k -Anonymity	5
3.3 Mô hình học liên kết (Federated Learning)	5
4 Bài toán k-Anonymity	6
4.1 Định nghĩa k -Anonymity	7
4.2 Các thuật toán k -Anonymity	9
4.2.1 Thuật toán Datafly	10
4.2.2 Thuật toán Incognito	11
4.2.3 Thuật toán Top-Down Greedy	11
4.2.4 Thuật toán dựa trên k-NN Clustering	12
4.2.5 Thuật toán Mondrian	12
4.3 Dữ liệu sử dụng	13
4.4 Hàm đánh giá	14
4.4.1 Discernibility Metric (DM)	14
4.4.2 Average Equivalence class size Metric (C_{AVG})	14
4.4.3 Normalized Certainty Penalty (NCP)	14
4.5 Các mô hình máy học	16
5 Thí nghiệm	17
5.1 Thử nghiệm thuật toán Ẩn danh	17
5.2 Thử nghiệm các mô hình máy học	20

6	Tổng kết	25
7	Phụ lục	28

1 Thông tin cá nhân

MSSV	Họ và tên	SDT	Email	Mức độ đóng góp
18120019	Nguyễn Hoàng Dũng	0703924495	18120019@student.hcmus.edu.vn	100%
18120040	Nguyễn Đăng Khoa	0865316054	18120040@student.hcmus.edu.vn	100%
18120043	Phạm Minh Khôi	0775479216	18120043@student.hcmus.edu.vn	100%
18120051	Nguyễn Hoàng Lân	0789416557	18120051@student.hcmus.edu.vn	100%



2 Giới thiệu bài toán

2.1 Vấn đề trong thực tế

Trong thực tế, bất kì thông tin cá nhân lưu trữ dưới dạng điện tử đều có thể đối mặt với nguy hiểm. Ví dụ như khi bạn mở tài khoản Facebook của bạn, bạn đã gửi những thông tin nhận dạng cá nhân như là họ tên, địa chỉ, ngày sinh, tình trạng hôn nhân, Những thông tin này mang tính nhạy cảm và có thể bị tấn công.

Xem xét một trường hợp về công ty Target, công ty bán lẻ lớn thứ 2 nước Mỹ, đã quyết định thực hiện một cách tiếp cận có vẻ trực quan nhằm điều chỉnh trải nghiệm mua sắm của khách hàng theo nhu cầu cụ thể của họ vào năm 2012. Mục tiêu của họ là tìm ra khách hàng nào có nhiều khả năng sắp có con và từ đó, gửi cho những khách hàng này các quảng cáo liên quan đến thai sản dựa trên khả năng được tính toán. Ý tưởng của Target có thể hiểu là một phương pháp tiếp cận theo hai hướng: 1. Lưu trữ hoặc kết hợp dữ liệu để phân tích xu hướng của người mua đang mang thai. 2. Áp dụng kỹ thuật hoặc thuật toán để tìm tương quan giữa các điểm dữ liệu của khách hàng mới với mô hình mua của khách hàng trước đó để xác định khả năng một người là sắp có con hay không. Sáng kiến của Target đã trở thành trung tâm của nhiều cuộc thảo luận khi giao thoa giữa quyền riêng tư và thuật toán máy học khi Target gửi các quảng cáo liên quan đến thai sản (xen lẫn quảng cáo từ các bộ phận khác nhau) cho một cô gái tuổi teen trước khi gia đình cô ấy biết cô ấy đang chuẩn bị có con [5]. Những lo ngại về quyền riêng tư này không chỉ áp dụng cho việc thu thập và lưu trữ dữ liệu cho các mục đích tiếp thị mà còn cho các ứng dụng khác nhau, từ dữ liệu điều tra dân số đến truyền thông xã hội.

Một số giải pháp ban đầu sẽ là ẩn đi những thông tin nhận dạng cá nhân khi dữ liệu được truy vấn. Tuy nhiên, kể cả khi những thông tin đó được ẩn đi, thì danh tính của bạn vẫn có khả năng bị lộ. Một nghiên cứu [17] đã chứng minh giả thiết này bằng cách phân tích bộ dữ liệu Netflix Prize năm 2006 [1]. Họ đã sử dụng bộ dữ liệu IMDb để xác định danh tính của những người theo dõi Netflix dựa trên lịch sử xem phim của họ. Những thông tin cá nhân nhạy cảm như quan điểm chính trị của họ đã có thể bị truy ra được. Hay khi một người nghiên cứu thực hiện khảo sát để thống kê số lượng nam giới có từng quan hệ tình dục với gái mại dâm không. Người đó khảo sát một nhóm người ngẫu nhiên. Liệu người đó có thực sự quan tâm từng cá nhân nào đã trả lời có? Có lẽ là không. Tuy nhiên những dữ liệu đã thu thập có thể làm lộ những cá nhân nào. Sự thật này đã lôi kéo những nhà nghiên cứu, và dẫn tới sự hình thành của bài toán Differential Privacy.

2.2 Định nghĩa Differential Privacy

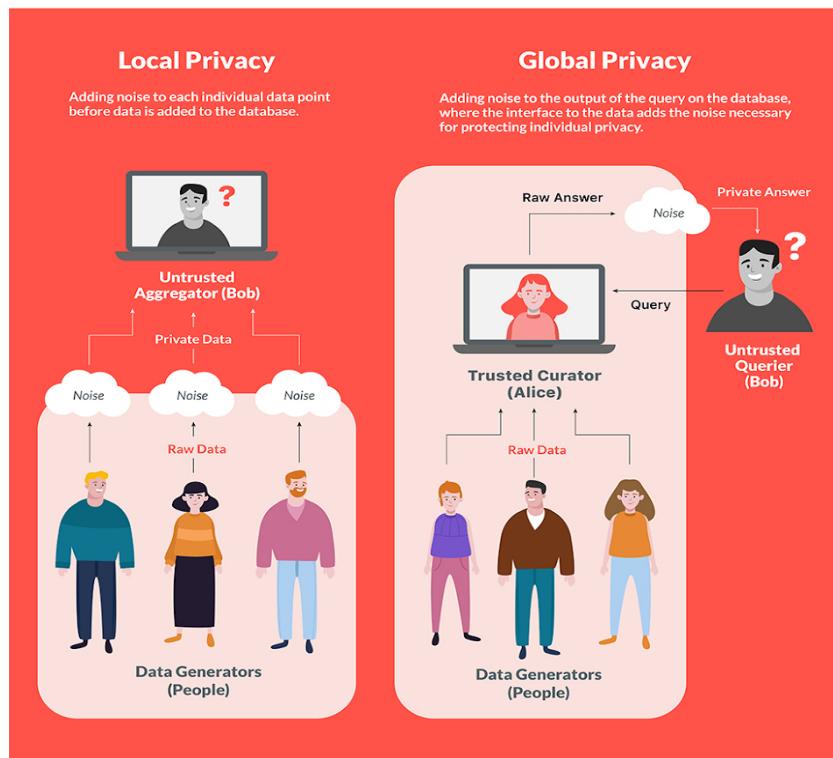
Differential Privacy (DP) là gì? DP là một định nghĩa toán học về tính bảo mật xung quanh nội dung của phân tích thống kê và máy học. Dựa trên định nghĩa này, DP là một tiêu chuẩn trong việc bảo vệ thông tin riêng tư, và cũng đã có nhiều công cụ dùng để phân tích dữ liệu nhạy cảm được xây dựng dựa trên nó [18].

Một thuật toán thỏa định nghĩa DP, khi mà nhìn vào kết quả của phép thống kê, ta không thể xác định cá nhân nào đã được thêm vào bộ dữ liệu; Mặt khác, nếu như khi thêm một cá nhân vào bộ dữ liệu, kết quả phép thống kê có sự thay đổi rõ rệt thì cá nhân đó có thể bị định danh. Thuật toán như thế này thì không thỏa DP. Ưu tiên

đầu tiên của DP bao gồm việc đảm bảo dữ liệu không bị tác động bởi những đối tượng khác trong bộ dữ liệu, đồng thời tối đa hóa độ tin cậy của dữ liệu khi truy vấn. Để tối ưu hóa mức độ bảo mật riêng tư mà vẫn để đảm bảo dữ liệu vẫn còn sử dụng được là một bài toán NP-hard, nên các phương pháp tiếp cận chỉ có thể mang tính xấp xỉ và cần sự đánh đổi (có được và mất).

Các phương pháp DP có thể chia thành 2 loại: Local DP và Global DP. Hình 1

- *Local Differential Privacy*: Các giá trị nhiễu được thêm vào mỗi điểm dữ liệu (có thể bởi người quản lý cơ sở dữ liệu hoặc bởi các cá nhân trước khi gửi đến server)
- *Global Differential Privacy*: Các giá trị nhiễu được thêm vào các thông tin trả về cho những truy vấn đến dữ liệu Tổng quát, global DP có thể dẫn đến kết quả truy vấn chính xác hơn local DP, trong khi độ bảo mật vẫn tương đương. Một khía cạnh khác để đạt được sự hiệu quả đó, người dùng cung cấp dữ liệu cho hệ thống cần tin tưởng hoàn toàn vào phía quản lý dữ liệu.



Hình 1: Hai loại Differential Privacy: Global và Local

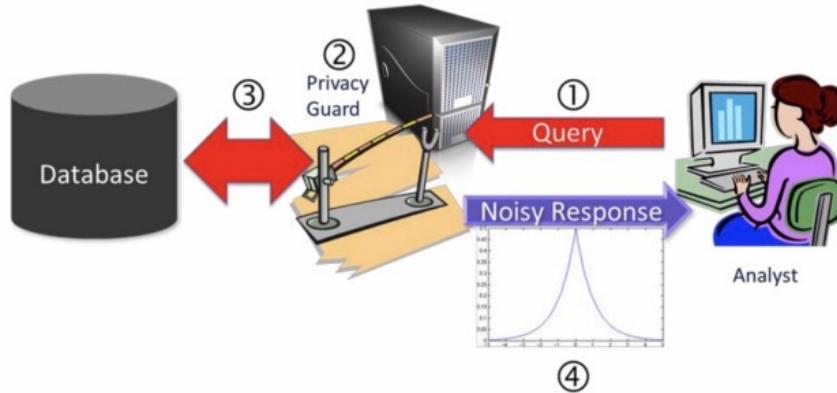
Ngoài ra, DP được phát biểu theo các ký tự toán học như sau: Một hàm ngẫu nhiên K đạt DP mức độ ϵ cho

mọi bộ dữ liệu D_1 và D_2 khác nhau nhiều nhất một phần tử với $S \subseteq Range(K)$ sao cho

$$Pr[K(D_1) \in S] \leq exp(\epsilon) \times Pr[K(D_2) \in S] [8] \quad (1)$$

2.3 Ứng dụng trong thực tế

Trong thực tế nhiều công ty công nghệ lớn đã áp dụng Differential privacy vào một số bài toán của họ. Vào năm 2016, Apple đã thông báo rằng công nghệ DP đã được áp dụng những phiên bản sau iOS10 để theo dõi thói quen sử dụng điện thoại của khách hàng nhưng không tác động đến các thông tin cá nhân. [3] Một số lập trình viên cho rằng việc này ảnh hưởng xấu đến họ vì khiến việc bán quảng cáo cá nhân khó khăn hơn. [2] Điều này chứng tỏ các quy chuẩn về tính riêng tư có thể bảo vệ khách hàng tốt hơn. Google áp dụng DP vào các hệ thống học máy bảo mật bậc nhất trên hạ tầng đám mây học của họ, đảm bảo dữ liệu chỉ nằm trên thiết bị cá nhân mà không cần truyền tải đi nơi khác. [4]



- Hình 2: DP trong thực tế [16]: 1. Nhà phân tích gửi câu truy vấn đến một ứng dụng.
 2. Lớp bảo vệ DP đánh giá mức độ tác động bảo mật của câu truy vấn đó bằng một thuật toán đặc biệt.
 3. Sau đó lớp bảo vệ này gửi câu truy vấn đến cơ sở dữ liệu và nhận lại một câu trả lời gốc.
 4. Sau đó lớp bảo vệ này sẽ thêm vào một lượng nhiễu thích hợp, dựa trên mức độ tác động bảo mật đã đánh giá trước đó, từ đó bảo vệ được tính bảo mật của những thông tin cá nhân liên quan. Câu phản hồi đã được bảo vệ này sẽ được gửi lại cho nhà phân tích

3 Một số thuật toán Differential Privacy

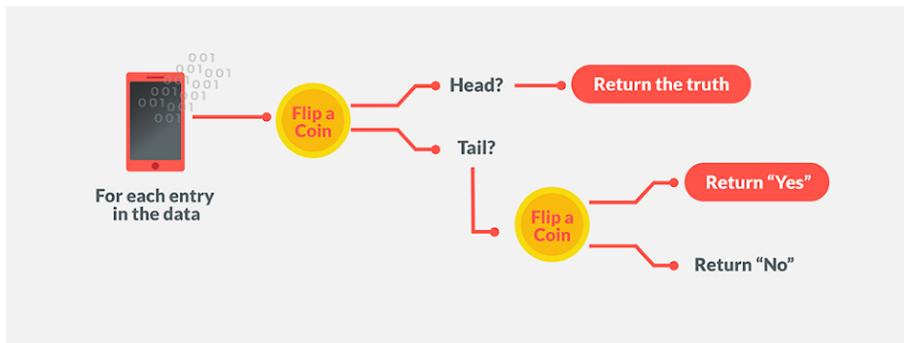
Chúng tôi cũng chia các phương pháp này ra thành 3 loại dựa trên cách hoạt động chung để tăng tính bảo mật cho dữ liệu của chúng: Nhóm các thuật toán thêm nhiễu kết quả truy vấn; Mô hình học liên kết và các thuật toán

đảm bảo tính chất K-Anonymity. Ở phần này chúng tôi sẽ xem xét qua một số thuật toán này, nhưng sau cùng sẽ chỉ tập trung phân tích sâu vào các thuật toán K-Anonymity, cũng là đề tài của báo cáo này.

3.1 Nhóm các thuật toán thêm nhiều vào kết quả truy vấn

3.1.1 Thuật toán Flip a coin

Thuật toán đơn giản nhất Tung đồng xu (Flip a coin) được mô tả trong hình 3 như sau: Một khi có sự truy vấn dữ liệu, thuật toán sẽ thực hiện tung một đồng xu, nếu là mặt ngửa thì trả lời dữ liệu gốc, ngược lại tung một đồng xu khác, nếu đồng này mặt ngửa thì trả lời “Yes” ngược lại là “No”.



Hình 3: Sơ đồ mô tả thuật toán Flip a coin

3.1.2 Phân phối Laplace

Hai cơ sở dữ liệu D_1 và D_2 được gọi là đồng nhất nếu chỉ khác nhau duy nhất 1 dòng và tồn tại 1 thủ tục $q_n()$ nào đó xáo trộn các kết quả nhận được khi chạy cùng query trên D_1 và D_2 , với xác suất không quá e^ϵ .

$$\frac{P[q_n(D_1) \in R]}{P[q_n(D_2) \in R]} \quad (2)$$

Với D_1 và D_2 là hai bộ dữ liệu, P là xác suất của một câu truy vấn từ hai bộ dữ liệu, q là phương pháp gây xáo trộn dữ liệu.

3.1.3 Cơ chế RAPPOR

Google áp dụng cơ chế RAPPOR để bảo mật thông tin khách hàng [10]. Cơ chế này cung cấp một sự đảm bảo sự an toàn cho các dữ liệu cá nhân đối với các truy vấn dạng thống kê từ phần mềm end-user hoặc client. RAPPOR được xây dựng trên ý tưởng của phương pháp Randomized Response [20] (phản hồi ngẫu nhiên), đây là một phương pháp được phát triển cho những cuộc khảo sát thông tin năm 1960s, dữ liệu sẽ được thêm nhiều vĩnh viễn vào dựa

trên các phép phân phối và xác suất. Những nhiễu này có thể là ‘1’, ‘0’ với xác suất $\frac{1}{f}$ hoặc một giá trị ngẫu nhiên có xác suất $1-f$.

$$B'_i = \begin{cases} 1, & \text{with probability } \frac{1}{2}f \\ 0, & \text{with probability } \frac{1}{2}f \\ B_i, & \text{with probability } 1 - f \end{cases} \quad (3)$$

Trong đó f là tham số điều chỉnh mức độ bảo mật. Sau đó nhiễu sẽ được lưu lại trong bộ nhớ và tái sử dụng cho tất cả truy vấn đối với giá trị này.

3.1.4 Thuật toán RNM

Report-noisy-max (RNM) được đề ra trong paper [9] là thuật toán đơn giản để xác định truy vấn đếm nào trong một danh sách các truy vấn và chọn riêng truy vấn có giá trị cao nhất: Thêm tạp âm Laplace được tạo độc lập $\frac{1}{\epsilon}$ vào mỗi số đếm và trả về chỉ số của số đếm nhiều lớn nhất.

3.1.5 Phương pháp PATE

Dựa trên RNM, phương pháp Private Aggregation of Teacher Ensembles (PATE) được dùng để giải quyết khả năng khai thác thông tin riêng tư bằng thuật toán Machine Learning, đây là cách tiếp cận áp dụng chung để cung cấp các đảm bảo quyền riêng tư mạnh mẽ cho dữ liệu huấn luyện.

Ý tưởng đầu tiên sau PATE là thuật toán RNM trên đầu ra các mô hình “nhạy cảm”, được gọi là “teachers” (Các mô hình nhạy cảm được đào tạo một mô hình khai thác chưa được gán nhãn). Bằng cách đó, việc áp dụng DP trên các phản hồi của teacher có thể được xem như một phương pháp ủy quyền để bảo vệ tính riêng tư của dữ liệu nhạy cảm và các giáo viên phải được huấn luyện trên “dữ liệu con rời rạc” của tập dữ liệu đào tạo.

Tại sao chúng phải rời rạc? Nếu mỗi giáo viên được đào tạo trên toàn bộ dữ liệu thì việc xóa một giáo viên không có bất kỳ ảnh hưởng nào đến việc tham gia bất kỳ chủ đề dữ liệu riêng lẻ nào trong kết quả tổng hợp, vì chủ đề dữ liệu cá nhân đó vẫn tham gia đào tạo các giáo viên khác. Điều này sẽ làm cho việc áp dụng Quyền riêng tư khác biệt đối với câu trả lời của giáo viên không còn là một proxy hợp lệ để bảo vệ quyền riêng tư của các chủ đề dữ liệu cá nhân trong dữ liệu nhạy cảm. Do đó, các tập dữ liệu huấn luyện phải rời rạc

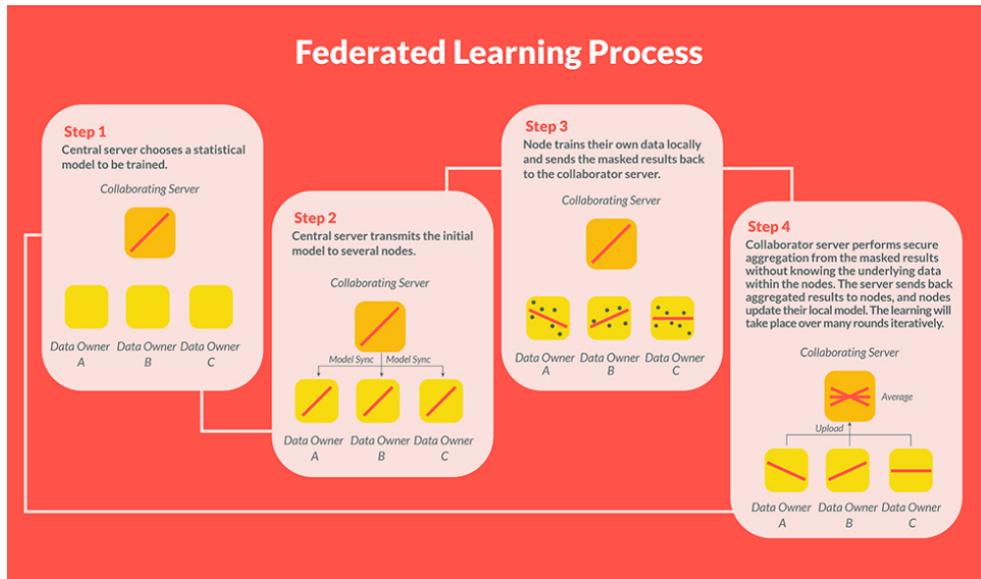
3.2 Nhóm các thuật toán k -Anonymity

Sẽ được giới thiệu ở phần 4.2

3.3 Mô hình học liên kết (Federated Learning)

Hiện nay, có một phương pháp phổ biến khác được sử dụng trong các mô hình máy học, học sâu đó là Federated Learning (FL). FL cho phép các mô hình này được huấn luyện qua sự phân quyền, nghĩa là dữ liệu của người dùng

sẽ luôn luôn nằm ở thiết bị của họ mà không cần truyền tải đến hệ thống server mà chỉ có trọng số của mô hình được truyền đi. Mô hình học máy ở server sẽ tổng hợp lại các thông tin đó và cập nhật lại trọng số của mình. Sau đó, theo định kỳ, server sẽ gửi trọng số của mô hình đến những thiết bị cá nhân để chúng cập nhật lại mô hình.



Hình 4: Quá trình thực hiện Federated Learning

Google hiện nay đang tiên phong trong việc sử dụng phương pháp này, một trong các ứng dụng đang được áp dụng mô hình này là Google Keyboard trên các thiết bị Android. Ngoài ra, Google đề xuất thêm một số phương pháp cải tiến để tăng tính bảo mật người dùng cũng như giảm chi phí của thuật toán [4].

4 Bài toán *k*-Anonymity

Trong báo cáo này, chúng tôi tập trung phân tích các phương pháp K-Anonymity khác nhau, so sánh độ hiệu quả của chúng trong việc tăng tính bảo mật của dữ liệu đồng thời thử nghiệm mức độ ảnh hưởng của phương pháp đến độ chính xác đối với các hệ thống mô hình Máy học hiện nay. Chúng tôi tham khảo và tận dụng lại nghiên cứu và code từ Slijepčević et. al [21], ngoài ra chúng tôi chỉ ra một số điểm chưa phù hợp của nghiên cứu trên (mô tả ở phần 5.2) và tiến hành cài đặt thêm một số phương pháp khác và thực hiện thêm thí nghiệm trên các mô hình máy học khác để đưa ra các đánh giá chi tiết hơn.

4.1 Định nghĩa *k*-Anonymity

K-Anonymity là một thuộc tính được định nghĩa cho các bộ dữ liệu ẩn danh. Khái niệm k-anonymity lần đầu tiên được đưa ra bởi Latanya Sweeney và Pierangela Samarati trong một bài báo xuất bản năm 1998 [19] là một nỗ lực để giải quyết vấn đề: "Với dữ liệu có cấu trúc trường dành riêng cho từng người, ta tạo ra một mẫu dữ liệu đảm bảo rằng những cá nhân có trong mẫu dữ liệu không thể bị xác định ngược lại đồng thời dữ liệu vẫn còn ý nghĩa trên thực tế." Một mẫu dữ liệu được coi là có thuộc tính k-anonymity nếu thông tin cho mỗi người có trong mẫu đó không thể phân biệt với ít nhất $k - 1$ cá nhân khác trong mẫu.

Một số định nghĩa:

- **Quasi-Identifier (QIDs)**: Là một nhóm các thuộc tính trong bảng dữ liệu, sao cho khi kết hợp các thuộc tính này lại có thể xác định từ một mẫu dữ liệu danh tính thật sự một cá nhân trong dữ liệu đó.
- **Sensitive attributes (SA)**: là các thuộc tính chứa các thông tin đặc biệt nhạy cảm của một cá nhân. Một cá nhân chắc chắn sẽ không muốn bị lộ thông tin này.
- **Non-Sensitive attributes (non-SA)** là những thuộc tính còn lại không có tính chất đặc biệt.
- **Equivalence class (EQ)** là một bộ các thực thể có giá trị các thuộc tính QID giống nhau

Ngoài ra, k-anonymity còn có các định nghĩa chặt chẽ hơn để tăng cường tính bảo mật dữ liệu:

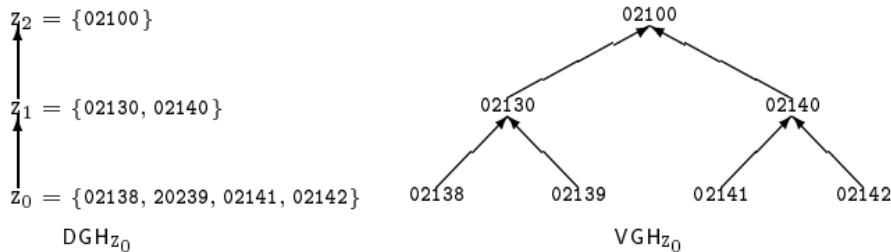
- **L-diversity**. l-diversity đảm bảo rằng mỗi nhóm đã đạt k-anonymity tồn tại thêm ít nhất l giá trị sensitive attributes khác nhau. Một vấn đề khiến k-anonymity có điểm yếu là khi một cá nhân đã thuộc một nhóm k-ẩn danh, nhưng nhóm này lại có chung giá trị của thuộc tính nhạy cảm (SA). Nếu biết được các cá nhân thuộc nhóm này, thông tin nhạy cảm của cá nhân sẽ bị lộ. Vấn đề này sẽ được giải quyết nếu ràng buộc l-diversity được áp dụng.
- **T-closeness**. T-closeness là một ràng buộc chặt chẽ hơn nữa cho l-diversity, bằng cách giảm mức độ chi tiết của phân phối dữ liệu. Mô hình t-closeness mở rộng ra từ l-diversity kết hợp thêm thông tin phân bố của dữ liệu, đảm bảo rằng phân bố xác suất của các thuộc tính nhạy cảm (SA) trong một nhóm k-ẩn danh sẽ tương đồng với phân bố của thuộc tính đó trong toàn bộ dữ liệu.

Việc ẩn danh bộ dữ liệu đảm bảo thông tin của một người không có khả năng bị truy xuất ngược lại để xác định người này. K-Anonymity [19] là một mô hình hỗ trợ việc này, dựa trên mô hình này đã phát triển lên các phương pháp như L-Diversity [15], T-Closeness [13]. Một bộ dữ liệu thỏa mãn k-anonymity, ta cần phải biến đổi các thuộc tính quasi-identifiers (QIDs) - là các thuộc tính như vị trí, tuổi, dân tộc, tôn giáo hoặc giới tính có thể dùng để xác định đích danh một cá nhân. Phép biến đổi này phải chắc chắn rằng mỗi bộ dữ liệu phải có chung giá trị với ít nhất $k-1$ bộ còn lại trong cơ sở dữ liệu.

Các mô hình k-anonymity đều có thể sử dụng chung phương pháp tổng quát hóa (generalization). Đây là phương pháp thay thế các giá trị của một thuộc tính bằng một giá trị tổng quát hơn; thường đạt được bằng cách tận dụng

phân cấp tổng quát theo miền hoặc theo giá trị (domain/value generalisation hierarchies). Phương pháp phân cấp tổng quát sử dụng mối quan hệ giữa các thuộc tính dựa trên miền của thuộc tính đó.

Ví Dụ về phương pháp tổng quát hóa:



Hình 5: Cây phân tầng tổng quát hóa dữ liệu

Như trong hình 5. Thay vì để các giá trị ZIP code rõ ràng như ở mức Z_0 , ta có thể tổng quát lên theo nhiều mức độ, như Z_1 (thay tất cả chữ số cuối thành 0) hoặc Z_2 (thay 2 chữ số cuối thành 0). Mức độ tổng quát hóa sẽ phụ thuộc vào số k . Ví dụ xét bảng sau:

Eth:E ₀	ZIP:Z ₀
asian	02138
asian	02138
asian	02142
asian	02142
black	02138
black	02141
black	02142
white	02138

PT

Eth:E ₁	ZIP:Z ₀
person	02138
person	02138
person	02142
person	02142
black	02138
black	02141
black	02142
person	02138

GT_[1,0]

Eth:E ₀	ZIP:Z ₁
asian	02130
asian	02130
asian	02140
asian	02140
black	02130
black	02140
black	02140
white	02130

GT_[0,1]

Eth:E ₀	ZIP:Z ₂
asian	02100
black	02100
black	02100
black	02100
white	02100

GT_[0,2]

Eth:E ₁	ZIP:Z ₁
person	02130
person	02130
person	02140
person	02140
black	02130
black	02130
black	02140
person	02140

GT_[1,1]

Hình 6: Một số kết quả sau khi k-anonymize

PT là bảng ban đầu cần đạt k-anonymity. Sau khi tổng quát đặc trưng Eth lên mức độ 1, tức là từ asian, black, white thì tổng quát thành person, thì vẫn chỉ có 1 person có zipcode là 02141, và có thể map 1:1 lại với PT. Nếu mong muốn k-anonymity > 1 thì cần phải tổng quát dữ liệu hơn nữa. Đối với phương pháp Nén, thay vì sẽ tiến hành tổng quát dữ liệu hơn nữa để đảm bảo k-anonymity thì ta sẽ bỏ luôn các thẻ hiện đó.

Các thuật toán Tổng quát hóa có thể chia thành hai loại [7]: toàn cục và địa phương. Thuật toán toàn cục áp dụng các bước tổng quát hóa giống nhau cho mỗi thuộc tính, khi đó các giá trị cập nhật của thuộc tính đều có giá trị giống nhau. Ngược lại, tổng quát hóa dạng địa phương chỉ áp dụng vào một số giá trị của thuộc tính, trong khi các giá trị khác vẫn giữ nguyên, việc này cho phép dữ liệu được xử lý chi tiết hơn để đạt được k-anonymity với ít biến dạng dữ liệu hơn. Hình 7 là một ví dụ về 2 loại này.

Row-id	Age	Zipcode
R1	24	53712
R2	25	53711
R3	30	53711
R4	30	53711
R5	32	53712
R6	32	53713

Row-id	Age	Zipcode
R1	[24-32]	[53712-53713]
R2	[25-30]	53711
R3	[25-30]	53711
R4	[25-30]	53711
R5	[24-32]	[53712-53713]
R6	[24-32]	[53712-53713]

Row-id	Age	Zipcode
R1	[24-30]	[53711-53712]
R2	[24-30]	[53711-53712]
R3	[24-30]	[53711-53712]
R4	[30-32]	[53711-53713]
R5	[30-32]	[53711-53713]
R6	[30-32]	[53711-53713]

Hình 7: Tổng quát địa phương và toàn cục. Từ trái sang phải: Dữ liệu gốc, sau khi tổng quát hóa địa phương, sau khi tổng quát hóa toàn cục

Thuật toán tổng quát hóa toàn cục có thể chia ra thành các thuật toán đơn chiều và đa chiều. Thuật đơn chiều áp dụng tổng quát hóa vào riêng lẻ các thuộc tính, ngược lại đa chiều xem xét chung một nhóm các thuộc tính để tìm một cách tổng quát phù hợp cho dữ liệu.

Age	Sex	Zipcode	Disease
25	Male	53711	Flu
25	Female	53712	Hepatitis
26	Male	53711	Brochitis
27	Male	53710	Broken Arm
27	Female	53712	AIDS
28	Male	53711	Hang Nail

Age	Sex	Zipcode	Disease
[25-28]	Male	[53710-53711]	Flu
[25-28]	Female	53712	Hepatitis
[25-28]	Male	[53710-53711]	Brochitis
[25-28]	Male	[53710-53711]	Broken Arm
[25-28]	Female	53712	AIDS
[25-28]	Male	[53710-53711]	Hang Nail

Age	Sex	Zipcode	Disease
[25-26]	Male	53711	Flu
[25-27]	Female	53712	Hepatitis
[25-26]	Male	53711	Brochitis
[27-28]	Male	[53710-53711]	Broken Arm
[25-27]	Female	53712	AIDS
[27-28]	Male	[53710-53711]	Hang Nail

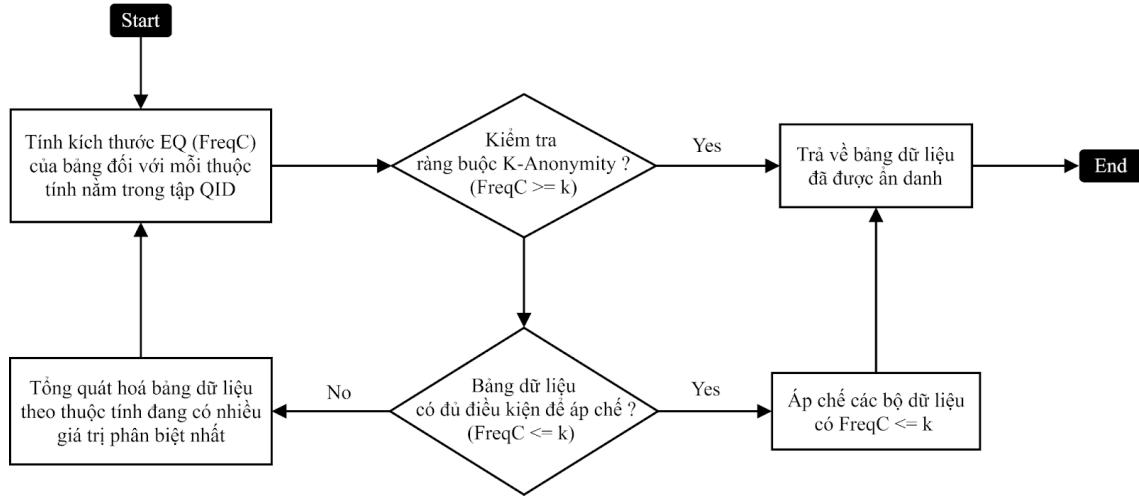
Hình 8: Tổng quát hóa đơn chiều và đa chiều. Từ trái sang phải: Dữ liệu gốc, phương pháp đơn chiều, phương pháp đa chiều

Trong hình 8, ví dụ k-anonymity với $k = 2$ trên dữ liệu các bệnh nhân sử dụng phương pháp tổng quát hóa đa và đơn chiều được thể hiện trong hình 25. Nhận thấy rằng sự khác nhau ở đây là thuộc tính Age và Zipcode trong phương pháp đa chiều có thể được ánh xạ vào các miền khác nhau. (chẳng hạn Age với giá trị 25 có thể được ánh xạ thành [25-26] hoặc [25-27] tùy thuộc vào giá trị Zipcode và Sex). Trong khi phương pháp đơn chiều ánh xạ vào chung một miền.

4.2 Các thuật toán k -Anonymity

Đã tồn tại khá nhiều các thuật toán được đề xuất trước đó để giải quyết bài toán k-anonymity, như là thuật toán Datafly, thuật toán Top-down Greedy, thuật toán dựa trên phương pháp phân cụm, hay thuật toán Mondrian và các phiên bản mở rộng của nó. Trong phạm vi báo cáo này, chúng tôi thực hiện, phân tích, cài đặt, đánh giá và so sánh các phương pháp này để xem khả năng ứng dụng của chúng vào thực tế. Như đã đề cập trước đó, báo cáo này phụ thuộc phần lớn vào nghiên cứu và mã nguồn được hoàn thành gần đây vào năm 2021 của Slijepčević et al. [21]

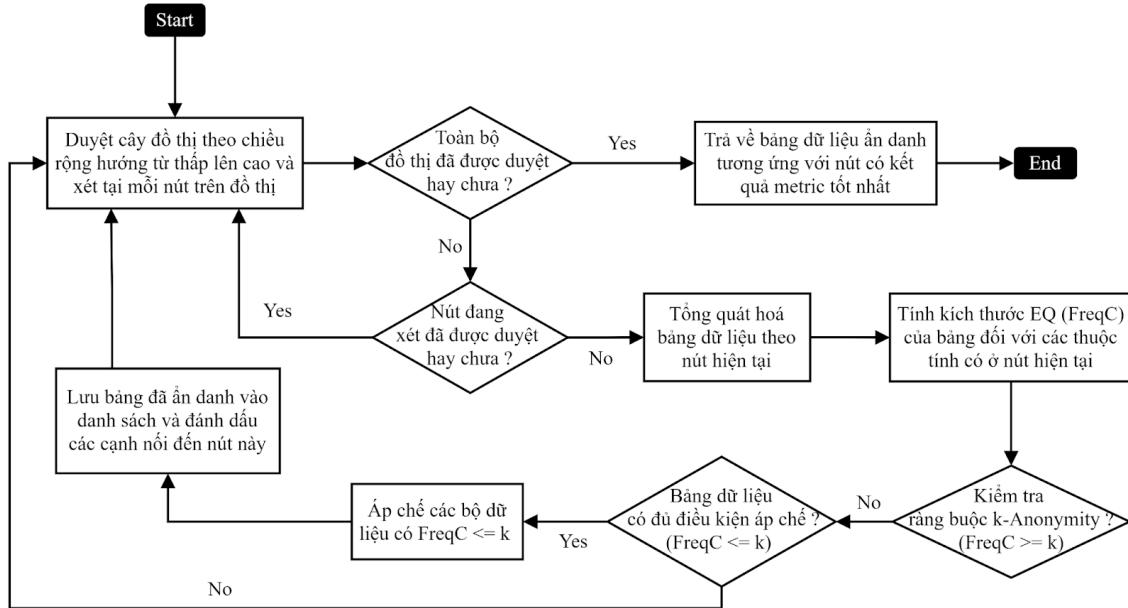
4.2.1 Thuật toán Datafly



Hình 9: Sơ đồ thuật toán Datafly

Datafly là thuật toán ẩn danh thường được áp dụng cho bộ dữ liệu y khoa, thuật toán được đề xuất bởi Latanya Arvette Sweeney vào năm 1997 [22]. Thuật toán giả định rằng giải pháp tốt nhất sẽ là những giải pháp tổng quát hóa các biến có nhiều giá trị khác biệt nhất. Sự ẩn danh đạt được bằng cách tự động lựa chọn các thuộc tính với nhiều giá trị khác biệt nhất và thực hiện các việc tổng quát hóa, thay thế, thêm, xóa, sửa các thông tin một cách hợp lý mà không mất đi nhiều thông tin quan trọng. Không gian tìm kiếm của thuật toán là toàn bộ mạng lưới thuộc tính. Tuy nhiên thuật toán chỉ duyệt qua một số đỉnh để đưa ra đáp án. Thuật toán hoạt động dựa trên toàn bộ cây phân tầng và tiếp cận theo phương pháp tham lam kết hợp heuristic. Thuật toán thực hiện tính toán tần suất của các QID và tổng quát hóa các thuộc tính có nhiều giá trị khác biệt nhất cho đến khi k-anonymity không còn thỏa. Trong khi thuật toán này rất hiệu quả về mặt về thời gian, tuy nhiên gặp phải hạn chế là có thể mắc kẹt ở cực tiểu địa phương.

4.2.2 Thuật toán Incognito



Hình 10: Sơ đồ thuật toán Incognito

Incognito được đề xuất bởi Lefevre et al. vào năm 2005 [12]. Phương pháp sử dụng hướng tiếp cận quy hoạch động, hoạt động dựa trên toàn bộ cây phân tầng và sử dụng phương pháp tổng quát hóa đơn chiều. Thuật xây dựng một đồ thị dựa trên các mức độ tổng quát của thuộc tính và duyệt qua đồ thị đó bằng phương pháp duyệt theo chiều rộng (breadth-first search) theo hướng bottom-up. Sau khi duyệt qua toàn bộ đồ thị sẽ tìm được một bảng dữ liệu ẩn danh tương ứng theo từng node đồ thị. Incognito có thể tìm được nhiều bảng dữ liệu thỏa k-anonymity, sau đó áp dụng các hàm đánh giá để tìm ra bảng có điểm tốt nhất.

4.2.3 Thuật toán Top-Down Greedy

Xu et al. [6] đề xuất một phương pháp tổng quát hóa địa phương heuristic đơn giản. Thuật toán nhận vào một bảng dữ liệu và phân vùng một cách để qui thành các lớp tương đương càng lúc càng cục bộ hơn. Với mục đích này, phân vùng nhị phân kết hợp với heuristic được sử dụng để chia nhỏ dữ liệu trong mỗi lần lặp.

Nghiên cứu còn đưa ra hàm đánh giá normalised certainty penalty (NCP). NCP kết hợp cả việc mất thông tin do ẩn danh hóa cũng như tầm quan trọng của các thuộc tính, sẽ được mô tả rõ ở phần 4.4.3. Các bộ giá trị ban đầu cho mỗi lớp tương đương được tìm thấy bằng cách chọn ngẫu nhiên một bộ giá trị u và tính NCP với mọi bộ giá trị v khác; sau đó, bộ v với NCP cao nhất được sử dụng làm điểm bắt đầu cho một lần lặp khác, khi này NCP sẽ được tính toán lại với tất cả các bộ khác. Quá trình này được lặp lại cho đến khi điểm NCP không thay đổi đáng

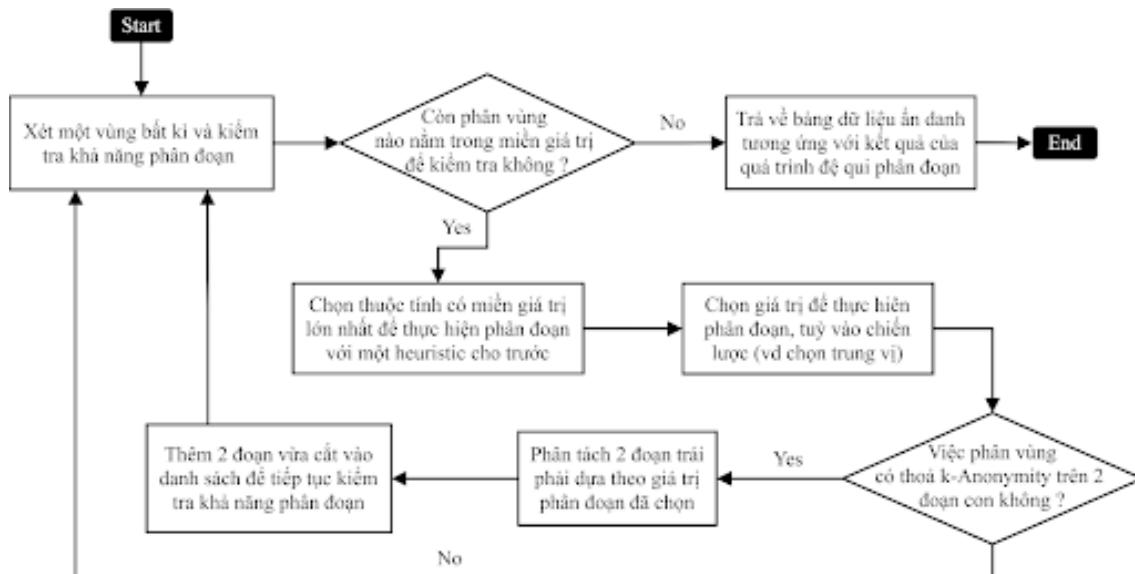
kể nữa, khi đó ta cố định hai bộ giá trị làm cơ sở cho sự phân vùng. Các bộ dữ liệu của bảng sau đó được gán cho một trong hai lớp tương đương sao cho giảm thiểu NCP tối đa.

Sau khi phân vùng hoàn tất, tất cả các lớp tương đương chứa ít hơn k phần tử được hậu xử lý để đạt được k -anonymity. Đối với mỗi lớp tương đương G như vậy, hai bước tiếp theo được áp dụng: Bước đầu tiên là tìm kiếm trong tất cả các lớp tương đương có kích thước ít nhất là $2k - |G|$, và tập con G_s các bộ có kích thước $k - |G|$ với độ lỗi NCP thấp nhất ($G \cup G_s$). Giải pháp có độ lỗi tổng thể nhỏ nhất sẽ được áp dụng; toàn bộ quá trình sau đó được lặp lại cho đến khi tất cả các lớp tương đương có kích thước ít nhất là k .

4.2.4 Thuật toán dựa trên k-NN Clustering

Một phương pháp khác để gom nhóm các bộ dữ liệu để các bộ thỏa k -anonymity là phương pháp gom cụm, hướng tới việc phân vùng các bộ dữ liệu thành lớp các bộ tương tự [14]. Cách thức phân cụm này cũng tương tự hướng tiếp cận của phân cụm k -láng giềng gần nhất. Ở mỗi vòng lặp, một bộ dữ liệu sẽ được chọn ngẫu nhiên và $k - 1$ bộ gần nhất dựa trên một hàm khoảng cách sẽ được chọn cùng. Những bộ này sẽ được gán vào một lớp tương đương và xóa bỏ khỏi dữ liệu gốc. Quy trình này được lặp lại tới khi tất cả bộ đã được xử lý và thuộc về một lớp. Thuật toán này sử dụng NCP (đã đề cập ở trên) làm hàm khoảng cách, khoảng cách được tính toán giữa các thuộc tính của các bộ.

4.2.5 Thuật toán Mondrian



Hình 11: Sơ đồ thuật toán Mondrian

Giữa các phương pháp áp dụng tổng quát hóa toàn cục, Mondrian [11] là phương pháp thành công về việc bảo quản thông tin và thời gian thực thi. Mondrian là một thuật toán tham lam giúp dữ liệu đạt được k-anonymity bằng cách phân vùng không gian miền thành các vùng đa chiều. Mondrian hoạt động theo kiểu top-down, thuật bắt đầu tổng quát hóa QIDs từ mức tổng quát cao nhất và đệ qui dần vào các vùng nhỏ hơn bằng các phép cắt đa chiều cho đến khi không thể cắt được nữa. Mỗi vòng lặp của thuật toán chọn ra một chiều (thuộc tính) để thực hiện việc cắt. Việc chọn sẽ ưu tiên chiều có miền giá trị lớn nhất. Vị trí cắt sẽ là vị trí trung vị của miền, phép cắt sẽ chia đôi miền này. Mondrian được có thể thực hiện cả phân vùng toàn cục hoặc địa phương. Thuật toán Mondrian gốc đã có thể xử lý các thuộc tính giá trị rời rạc, LeFevre et al. [11] đề ra một phiên bản mở rộng hơn bằng cách tận dụng cây Phân tầng tổng quát giá trị: Basic Mondrian. Hoặc phiên bản state-of-the-art Relaxed Mondrian sử dụng tổng quát hóa cục bộ.

4.3 Dữ liệu sử dụng

Các bộ dữ liệu chúng tôi sử dụng để thử nghiệm bao gồm 4 bộ khác nhau:

The Adult Dataset (ADULT)¹ (hay Census Income Dataset) chứa 30,162 bộ dữ liệu trích từ cơ sở dữ liệu thống kê dân số của Hoa Kỳ năm 1994. Các thuộc tính bao gồm giới tính, tuổi tác, chủng tộc, tình trạng hôn nhân, trình độ giáo dục, quốc tịch, tầng cấp, nghề nghiệp và mức lương; chúng tôi sử dụng mức lương để phân lớp nhị phân, còn những thuộc tính còn lại là các QIDs. Cây phân tầng tổng quát được mô tả trong phần Phụ lục.

The California Housing Prices Dataset (CAHOUSING)² chứa 20,640 bộ. Chúng tôi sử dụng các thuộc tính tuổi của ngôi nhà, giá trị ngôi nhà, thu nhập của chủ nhà và tọa độ (kinh, vĩ độ) là các QIDs và vị trí của nhà so với biển là giá trị cần dự đoán. Cây phân tầng tổng quát được mô tả trong phần Phụ lục.

The Contraceptive Method Choice Dataset (CMC)³ chứa 1,473 dòng. Bộ dữ liệu là tập con từ một cuộc khảo sát biện pháp tránh thai ở Indonesia năm 1987, bao gồm các đặc điểm nhân khẩu học và kinh tế học của những phụ nữ chưa mang thai về các biện pháp họ sử dụng. Chúng tôi chọn các thuộc tính như tuổi, trình độ học vấn và số con của họ làm QIDs và biện pháp tránh thai họ sử dụng làm nhãn dự đoán (gồm 3 lớp: không sử dụng, biện pháp ngắn hạn, biện pháp dài hạn). Cây phân tầng tổng quát được mô tả trong phần Phụ lục.

The Mammographic Mass Dataset (MGM)⁴ bao gồm 830 bộ với dữ liệu lấy từ phân tích chụp nhũ ảnh. Bộ dữ liệu gồm thông tin của bệnh nhân và kết quả xét nghiệm liệu tổn thương mô của họ là ác tính hay lành tính. Chúng tôi sử dụng tuổi, hình dạng, điểm đánh giá, biên độ và mật độ là QIDs và nhãn là mức độ nghiêm trọng của khối u. Cây phân tầng tổng quát được mô tả trong phần Phụ lục.

INFORMS Data Mining Dataset (INFORMS)⁵ bao gồm 102,580 thông tin về bệnh nhân của một bệnh viện ẩn danh, được công bố trong một cuộc thi về khai thác thông tin. Bộ dữ liệu gồm thông tin bệnh nhân như

¹<https://www.kaggle.com/uciml/adult-census-income>

²<https://www.kaggle.com/camnugent/california-housing-prices>

³<https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice>

⁴<https://archive.ics.uci.edu/ml/datasets/Mammographic+Mass>

⁵<https://sites.google.com/site/informsdataminingcontest/home>

giới tính, ngày sinh, chủng tộc, tình trạng học vấn, hôn nhân,... Tất cả đều được số hóa nhằm tránh việc lộ danh tính bệnh nhân. Để đảm bảo hơn về tính bảo mật cho bộ dữ liệu, nhóm thử nghiệm các phương pháp ẩn danh trên bộ này và đánh giá chúng.

Italia Healthcare Dataset là một bộ dữ liệu nhỏ nhom sưu tầm được gồm 100 mẫu dữ liệu về bệnh nhân mắc bệnh hiếm nghèo ở nước Ý. Nhóm dự định sử dụng bộ dữ liệu này để đánh giá các thuật toán k-anonymity có độ phức tạp cao, thời gian chạy không khả thi trên các dữ liệu lớn như thuật toán Datafly hay Cluster-based.

4.4 Hàm đánh giá

Chúng tôi xem xét một số hàm đánh giá tiêu chuẩn cho bài toán k-anonymity. Các hàm cho thấy độ mất mát thông tin cũng như khả năng tối ưu của việc tổng quát hóa, ngoài ra còn đánh giá liệu sau khi ẩn danh thì dữ liệu có mức "ý nghĩa" bao nhiêu.

4.4.1 Discernibility Metric (DM)

Điểm đánh giá discernibility (độ rõ ràng) được sử dụng để tính toán khả năng phân biệt giữa các thực thể của dữ liệu trong bảng T . Một giá trị lỗi được gán cho mỗi thực thể của dữ liệu được tính bằng kích thước của EQ mà thực thể đó thuộc vào. Nếu thực thể bị nén ra khỏi bộ dữ liệu thì độ lỗi bằng với kích thước của bảng. Ý tưởng dồn sau metric này là với EQ càng lớn nghĩa là độ mất mát thông tin càng lớn, do đó tối thiểu metric này là mục tiêu của các thuật toán ẩn danh. Biểu diễn theo công thức, DM của một bảng ẩn danh T được mô tả như sau [7]:

$$DM(T) = \sum_{EQ \subset T, |EQ| \geq k} |EQ|^2 + \sum_{EQ \subset T, |EQ| < k} |EQ| * |T| \quad (4)$$

trong đó $|EQ|$ là kích thước của equivalence class, $|T|$ là kích thước của bảng dữ liệu.

4.4.2 Average Equivalence class size Metric (C_{AVG})

Metric này mô tả việc tạo ra các EQ tốt như thế nào, với mỗi thực thể của bảng dữ liệu được tổng quát hóa thành các nhóm EQ kích thước k . Do đó, độ lỗi này càng thấp thì thuật toán ẩn danh càng tốt: giá trị bằng 1 ám chỉ độ tổng quát hóa tối ưu khi kích thước của tất cả các EQ bằng giá trị k . Điểm đánh giá C_{AVG} của bảng đã ẩn danh T được tính như sau [7]:

$$C_{AVG}(T) = \frac{|T|}{|EQs| \times k} \quad (5)$$

với $|T|$ là số thực thể của bảng, $|EQs|$ là tổng số lượng bộ EQ trong dữ liệu và k là ràng buộc k-anonymity.

4.4.3 Normalized Certainty Penalty (NCP)

Các nghiên cứu trước đây đã đề xuất ra các hàm để đánh giá hiệu quả bảo mật cũng như lượng mất mát thông tin khi sử dụng các thuật toán ẩn danh dữ liệu. Tuy nhiên chúng ta có thể có một metric đánh giá được độ hữu

dụng của dữ liệu bị ẩn danh không? NCP được giới thiệu cho công việc này bởi Vanessa Ayala-Rivera et al. [7] (được gọi là **Generalized Information Loss (GenILoss)** trong nghiên cứu của họ), xem xét 2 yếu tố sau:

- Mất thông tin do ẩn danh. Khi một bản ghi được ẩn danh, nó được khái quát hóa với các QID của nó. Metric phải đo lường được sự mất mát thông tin của sự tổng quát hóa đối với dữ liệu ban đầu.
- Tầm quan trọng của các thuộc tính. Chẳng hạn như trong phân tích dữ liệu, chẳng hạn như truy vấn tổng hợp, các thuộc tính khác nhau có thể mang tầm quan trọng khác nhau trong phân tích dữ liệu. Việc ẩn danh hóa để tối ưu tính hữu dụng của dữ liệu có thể tăng chất lượng của việc phân tích dữ liệu sau đó.

NCP được tính dựa trên cây phân tầng tổng quát hóa, được hiểu đơn giản là dùng khoảng giá trị của thuộc tính của các bộ chia cho khoảng giá trị của thuộc tính của toàn bộ dữ liệu. Do các thuộc tính có hai loại là rời rạc (ví dụ như zipcode, giới tính) và liên tục (số tuổi, thu nhập), nên khoảng giá trị của chúng sẽ được tính toán khác nhau. Sau đây sẽ mô tả cách tính cho từng loại thuộc tính.

Đối với thuộc tính liên tục, gọi T là một bảng dữ liệu và các QID (A_1, \dots, A_n) là các thuộc tính liên tục. Giả sử một bộ $t = (x_1, \dots, x_n)$ được tổng quát hóa thành $t' = ([l_1, r_1], \dots, [l_n, r_n])$ sao cho $l_i \leq x_i \leq r_i$ ($1 \leq i \leq n$). Khi này NCP của 1 thuộc tính liên tục A_i của thực thể t là:

$$NCP_{A_i}(t) = \frac{r_i - l_i}{|A_i|} \quad (6)$$

với $|A_i|$ là hiệu của giá trị max và min của miền liên tục A_i

Đối với thuộc tính rời rạc, giả sử một bộ t có giá trị v_i của thuộc tính rời rạc A_i . Giả sử thêm giá trị đó được tổng quát hóa thành khoảng $u = v_l, \dots, v_r$ với $u = v_l, \dots, v_r$ là tầng trên của v_i trong cây phân tầng. NCP khi này của 1 thuộc tính rời rạc của thực thể t là:

$$NCP_{A_i}(t) = \frac{\text{leaf}(u)}{|A_i|} \quad (7)$$

với $|A_i|$ là số giá trị khác biệt của thuộc tính A_i , và $\text{leaf}(u)$ là số node lá thuộc nhánh con của node u .

Kết hợp 2 biểu thức trên, đối với một bảng dữ liệu tổng quát T , tổng độ lỗi được tính như sau [7]:

$$NCP(T) = \frac{1}{|T| \times n} \sum_{t \in T} \sum_{i=1}^n NCP_{A_i}(t) \quad (8)$$

với n là số lượng thuộc tính QID, A_i là thuộc tính thứ i , $|T|$ là tổng số lượng thực thể và t là các thực thể của bảng T .

Khoảng giá trị của NCP nằm trong khoảng $[0, 1]$, với 0 nghĩa là không có mất mát thông tin, 1 nghĩa là tất cả thông tin mất hết. Một thuật toán ẩn danh tối ưu độ hữu dụng của dữ liệu khi tối thiểu được hàm lỗi NCP này.

4.5 Các mô hình máy học

Có thể thấy, trong các định nghĩa của các metric trên, độ lỗi phụ thuộc nhiều vào số lượng EQ trong bộ dữ liệu, đồng nghĩa việc metric chỉ xem xét đến độ hiệu quả của thuật toán ẩn danh dựa trên số lượng thông tin mất mát. Do đó chúng tôi đề xuất bổ sung 1 phương pháp để đánh giá bằng cách sử dụng các mô hình máy học. Dựa vào tính chất các mô hình máy học thường học được sự phân bố và ý nghĩa của dữ liệu, ta có thể áp dụng các mô hình vào bộ dữ liệu trước và sau khi ẩn danh để xem xét độ chênh lệch hiệu năng của mô hình này. Nếu độ chính xác của mô hình phân loại không giảm sau khi ẩn danh, có thể phần nào nói rằng thuật ẩn danh dữ liệu không ảnh hưởng đến chất lượng của chúng; ngược lại độ chính xác giảm đồng nghĩa các thông tin quan trọng cho việc phân loại đã bị triệt tiêu. Các mô hình chúng tôi thực hiện thử nghiệm là:

K-nearest neighbors (KNN) KNN là một phương pháp phân loại đơn giản mà không cần tham số để huấn luyện. Để xác định nhãn một mẫu dữ liệu, quyết định sẽ được đưa ra dựa trên nhãn xuất hiện nhiều nhất trong k mẫu láng giềng gần nhất của mẫu đó trong bộ dữ liệu huấn luyện. Khoảng cách giữa các mẫu được tính bằng các độ đánh giá tương đồng (ví dụ khoảng cách Euclidean, độ tương đồng Cosine, ...). Nếu dữ liệu huấn luyện đủ lớn và đủ đa dạng, kết quả của phương pháp này sẽ tương đối tốt, nhưng nhược điểm của phương pháp là nhạy cảm với số lượng láng giềng xem xét.

Support Vector Machines (SVMs) SVMs là mô hình máy học phổ biến dùng cho việc phân loại. SVMs đạt hiệu quả cao đối với dữ liệu đầu vào nhiều chiều. Với một bộ dữ liệu huấn luyện, thuật toán SVM học và xây dựng một mô hình để ánh xạ các mẫu dữ liệu mới thành các điểm trên cùng một chiều không gian sao cho khoảng cách giữa các mẫu có nhãn giống nhau thì gần nhau; ngược lại, các mẫu khác nhãn nhau càng xa nhau. Mô hình này có linh hoạt cao nhờ vào các hàm kernel (hàm tuyến tính, hàm đa thức hay hàm cơ sở xuyên tâm) được sử dụng như các hàm quyết định, việc này giúp mô hình có thể phân lớp dữ liệu phi tuyến tính tốt. Nhược điểm của mô hình là nhạy cảm với việc lựa chọn tham số cho mô hình.

Random Forests (RFs) RFs là một phương pháp học tổng hợp cho việc phân loại, hoạt động bằng cách tạo ra và huấn luyện nhiều bộ cây quyết định dựa trên bộ dữ liệu huấn luyện. Khi thực hiện dự đoán, mô hình sẽ kết hợp tất cả kết quả đầu ra của các cây quyết định để đưa ra kết quả cuối cùng. RFs khá nhạy cảm với số lượng cây quyết định và tính chất của bộ dữ liệu.

Chúng tôi cài đặt những mô hình này sử dụng thư viện SKlearn, sau đó thực hiện thử nghiệm huấn luyện các mô hình trên các bộ dữ liệu gốc và bộ dữ liệu sau khi ẩn danh. Sau khi hoàn tất huấn luyện, các mô hình được dùng để đánh giá trên bộ dữ liệu đánh giá. Các điểm metric là Accuracy và F1 sẽ được áp dụng để đánh giá hiệu suất mô hình.

5 Thí nghiệm

5.1 Thủ nghiệm thuật toán Ẩn danh

Chúng tôi tiến hành thử nghiệm 5 phương pháp ẩn danh k-anonymity (Classic Mondrian, Basic Mondrian, Datafly, Top-Down Greedy, Clustering) trên 6 bộ dữ liệu (ADULT, CAHOUSING, CMC, MGM, ITALIA, INFORMS) và quan sát các điểm metric của từng cặp. Giá trị k được chọn lần lượt trong khoảng [10,100] với bước nhảy bằng 10 đối với thử nghiệm trong ảnh 14, khoảng [2,30] bước nhảy bằng 2 đối với thử nghiệm trong ảnh 15 để quan sát khả năng của thuật toán ẩn danh với ràng buộc càng lúc càng khó hơn. Chúng tôi cũng cung cấp công khai mã nguồn sử dụng cho báo cáo này trên Github⁶

id	age	city_birth	zip_code	disease
1	55	San Giovanni	17049	Cancer
2	23	Comina	26151	AIDS
3	17	Marina Di Camerota	73015	AIDS
4	62	San Giovanni	17028	Autism
5	40	Marina Di Camerota	58014	Autism

Hình 12: Một vài mẫu dữ liệu trong tập ITALIA

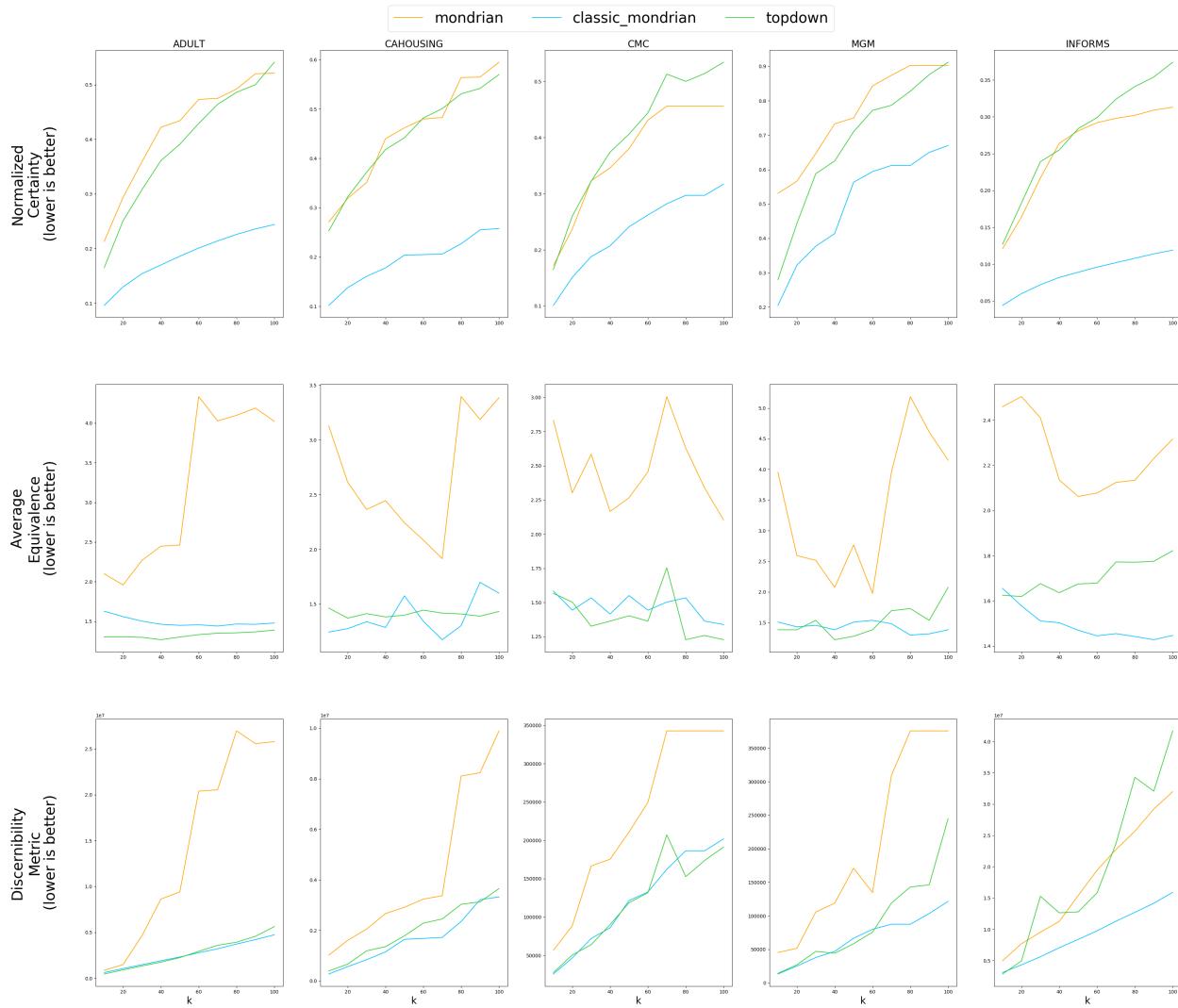
id	age	city_birth	zip_code	disease
1	50~100	San Giovanni	17***	Cancer
2	0~50	Italy	*****	AIDS
3	0~50	Italy	*****	AIDS
4	50~100	San Giovanni	17***	Autism
5	0~50	Italy	*****	Autism

Hình 13: Kết quả sau khi sử dụng thuật toán ẩn danh với $k = 2$

Hình 12 và hình 13 mô tả đơn giản một mẫu nhỏ dữ liệu trước và sau khi sử dụng thuật toán ẩn danh. Để đảm bảo dữ liệu đạt tiêu chuẩn k-anonymity, các giá trị được tổng quát hóa dựa trên cây phân tầng giá trị, trở thành các giá trị mang ý nghĩa bao quát hơn. Nhóm tiến hành đánh giá các thuật toán k-anonymity trên các bộ dữ liệu

⁶<https://github.com/kaylode/k-anonymity>

và so sánh độ hiệu quả ẩn danh của chúng với nhau. Hình 14 mô tả thử nghiệm thực tế của nhóm trên các tổ hợp các bộ dữ liệu và thuật toán.



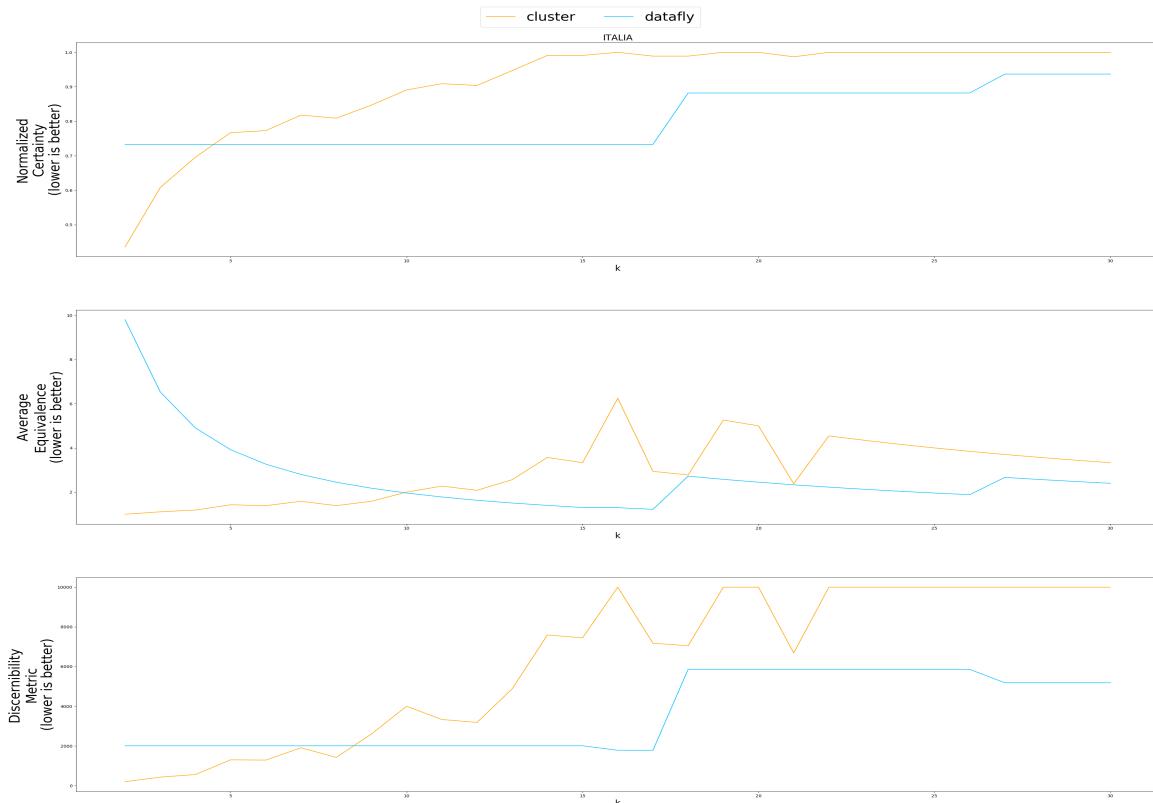
Hình 14: Điểm đánh giá các thuật toán ẩn danh với giá trị k khác nhau

Một số nhận xét có thể đưa ra như sau từ hình 14. Với metric Discernibility giá trị k càng lớn khiến điểm càng tăng do metric này phụ thuộc vào kích thước EQ. Điều tương tự cũng xảy ra với metric Average Equivalence, khi metric này phụ thuộc vào số lượng EQ. Một điểm đặc biệt quan sát được của các thuật toán ở 4 bộ dữ liệu thực tế đầu tiên là độ đắnh đổi giữa độ lỗi metric và giá trị k là rất đáng chấp nhận được, ở giá trị $k = 100$ là một giá trị

khá cao so với thực tế áp dụng, điểm đánh giá NCP không vượt quá 0.6 ở đa số các bộ cho thấy rằng lượng thuộc tính bị tổng quát hóa không quá nhiều.

Ta rút ra được rằng thuật toán Classic Mondrian cho kết quả trên các metric rất tốt so với các thuật còn lại, tuy nhiên dữ liệu đầu ra của thuật toán này không mang nhiều ý nghĩa do không tận dụng được cây phân tầng giá trị, trong khi các thuật còn lại thì sử dụng. Kết quả cho thấy 3 thuật toán này đều hoạt động khá hiệu quả với các bộ dữ liệu, trong đó Top-Down Greedy nổi bật nhất.

Dối với 2 thuật toán Datafly và Cluster được thống kê ở hình 15, do có độ phức tạp cao nên chỉ được thực nghiệm trên một bộ dữ liệu nhỏ là ITALIA. Với các giá trị k càng cao thì thuật toán áp dụng quy hoạch động Datafly đạt hiệu quả nổi bật rõ ràng hơn so với thuật toán phân cụm. Tuy nhiên giá trị NCP của cả hai đều tiệm cận nhanh đến 1 khi k đạt giá trị vượt quá 20 (với Cluster là 15), đồng nghĩa việc dữ liệu đã mất hết toàn bộ thông tin qua việc tổng quát hóa. Do đó 2 thuật toán này không phù hợp khi áp dụng vào các bài toán thực tế.



Hình 15: Điểm đánh giá các thuật toán ẩn danh với giá trị k khác nhau

5.2 Thủ nghiệm các mô hình máy học

Để sử dụng hai bộ [12](#) và [13](#) dựa vào các mô hình máy học yêu cầu việc tiền xử lý, số hóa các giá trị. Đối với các thuộc tính liên tục, ta có thể giữ nguyên để đưa vào mô hình, tuy nhiên để xử lý giá trị liên tục đã bị tổng quát hóa, nhóm lấy giá trị trung vị của khoảng này, ví dụ từ 50-100 lấy trung vị là 75 xem ở [hình 17](#). Đối với các giá trị là thuộc tính rời rạc, nhóm tiến hành one-hot encoding thuộc tính này. Giả sử ở [bảng 16](#), miền giá trị của thuộc tính city_birth là (San Giovanni, Comina, Marina Di Camerota), khi one-hot encoding sẽ thành 3 thuộc tính riêng lẻ (city_birth_SanGiovanni, city_birth_Comina, city_birth_MarinaDiCamerota), với giá trị 0 hoặc 1. Để qui đổi giá trị thuộc tính rời rạc tổng quát từ [bảng 13](#) để đưa vào mô hình, cách tương tự cũng được áp dụng kết hợp thêm thông tin từ cây phân tầng giá trị. Cụ thể hơn, các giá trị ở node có mức tổng quát cao sẽ phân tách thành các node lá của nó, chẳng hạn ở [bảng 13](#) giá trị Italy sẽ phân tích thành tất cả các nút lá của nó (city_birth_SanGiovanni, city_birth_Comina, city_birth_MarinaDiCamerota) như thấy ở [bảng 17](#).

id	age	city_birth_San Giovanni	city_birth_Comina (La)	city_birth_Marina Di Camerota	zip_code	disease
1	55	1	0	0	17049	Cancer
2	23	0	1	0	26151	AIDS
3	17	0	0	1	73015	AIDS
4	62	1	0	0	17028	Autism
5	40	0	0	1	58014	Autism

Hình 16: Dữ liệu đầu vào của các mô hình máy học sau khi one-hot encode từ [hình 12](#)

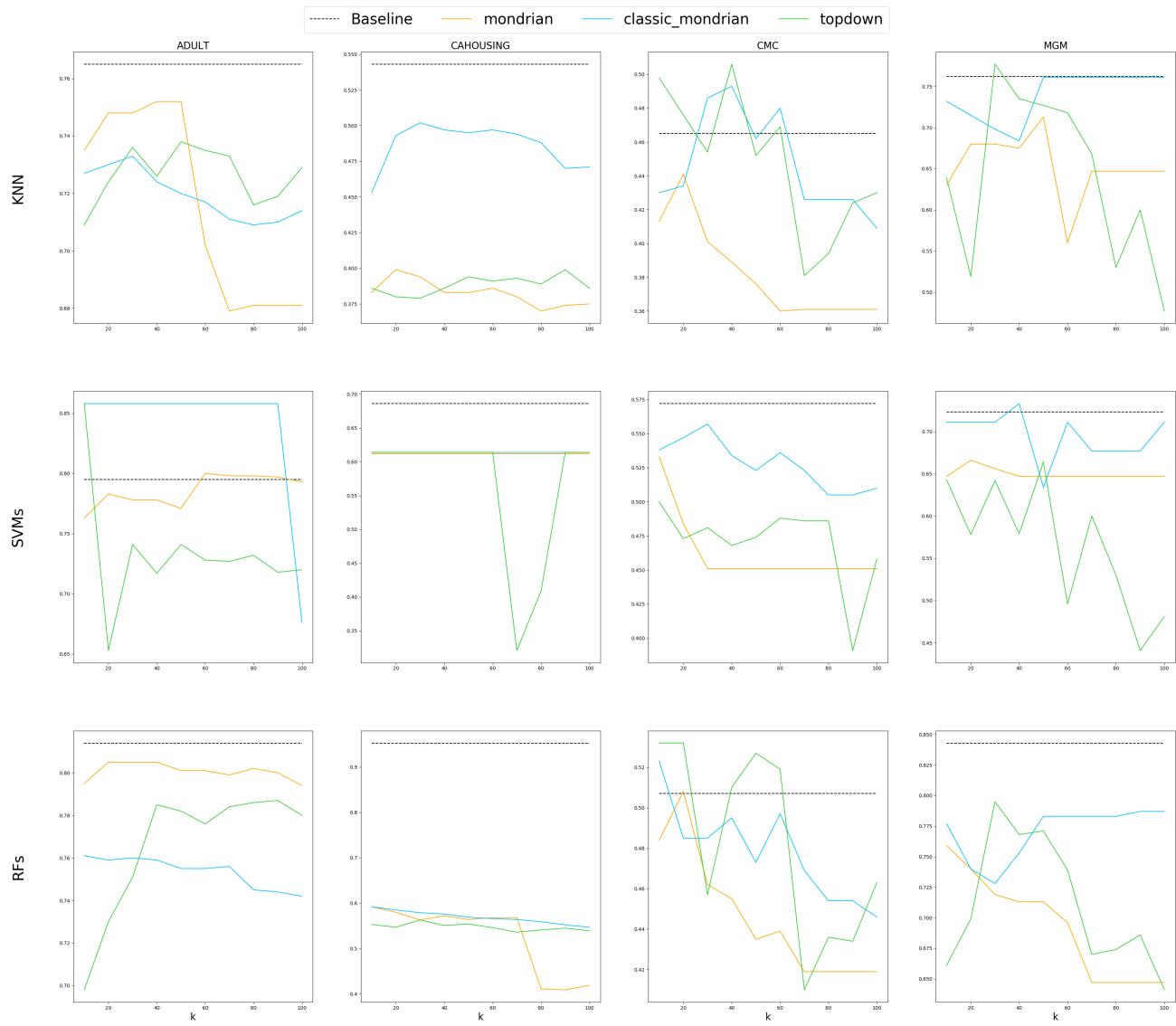
id	age	city_birth_San Giovanni	city_birth_Co mina (La)	city_birth_Marina Di Camerota	zip_code	disease
1	75	1	0	0	17499.5	Cancer
2	25	1	1	1	49999.5	AIDS
3	25	1	1	1	49999.5	AIDS
4	75	1	0	0	17499.5	Autism
5	25	1	1	1	49999.5	Autism

Hình 17: Dữ liệu đầu vào của các mô hình máy học sau khi quy đổi các giá trị tổng quát hóa từ hình 13

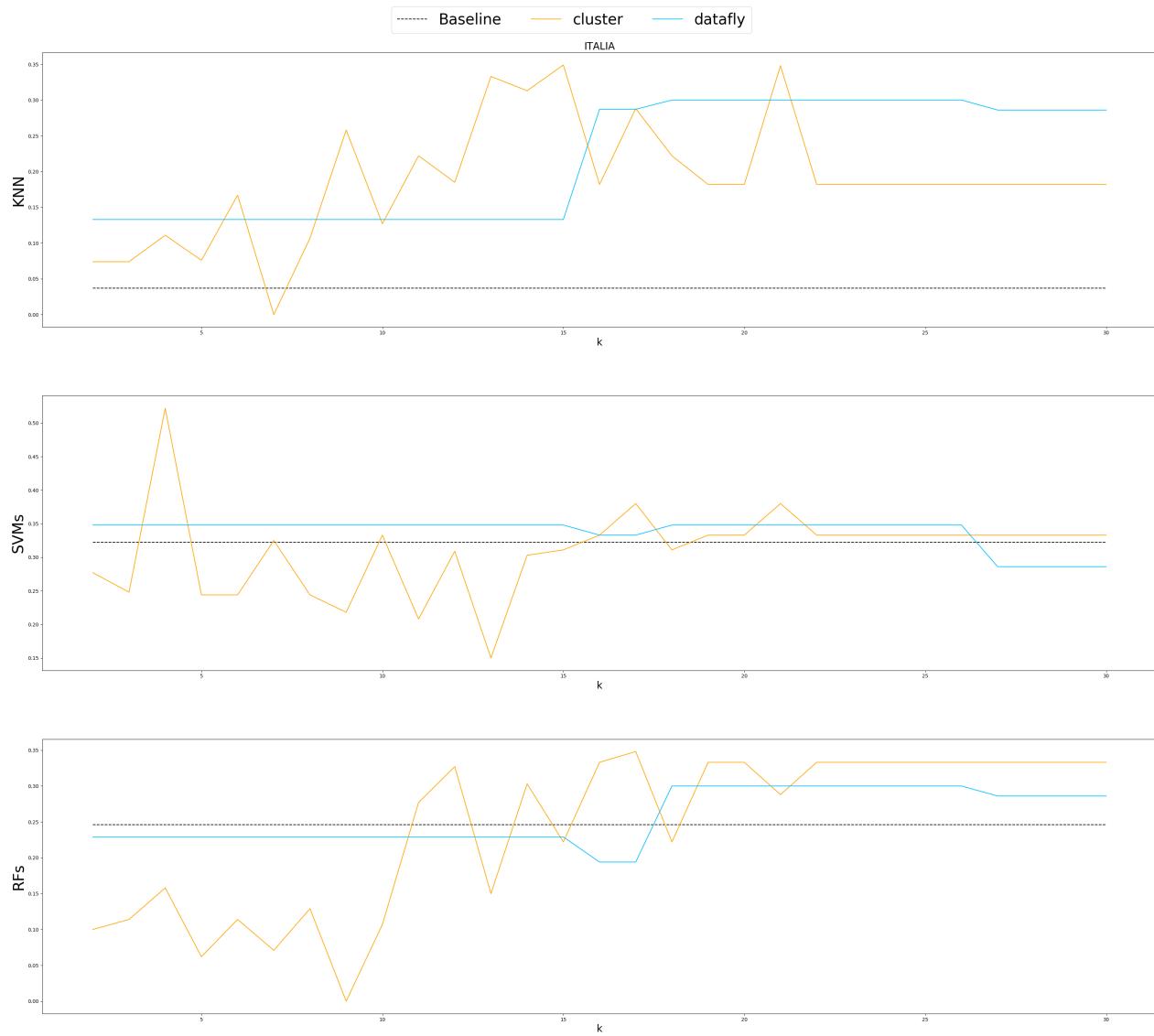
Có nhiều cách đơn giản hơn để quy đổi từ giá trị tổng quát sang dạng one-hot encoding, như trong nghiên cứu của Slijepčević et. al [21] là nghiên cứu đã truyền cảm hứng báo cáo này, họ thực hiện việc này bằng cách one-hot encoding trực tiếp trên các giá trị tổng quát hóa, nghĩa là ví dụ trên sẽ là city_birth_Italy như hình 19. Chúng tôi thấy rằng phương pháp này sẽ dẫn đến hình thành một chiều không gian rất lớn với các giá trị one-hot encode (curse of dimensionality) ngăn cản mô hình phân lớp học tốt. Nhóm nhận thấy rằng các này không thực sự mô tả đúng sự tương đồng giữa hai bảng dữ liệu trước và sau khi ẩn danh. Do đó không phù hợp để đánh giá mức độ ảnh hưởng của việc ẩn danh đối với hiệu suất mô hình máy học.

id	age	city_birth_San Giovanni	city_birth_Italy	zip_code	disease
1	75	1	0	17499.5	Cancer
2	25	0	1	49999.5	AIDS
3	25	0	1	49999.5	AIDS
4	75	1	0	17499.5	Autism
5	25	0	1	49999.5	Autism

Hình 18: Dữ liệu đầu vào của các mô hình máy học sử dụng one-hot encoding phiên bản của Slijepčević



Hình 19: F1-Score của các mô hình máy học trên các bộ dữ liệu, trong đó các đường có nhãn baseline nghĩa là hiệu suất mô hình dự đoán trên dữ liệu gốc, các đường còn lại là dữ liệu bị ẩn danh bởi các thuật toán tương ứng



Hình 20: F1-Score của các mô hình máy học trên các bộ dữ liệu, trong đó các đường có nhãn baseline nghĩa là hiệu suất mô hình dự đoán trên dữ liệu gốc, các đường còn lại là dữ liệu bị ẩn danh bởi các thuật toán tương ứng

Có thể quan sát ở hình 19, trong đa số trường hợp, các mô hình đều giảm độ hiệu quả của chúng so với trước khi áp dụng ẩn danh dữ liệu, điều này cũng là giả định lý thuyết ban đầu của nhóm. Song, vẫn tồn tại một số trường hợp nhóm gọi là nhiều. Những trường hợp này có thể thấy là khi giá trị k tăng, hiệu suất của mô hình cũng

tăng dần so với giá trị k trước đó và có khi vượt qua cả ngưỡng của mô hình baseline. Đây là một phát hiện thú vị rằng trong một số trường hợp hiếm, dữ liệu bị ẩn danh cho kết quả cao hơn dữ liệu gốc khi được phân loại bằng mô hình, một số trường hợp khác có hiện tượng giá trị k đồng biến với hiệu suất phân loại. Chúng tôi phân tích và có gắng tìm ra lời giải thích cho các trường hợp này, điều này có thể xảy ra với nguyên nhân là do khi cây phân tầng của các bộ dữ liệu hoạt động như một metadata giúp thuật toán ẩn danh đồng bộ được với cấu trúc các lớp của bộ dữ liệu, do đó giúp mô hình phân loại chia cắt tốt hơn các lớp của bộ dữ liệu. Tuy nhiên nếu nhìn nhận một cách tổng quát thì giá trị k càng lớn hơn thì hiệu suất giảm rõ rệt. Điều tương tự cũng xảy ra đối với 2 thuật toán Cluster-based và Datafly trên bộ dữ liệu ITALIA, tuy nhiên bộ dữ liệu này có kích thước khá nhỏ nên không thể rút ra được kết luận chắc chắn. Qua các thử nghiệm này, nhóm nhận xét rằng thuật toán Basic Mondrian và Top-Down Greedy hoạt động tương đối hiệu quả giống nhau và tốt hơn các thuật còn lại. Nói tóm lại, khi k càng tăng, tức dữ liệu càng bị ràng buộc chặt chẽ hơn, thông tin bị mất mát đi nhiều hơn về mặt ngữ nghĩa sẽ khiến khả năng học đặc trưng phân lớp của mô hình máy học giảm, điều này phù hợp với lý thuyết mà chúng tôi giả định ban đầu.

Cuối cùng, chúng tôi muốn đưa ra một số kết luận về tác động qua lại giữa phương pháp ẩn danh và mô hình máy học. Chúng tôi đã chỉ ra rằng việc ẩn danh và mô hình máy học phụ thuộc nhiều vào nhau. Cả hai cùng cố gắng giải quyết các mục tiêu khác nhau và có phần mâu thuẫn nhau, khiến cho sự kết hợp giữa chúng trở nên phức tạp. Dữ liệu bị ẩn danh có thể bị mất mát thông tin quan trọng cần để giải quyết bài toán phân loại ra khỏi tập dữ liệu và do đó có thể làm giảm hiệu suất phân loại của mô hình.

Bằng việc so sánh nhiều mô hình phân loại khác nhau, nghiên cứu cung cấp một cái nhìn tổng thể về sự nhạy cảm của việc thông tin bị thay đổi (bằng cách tổng quát hóa hoặc loại bỏ) có thể ảnh hưởng đến chất lượng dữ liệu.

Nghiên cứu cũng chỉ ra được rằng việc lựa chọn các thuộc tính để tiến hành tổng quát hóa đóng một phần quan trọng đối với hiệu quả của việc phân loại. Nếu những giá trị nhãn (target variable) có độ tương quan cao với các QID bị tổng quát hóa mạnh thì sẽ gây ảnh hưởng đến khả năng học của mô hình huấn luyện. Với mỗi bộ dữ liệu khác nhau tồn tại nhiều cách để đạt được k-anonymity khác nhau. Trong những trường hợp đó, thuật toán nên được chọn là thuật toán mang lại sự tổng quát hóa ít đối với các thuộc tính tương quan với nhãn.

Theo quan sát từ các thử nghiệm trên, có thể nhận xét rằng phương pháp ẩn danh Basic Mondrian và Top-Down Greedy là một trong các phương pháp tốt trong việc giữ lại các thông tin về mặt ý nghĩa đối với bộ dữ liệu, trong khi Top-Down vượt trội hơn về mặt ít mất mát thông tin. Chúng tôi cũng nhận thấy rằng việc lựa chọn giá trị k cũng quan trọng, việc sử dụng $k = 100$ như thử nghiệm (giá trị này cao hơn nhiều so với giá trị được áp dụng trong thực tế) vẫn cho kết quả đánh đổi về hiệu suất mô hình trong mức tương đối chấp nhận được. Tuy nhiên trong thực tế các bộ dữ liệu có độ lớn lớn hơn rất nhiều so với những bộ mà nhóm thử nghiệm. Do đó cần phải thử nghiệm nhiều thuật toán với bộ dữ liệu lớn hơn để có thể đưa ra những đánh giá khách quan hơn.

6 Tổng kết

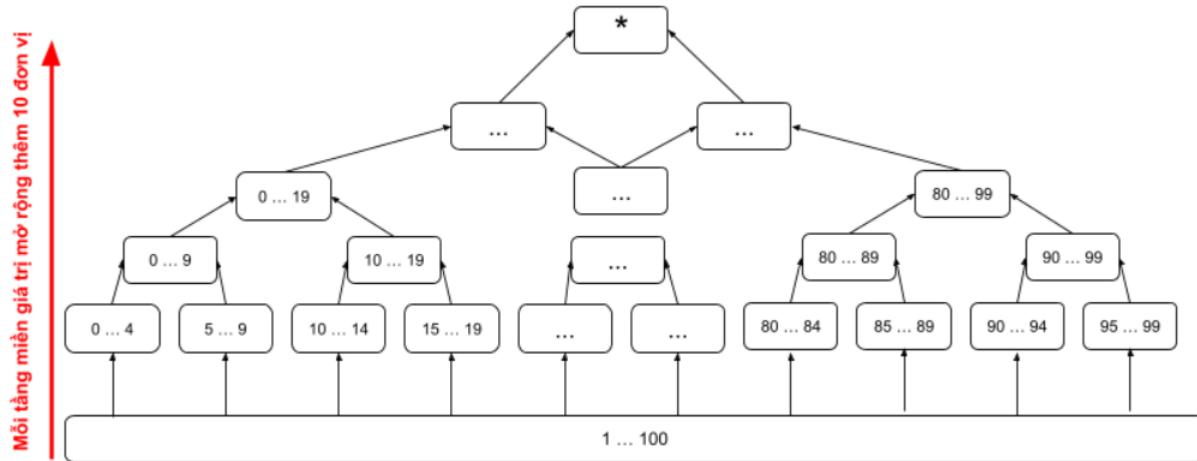
Trong thực tế, bảo mật thông tin mang một ý nghĩa quan trọng và là một mối quan tâm lớn đối với các bài toán liên quan đến dữ liệu. Thông tin một cá nhân có thể bị lộ bởi các hình thức tấn công danh tính nếu không được bảo vệ đúng cách. Do đó sự tồn tại các thuật toán ẩn danh dữ liệu giúp tăng cường an ninh cho các cá nhân tham gia vào bộ dữ liệu, đồng thời phải đảm bảo rằng việc mất mát thông tin là tối thiểu tức dữ liệu vẫn còn tính tiện dụng. Ngoài ra, với sự phát triển của các mô hình máy học ngày nay, vấn đề này càng được chú ý hơn bao giờ hết, lý do rằng dữ liệu đóng một vai trò quan trọng đối với hiệu suất của các mô hình. Qua nghiên cứu này, chúng tôi tập trung vào việc thực hiện so sánh sự hiệu quả về mặt mất mát thông tin giữa các thuật toán k-anonymity phổ biến. Đồng thời phân tích mối liên hệ qua lại giữa mức độ bảo mật thông tin và hiệu quả của mô hình máy học, từ đó đánh giá khả năng trade-off giữa độ tiện dụng và bảo mật của dữ liệu.

Tài liệu

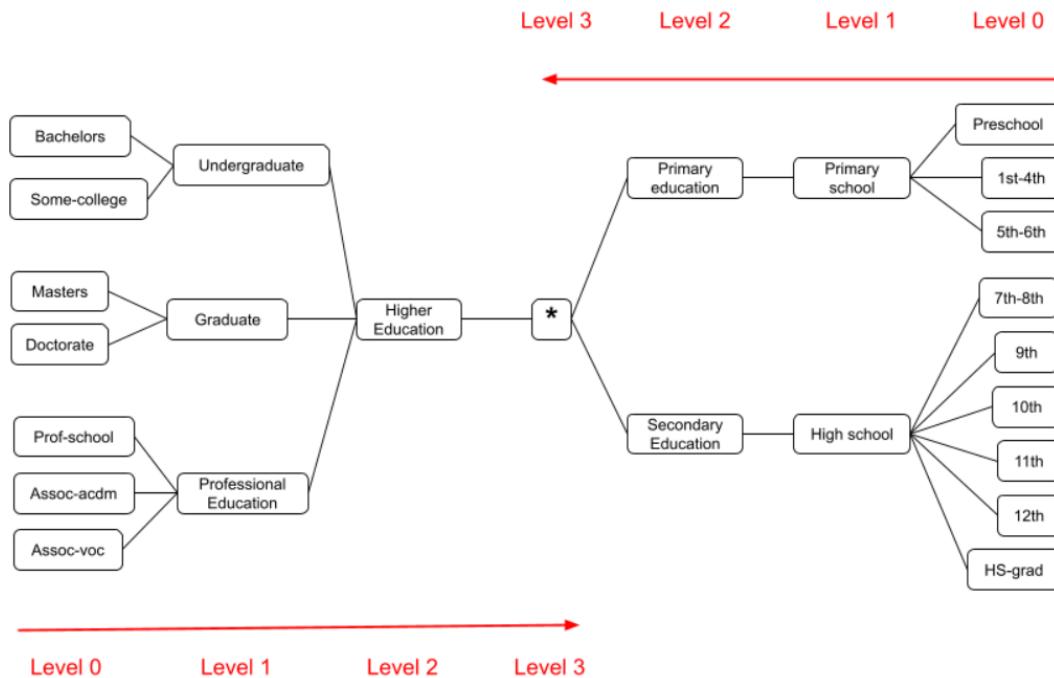
- [1] And if you liked the movie, a netflix contest may reward you handsomely. <https://www.nytimes.com/2006/10/02/technology/02netflix.html>.
- [2] Apple delays privacy change amid app publishers concerns. <https://www.wsj.com/articles/apple-delays-privacy-change-amid-app-publishers-concerns-11599164713>.
- [3] Differential privacy overview. https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf.
- [4] Federated learning collaborative. <https://ai.googleblog.com/2017/04/federated-learning-collaborative.html>.
- [5] How target figured out a teen girl was pregnant before her father did. <https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/?sh=1a16b13b6668>.
- [6] KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2006. Association for Computing Machinery.
- [7] Vanessa Ayala-Rivera, Patrick McDonagh, Thomas Cerqueus, and Liam Murphy. A systematic comparison and evaluation of k-anonymization algorithms for practitioners. *Trans. Data Privacy*, 7(3):337–370, December 2014.
- [8] Dwork C. Differential privacy: A survey of results. 2008.
- [9] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. 9(3–4):211–407, August 2014.
- [10] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor. *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, Nov 2014.
- [11] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*, pages 25–25, 2006.
- [12] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, SIGMOD '05, page 49–60, New York, NY, USA, 2005. Association for Computing Machinery.
- [13] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115, 2007.

- [14] Jun-Lin Lin and Meng-Cheng Wei. An efficient clustering method for k-anonymization. In *Proceedings of the 2008 International Workshop on Privacy and Anonymity in Information Society*, PAIS '08, page 46–50, New York, NY, USA, 2008. Association for Computing Machinery.
- [15] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. ℓ -diversity: Privacy beyond k -anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1):3–es, March 2007.
- [16] Microsoft. Microsoft - differential privacy for everyone, 2020.
- [17] Arvind Narayanan and Vitaly Shmatikov. How to break anonymity of the netflix prize dataset. *arXiv preprint cs/0610105*, 2006.
- [18] Kobbi Nissim. Differential privacy: A primer for a non-technical audience, 2018.
- [19] Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, 1998.
- [20] Warner SL. Randomized response: a survey technique for eliminating evasive answer bias. 1965.
- [21] Djordje Slijepčević, Maximilian Henzl, Lukas Daniel Klausner, Tobias Dam, Peter Kieseberg, and Matthias Zeppelzauer. k -anonymity in practice: How generalisation and suppression affect machine learning classifiers, 2021.
- [22] L. Sweeney. Datafly: A system for providing anonymity in medical data. In *DBSec*, 1997.

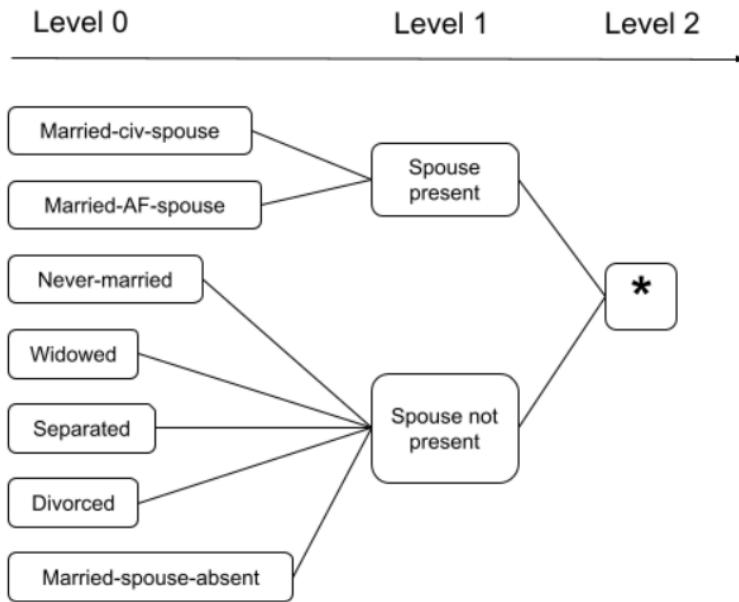
7 Phụ lục



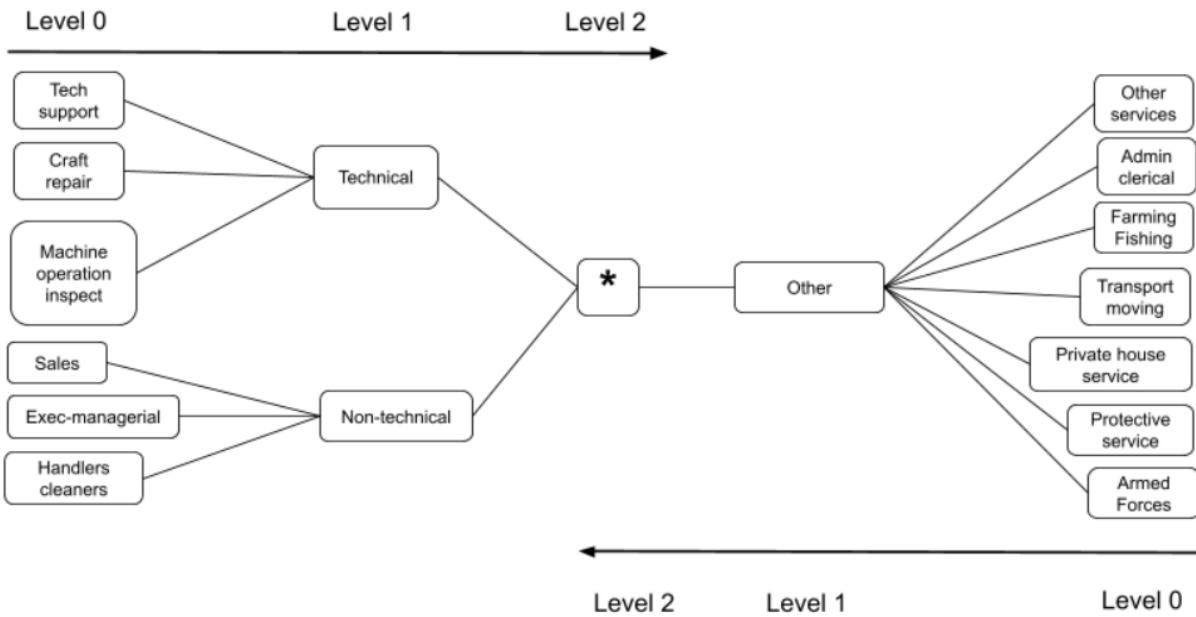
Hình 21: Cây phân tầng tuổi của dân số trong bộ Adult



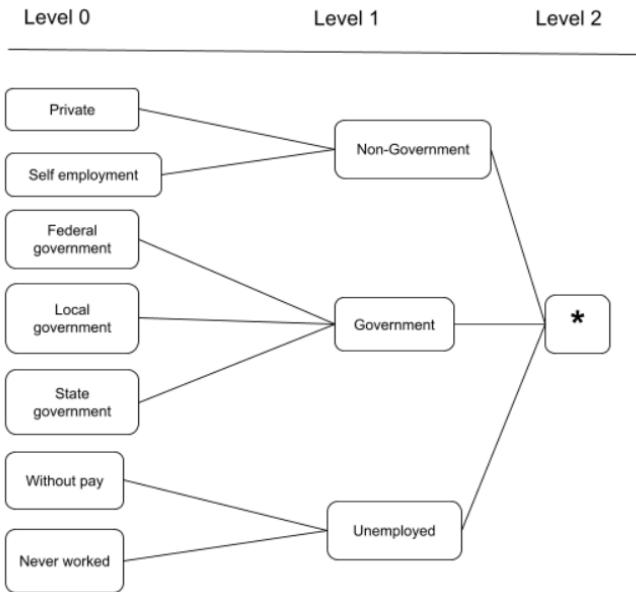
Hình 22: Cây phân tầng trình độ học vấn của dân số trong bộ Adult



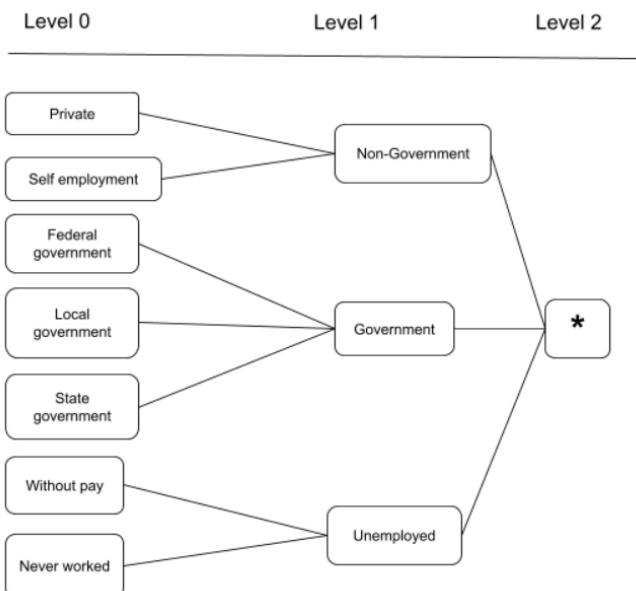
Hình 23: Cây phân tầng tình trạng hôn nhân của dân số trong bộ Adult



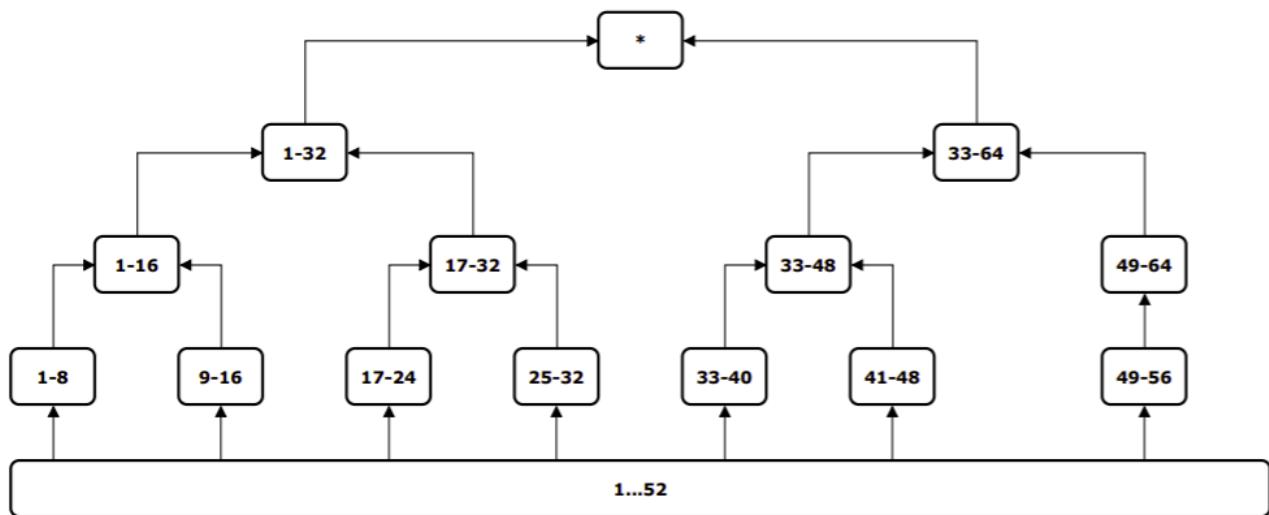
Hình 24: Cây phân tầng nghề nghiệp của dân số trong bộ Adult



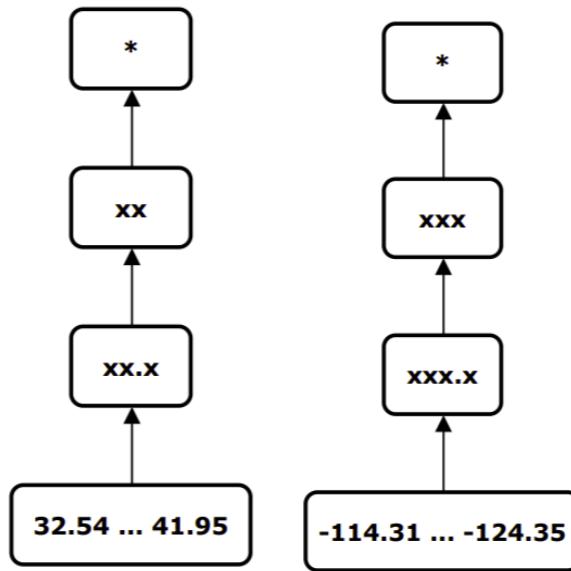
Hình 25: Cây phân tầng lớp của dân số trong bộ Adult



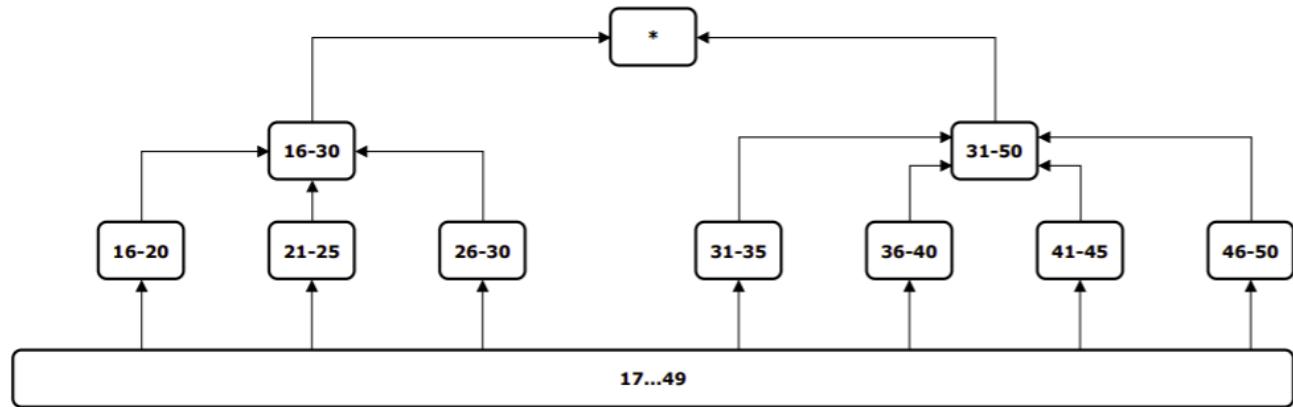
Hình 26: Cây phân tầng giá trị của ngôi nhà trong bộ California Housing Prices



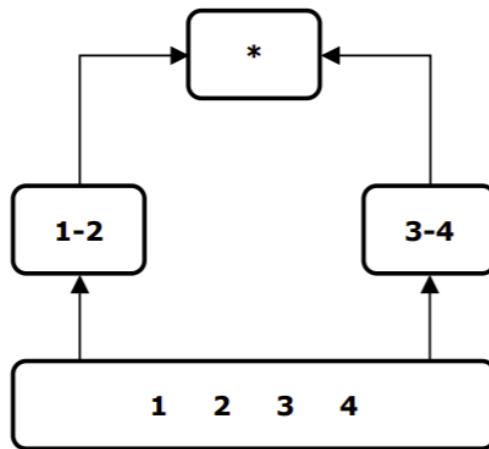
Hình 27: Cây phân tầng thu nhập của chủ nhà trong bộ California Housing Prices



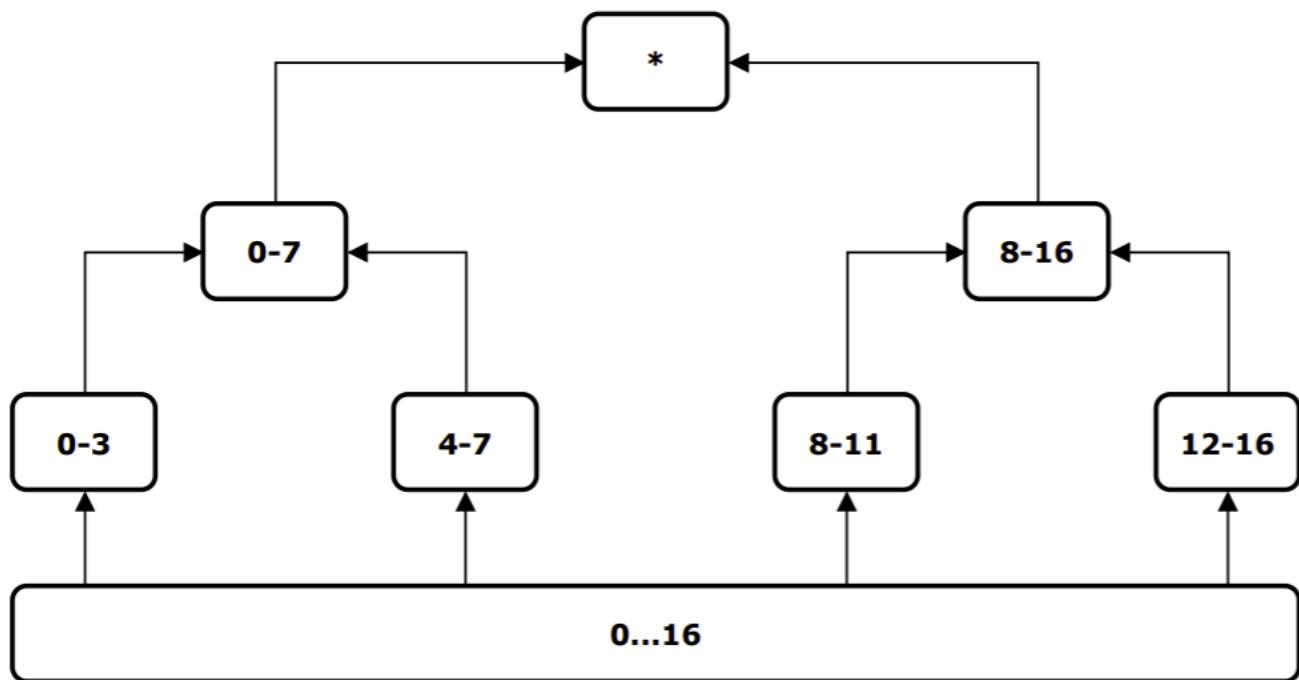
Hình 28: Cây phân tầng kinh độ, vĩ độ của ngôi nhà trong bộ California Housing Prices



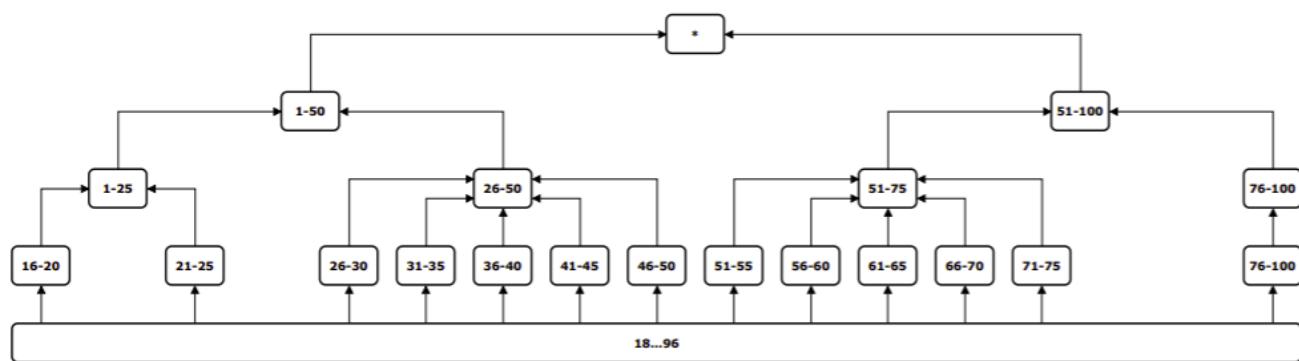
Hình 29: Cây phân tầng tuổi của người vợ trong bộ Contraceptive Method Choice



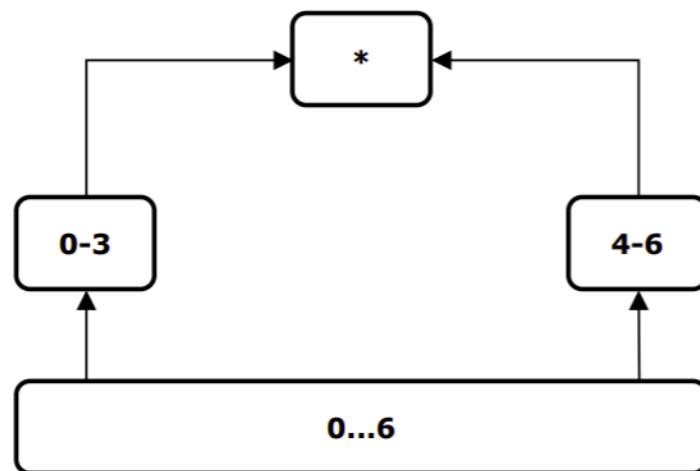
Hình 30: Cây phân tầng trình độ học vấn của người vợ trong bộ Contraceptive Method Choice



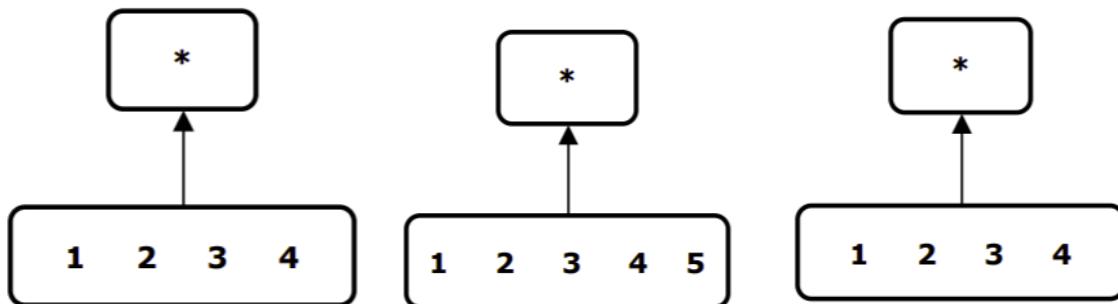
Hình 31: Cây phân tầng số lượng con của người vợ trong bộ Contraceptive Method Choice



Hình 32: Cây phân tầng tuổi của bệnh nhân trong bộ Mammographic Mass



Hình 33: Cây phân tầng điểm đánh giá của bệnh nhân trong bộ Mammographic Mass



Hình 34: Cây phân tầng mật độ, biên độ, hình dạng của khối u của bệnh nhân trong bộ Mammographic Mass