

Evaluation of Anonymization Techniques Through Privacy-Utility Trade-off Analysis

Ninawa Ismaeel* 

Dr. Basel Hasan**

(Received 28 / 4 / 2026. Accepted 23 / 7 / 2026)

□ ABSTRACT □

Balancing data privacy and utility remains a fundamental challenge in privacy-preserving data publishing. Although numerous anonymization techniques have been proposed, many existing studies rely on theoretical comparisons or limited experimental evaluations, often lacking a unified framework for jointly assessing privacy and utility.

This paper presents a systematic experimental evaluation of four widely used anonymization models: k -anonymity, ℓ -diversity, t -closeness, and k -map. The evaluation is conducted using ARX tool on real-world Adult dataset. Furthermore, k -anonymity and ℓ -diversity models are evaluated on healthcare dataset (Heart Disease). Privacy and utility metrics are analyzed within a unified methodological framework. To enhance evaluation consistency and interpretability, we propose a Unified Utility Measure (UUM) that integrates several utility metrics into a single aggregated score. This approach reduces inconsistencies from separate metrics and provides a comprehensive assessment of data utility.

The results reveal a clear inverse relationship between privacy and utility, where stricter privacy requirements significantly degrade data utility. Furthermore, optimal parameters are empirically identified for each model to achieve an effective trade-off. Finally, practical guidelines are provided to support informed decision-making regarding anonymization.

Keywords: Anonymization, Privacy, Data Utility, Privacy, Preserving Techniques.

Copyright



:Latakia University journal (formerly Tishreen) -Syria, The authors retain the copyright under a CC BY-NC-SA 04

* Master's Student, Department of Software Engineering and Information Systems, Faculty of Informatics Engineering, Latakia University (formerly Tishreen), Latakia, Syria.

Email: ninawa.ismaeel@latakia-univ.edu.sy

** Assistant Professor, Department of Software Engineering and Information Systems, Faculty of Informatics Engineering, Latakia University (formerly Tishreen), Latakia, Syria.

Email: basel.hasan@latakia-univ.edu.sy

تقييم تقنيات إخفاء الهوية من خلال تحليل المقايضة بين الخصوصية والفائدة

نينوى اسماعيل *

د. باسل حسن **

(تاريخ الإيداع 28 / 4 / 2026. قُبل للنشر في 23 / 7 / 2026)

□ ملخص □

لا يزال تحقيق التوازن بين خصوصية البيانات وفائدتها يمثل تحدياً أساسياً في نشر البيانات مع الحفاظ على الخصوصية. وعلى الرغم من اقتراح العديد من تقنيات إخفاء الهوية، فإن الكثير من الدراسات الحالية تعتمد على مقارنات نظرية أو تقييمات تجريبية محدودة، وغالباً ما تقتصر لإطار موحد لتقييم الخصوصية والفائدة معاً. تقدم هذه الورقة البحثية تقييماً تجريبياً منهجياً لأربعة من نماذج إخفاء الهوية واسعة الاستخدام: k -anonymity، ℓ -diversity، k -map و t -closeness. تم إجراء التقييم باستخدام أداة ARX على مجموعة بيانات Adult من العالم الحقيقي، علاوة على ذلك، جرى تقييم نموذجي k -anonymity و ℓ -diversity على مجموعة بيانات الرعاية الصحية (Heart Disease).

تم تحليل مقاييس الخصوصية والفائدة ضمن إطار منهجي موحد. ولتعزيز اتساق التقييم وقابليته للتفسير، نقترح مقياس فائدة موحد (UUM) يدمج عدة مقاييس للفائدة في درجة مجمعة واحدة. يقلل هذا النهج من التناقضات المرتبطة بالمقاييس المنفصلة ويقدم تقييماً أكثر شمولية لفائدة البيانات.

تكشف النتائج عن علاقة عكسية واضحة بين الخصوصية والفائدة، حيث تؤدي متطلبات الخصوصية الأكثر صرامة إلى تدهور كبير في فائدة البيانات. كذلك، تم تحديد المعلمات المثلى تجريبياً لكل نموذج لتحقيق توازن فعال. وأخيراً، تم تقديم إرشادات عملية لدعم اتخاذ قرارات مستنيرة بشأن إخفاء الهوية.

الكلمات المفتاحية: إخفاء الهوية، الخصوصية، فائدة البيانات، تقنيات الحفاظ على الخصوصية.



حقوق النشر : مجلة جامعة اللاذقية (تشرين سابقاً) - سورية، يحتفظ المؤلفون بحقوق النشر

بموجب الترخيص CC BY-NC-SA 04

* طالبة ماجستير، قسم البرمجيات ونظم المعلومات، كلية الهندسة المعلوماتية، جامعة اللاذقية (تشرين سابقاً)، اللاذقية، سورية.

Email: ninawa.ismaeel@latakia-univ.edu.sy

** مدرس، قسم البرمجيات ونظم المعلومات، كلية الهندسة المعلوماتية، جامعة اللاذقية (تشرين سابقاً)، اللاذقية، سورية.

Email: basel.hasan@latakia-univ.edu.sy

1. مقدمة:

في السنوات الأخيرة، تجاوز النمو الهائل للمعلومات الرقمية التوقعات مع التطور التكنولوجي، فقد وصلت كمية البيانات التي يتم جمعها ومشاركتها إلى مستويات مذهلة، حيث توفر سجلات مفصلة عن الأفراد التي تُستخدم في تطبيقات العالم الحقيقي [1]. وقد تحتوي هذه البيانات الشخصية على معلومات تتيح التعرف على الفرد، وقد يؤدي نشرها للعامة إلى انتهاك الخصوصية [2].

تُعد الخصوصية حقاً أساسياً من حقوق الإنسان. والتي تمثل قدرة الفرد على تحديد بياناته التي يمكن مشاركتها، وتطبيق ضوابط الوصول إليها [3]. ومع التقدم السريع للتكنولوجيا، أصبح ضمانها أكثر تعقيداً [4]. ونتيجة لذلك، أثارت الخصوصية العديد من الإشكالات القانونية حول العالم، مما زاد اهتمام وقلق الأفراد حول كيفية تعامل المنظمات والمؤسسات مع بياناتهم ومعلوماتهم الخاصة. ومؤخراً، وضعت عدة تشريعات لحماية خصوصية أصحاب البيانات، وقد ظهرت اللائحة العامة لحماية البيانات (GDPR) General Data Protection Regulation لتوحيد قوانين الخصوصية في أوروبا [1]. وعقب صدور هذه اللائحة، تم اتخاذ عدة جهود تشريعية مماثلة، مثل قانون خصوصية المستهلك في كاليفورنيا (CCPA) California Consumer Privacy Act لعام 2018، والذي يمنح سكان كاليفورنيا الحق في معرفة البيانات الشخصية التي يتم جمعها عنهم وطلب حذفها [5]. وبناءً على ذلك، كان أحد الأسئلة التي دارت حولها الكثير من النقاشات هو كيفية الحفاظ على خصوصية البيانات مع الإبقاء عليها صالحة للتحليل [2]. وعلى الرغم من أن الحل الأمثل يتمثل في بيانات مُخفاة الهوية تحقق أقصى درجات الخصوصية والفائدة، فإن هذا السيناريو يكاد يكون مستحيلاً تحقيقه عملياً [1].

في مواجهة هذا التحدي المزدوج، تم تطوير العديد من تقنيات الحفاظ على الخصوصية، لكن معظمها يعتمد على إخفاء الهوية (Anonymization) [3]، وهو الحل العملي والأكثر استخداماً لحماية خصوصية المستخدم أثناء نشر البيانات [6]. ولا يقتصر دور إخفاء هوية البيانات على الامتثال لقوانين الخصوصية مثل GDPR و CCPA فحسب؛ بل إنه يجسد أيضاً التزاماً بالممارسات الأخلاقية في التعامل مع البيانات، حيث إن هدفه الأساسي هو تعديل البيانات الشخصية بطريقة لا يمكن من خلالها تحديد هوية الأفراد، بشكل مباشر أو غير مباشر، مع الحفاظ على فائدة البيانات للتحليل واتخاذ القرار [4].

تواجه الأساليب الأخرى العديد من القيود، فالخصوصية التفاضلية تؤدي إلى تدهور ملحوظ في فائدة البيانات المُخفاة الهوية [6]، كما أن دلالاتها معقدة، حيث يصعب شرحها لصناع القرار، مثل لجان الأخلاقيات وصناع السياسات [7]. وتتمثل إحدى العيوب الرئيسية للتشفير في الصعوبة اللاحقة في معالجة البيانات وتحليلها وتفسير النتائج [1].

قد اقترح الباحثون عدداً من تقنيات إخفاء الهوية [8]، أكثرها شيوعاً: k -anonymity، ℓ -diversity، k -map، و t -closeness، لذا يقيم هذا البحث الأثر الفعلي لهذه التقنيات، بالتركيز على خطرين لإعادة تحديد الهوية (Re-identification): خطر المدعي العام (Prosecutor Risk) الذي يفترض توفر معرفة مسبقة بوجود الفرد في مجموعة البيانات، وخطر الصحفي (Journalist Risk) الذي يفترض غياب هذه المعرفة، بالإضافة إلى محاكاة سيناريو الكشف عن الهوية (Identity Disclosure) الذي يحدث عندما يتمكن المهاجم من ربط السجل بفرد معين عن طريق مطابقة قيم شبه المعرفات (Quasi-Identifiers)، وسيناريو الكشف عن

السمات (Attribute Disclosure) عندما يتمكن المهاجم من استنتاج خصائص جديدة عن الفرد من البيانات المنشورة [1]. وبناءً على ذلك، يعتمد هذا البحث معدل خطر المدعي العام معياراً أساسياً لتقييم الخصوصية.

2- أهمية البحث وأهدافه:

إن العثور على أفضل مقايضة بين الخصوصية والفائدة لا يزال يشكل تحدياً. فمن ناحية، البيانات ذات الجودة المنخفضة والتي تتمتع بمستوى خصوصية عالٍ تجعل تفسير النتائج واستخلاص المعرفة أقل فعالية، ومن ناحية أخرى، قد يؤدي مستوى الخصوصية المنخفض إلى إعادة تحديد الهوية [1]. في ظل هذا التحدي، تتساءل المؤسسات عن الخيار الأمثل الذي يحقق أعلى قدر من الخصوصية بأقل تضحية ممكنة في فائدة البيانات.

تنتم الأبحاث الحالية في هذا المجال بغلبة الدراسات النظرية التي تقارن بين التقنيات دون تطبيق تجريبي شامل. لسد هذه الفجوة، يقدم هذا البحث تقيماً تجريبياً منهجياً لتقنيات إخفاء الهوية بهدف:

- تحديد العلاقة بين خصوصية البيانات وفائدتها عند تطبيق تقنيات إخفاء الهوية.
- اقتراح مقياس فائدة موحد يجمع بين عدة مقاييس لتقييم شامل وموضوعي لأثر إخفاء الهوية على فائدة البيانات.
- تحديد القيم المثلى لمعاملات الإخفاء التي تحقق أفضل مقايضة (Optimal Trade-off) بين الخصوصية وفائدة البيانات ضمن الإعدادات التجريبية الخاصة بهذه الدراسة.
- تقديم توصيات مدعمة بنتائج تجريبية باستخدام أداة ARX، تعزز أفضل الممارسات عند التعامل مع البيانات الحساسة، وتقدم دليلاً واضحاً لصناع القرار والمحللين لتقييم الأثر الحقيقي لخيارات إخفاء الهوية على قابلية استخدام البيانات.

3- طرائق البحث ومواده:

تم استخدام مجموعة بيانات من العالم الحقيقي Adult dataset [9]، والتي تمثل بيانات التعداد السكاني الأمريكي لعام 1994، وتُعد المعيار المرجعي الفعلي لتقييم خوارزميات إخفاء الهوية [10]. تتضمن المجموعة الأصلية 48842 سجلاً و15 سمة (عموداً). شملت المعالجة المسبقة إزالة القيم المفقودة ومعالجة القيم غير الصالحة، مما أدى للإبقاء على 45222 سجلاً صالحاً. تم اختيار 9 سمات لإجراء التجارب، والتي استخدمت بشكل شائع في العديد من الدراسات السابقة مثل: [10] وذلك من أجل تقليل الأبعاد (عدد السمات) التي قد تؤدي زيادتها إلى تعميم مفرط وفقدان غير مبرر لفائدة البيانات التحليلية، مع التركيز على السمات الديموغرافية الأكثر ارتباطاً بإخفاء الهوية.

ولدراسة أثر تغير خصائص مجموعة البيانات على تقييم النماذج، تم استخدام مجموعة البيانات الصحية المعروفة Heart Disease Dataset [11]، التي تخص مرضى القلب، حيث تحتوي على 297 سجلاً و14 سمة، وذلك بعد إجراء المعالجة المسبقة والتي شملت إزالة القيم المفقودة.

استُخدمت أداة ARX [10] لتنفيذ التجارب، وهي إطار عمل شامل مفتوح المصدر تستخدم على نطاق واسع في تطبيق تقنيات إخفاء الهوية وتقييم تأثيرها على الخصوصية والفائدة. بعد بناء التسلسلات الهرمية للتعميم لشبه المعارف تم

تطبيق التقنيات المختارة باستخدام قيم مختلفة لمعاملاتها، ثم تحليل النتائج وتقييمها باستخدام مجموعة من مقاييس خصوصية البيانات وفائدتها، ومقارنة هذه النتائج فيما بينها للوصول لأفضل مقايضة، ونهي البحث بتقديم التوصيات اعتماداً على النتائج المستخلصة.

الخلفية النظرية والمفاهيم الأساسية

3-1- أنواع السمات وكيفية التعامل معها في إخفاء الهوية

يُعد تصنيف السمات وفقاً لدورها في إعادة تحديد الهوية خطوة أساسية لاختيار تقنيات إخفاء الهوية المناسبة، حيث تصنف إلى: **المعرفات (Identifiers)** التي تُعرّف الفرد بشكل مباشر، كالاسم، ورقم الضمان الاجتماعي [1]، تُزال بالكامل أثناء عملية إخفاء الهوية [8].

شبه المعرفات (Quasi-Identifiers-QI) التي قد تُنتج عند جمعها معاً بصمة فريدة يمكن أن تؤدي إلى إعادة تحديد هوية الفرد، مثل تاريخ الميلاد والجنس [1]. تُطبق على هذه السمات تقنيات إخفاء الهوية، مثل التعميم (Generalization) أو القمع (Suppression) [8]. **السمات الحساسة (Sensitive)** وهي المعلومات التي يرغب الفرد في عدم كشفها للآخرين، مثل معلومات الأمراض والراتب. عادةً لا تُعدّل لأنها تُستخدم لاحقاً في تحليل البيانات [6]، ولكن يمكن حمايتها باستخدام نماذج إخفاء الهوية، مثل k -closeness و l -diversity [12]، أخيراً، **السمات غير الحساسة (Non-Sensitive)** القابلة للنشر للعامة لأنها لا تشكل أي خطر على الخصوصية، مثل الوزن، لون الشعر، والطول. لا تُجمع في معظم الحالات، وإن جُمعت تُنشر كما هي دون تعديل [8].

3-2- تقنيات إخفاء الهوية

يمكن حماية خصوصية البيانات من خلال تطبيق تقنيات إخفاء الهوية التي تحوّل البيانات لضمان تلبية متطلبات الخصوصية، بينما يتم قياس وتقليل الانخفاض في فائدة البيانات [7]. ويُعد التعميم (Generalization) والقمع (Suppression) من أبرز هذه التقنيات [5]، حيث يستبدل التعميم قيم شبه المعرفات بقيم أقل تحديداً تتوافق مع البيانات الأصلية [12]، مما يحدّ من إمكانية تمييز بيانات أي فرد بشكل فريد عند تجميع البيانات، ومع ذلك، فإن الإفراط في التعميم يؤدي إلى انخفاض فائدة البيانات، فبدلاً من استخدام تاريخ ميلاد دقيق، قد تعرض مجموعة البيانات المعممة فئة عمرية (مثل 30-40). أما القمع فيحذف أجزاء معينة من المعلومات لتجنب إعادة تحديد الهوية [5]، ويُستخدم بشكل أساسي للتقليل من التعميم المفرط عبر إزالة القيم المتطرفة (Outliers) التي تكون بعيدة عن السجلات الأخرى وعددها قليل، مما يجعل من غير الممكن تجميعها ضمن نفس فئات التكافؤ [12]. فيما يلي أبرز نماذج إخفاء الهوية المستخدمة على نطاق واسع، والتي تعتمد بشكل أساسي على تقنيتي التعميم والقمع، حيث سيتم تقييمها في هذا البحث:

1) k -anonymity

يُعد نموذج k -anonymity [13]، المعيار الأكثر انتشاراً للخصوصية، حيث يضمن أن كل سجل من البيانات لا يمكن تمييزه عن $k-1$ على الأقل من السجلات الأخرى فيما يتعلق بشبه المعرفات، وذلك بهدف الحماية من الكشف عن الهوية، ومن هجمات الربط (Linking Attack) [14] لكنه لا يأخذ في الاعتبار الكشف عن السمات الحساسة [4].

تُعرّف مجموعة السجلات التي لا يمكن تمييزها عن بعضها البعض بناءً على قيم شبه المعرفات بفئة التكافؤ (Equivalence Class) [1].

(2) k -map :

يشبه نموذج k -map [15] نموذج k -anonymity إلى حد كبير، إلا أن k -map يحسب مستوى الخطر اعتماداً على بيانات المجتمع الإحصائي الكلي (Population) الذي سُحبت منه العينة، وليس فقط على السجلات الموجودة في قاعدة البيانات محل الدراسة [1].

(3) ℓ -diversity :

يُعد نموذج ℓ -diversity [16] امتداداً لنموذج k -anonymity، حيث يعالج بعض قيوده من خلال توفير حماية ضد هجمات الكشف عن السمات، إذ يشترط أن تحتوي كل فئة تكافؤ على ℓ قيم مميزة على الأقل للسمة الحساسة [4]، للحد من ظهور الفئات ذات التنوع المنخفض وإبقاء المهاجم في حالة من عدم اليقين [2].

(4) t -closeness :

يُعد نموذج t -closeness [17] تحسيناً لنموذج ℓ -diversity [3]، إذ يضمن ألا تزيد المسافة بين توزيع القيم الحساسة في كل فئة تكافؤ والتوزيع الكلي عن حد معين t ، حيث $0 \leq t < 1$ [1]، وذلك بهدف الحفاظ على الخصائص الإحصائية للبيانات بشكل أكثر فعالية [4]. من بين دوال قياس المسافة t ، يُعدّ Earth Mover's Distance (EMD) [17] الأكثر استخداماً في الأدبيات. على سبيل المثال، يوضّح الجدول 1 مجموعة البيانات السكانية الأصلية المُراد نشرها لأغراض الدراسات الإحصائية، والتي تضم سبعة أفراد، حيث يمثل كل سجل بيانات فرد واحد.

الجدول 2. مجموعة البيانات مُخفاة الهوية

الاسم	العمر	الجنس	البلد الأصل	الدخل
إيفا	22	أنثى	الولايات المتحدة	$\leq 50k$
فيولا	23	أنثى	الولايات المتحدة	$> 50k$
هيلين	24	أنثى	كندا	$\leq 50k$
آدم	32	ذكر	اليابان	$\leq 50k$
بوب	30	ذكر	الهند	$\leq 50k$
داني	34	ذكر	الفلبين	$\leq 50k$
جان	31	ذكر	الصين	$> 50k$

الجدول 1. مجموعة البيانات الأصلية

العمر	الجنس	البلد الأصل	الدخل
[20,25]	أنثى	أمريكا الشمالية	$\leq 50k$
[20,25]	أنثى	أمريكا الشمالية	$> 50k$
[20,25]	أنثى	أمريكا الشمالية	$\leq 50k$
[30,35]	ذكر	آسيا	$\leq 50k$
[30,35]	ذكر	آسيا	$\leq 50k$
[30,35]	ذكر	آسيا	$\leq 50k$
[30,35]	ذكر	آسيا	$> 50k$

ويعرض الجدول 2 النسخة مُخفاة الهوية من الجدول 1، والتي تحقق كلاً من النماذج: k -anonymity عند $k=3$ ، بوجود ثلاثة سجلات على الأقل في كل فئة تكافؤ، و ℓ -diversity عند $\ell=2$ ، بوجود قيمتين مختلفتين على الأقل من السمة الحساسة (الدخل) في كل فئة تكافؤ، والنموذج t -closeness عند $t=0.0477$ ، المحسوبة باستخدام مقياس EMD، لضمان ألا تتجاوز المسافة بين توزيع قيم السمة الحساسة في الجدول الأصلي وتوزيعها داخل فئات التكافؤ هذه العتبة. ولتحقيق ذلك تمت إزالة الأسماء باعتبارها معرفات مباشرة. بينما عُممت شبه المعرفات، عبر تحويل سمة العمر إلى فترات زمنية مدتها خمس سنوات، واستُبدلت قيم البلد الأصل بالقارات التي تنتمي إليها.

3-3 - النماذج الرياضية والتوليدية الحديثة

على الرغم من القدرة العالية لنماذج إخفاء الهوية السابقة على حماية البيانات عبر تقنيات التعميم والقمع، إلا أن تعقّد الهجمات السيبرانية، وتطوّر الهجمات المعتمدة على المعرفة الخلفية للمهاجم أدت إلى ظهور الخصوصية التفاضلية (Differential Privacy-DP) كإطار رياضي صارم يوفر حماية مستقلة عن تلك المعرفة [18].

أُقترح نموذج ϵ -DP لأول مرة في [19] لحماية مخرجات الاستعلامات التفاعلية المُجرّاة على مجموعة بيانات حساسة وغير منشورة. حيث يضمن أن وجود سجل أي فرد أو غيابه في مجموعة البيانات غير قابل للتمييز بناءً على إجابة الاستعلام المنشورة، وهذا ما تحدده ميزانية الخصوصية ϵ ، فكلما صغرت قيمتها زادت الخصوصية. ورغم رصانة هذا النموذج الرياضية، إلا أن تطبيقه يتطلب ضجيجاً ضخماً يتسبب في تدهور حاد وفقدان غير مقبول لفائدة البيانات [20].

ولمواجهة معضلة المقايضة الصعبة هذه، اتجهت الأبحاث نحو تقنيات البيانات الاصطناعية (Synthetic Data) كآليات مبتكرة لإخفاء الهوية. والتي تعتمد على بناء نماذج احتمالية تستخلص التوزيعات الإحصائية والارتباطات بين السمات من البيانات الحقيقية، دون الحفاظ على أي صلة بالأفراد الحقيقيين [21]. ومع الطفرة الحالية في الذكاء الاصطناعي والتعلم العميق، توسعت كفاءة المولدات عبر شبكات الخصومة التوليدية (GANs) ومحوّلات الترميز التلقائي المتغيرة (VAEs)، والتي أثبتت تفوقاً استثنائياً في الجمع بين فائدة البيانات وخصوصيتها، بما يتوافق مع التشريعات الصارمة مثل GDPR.

ومع ذلك، فإن النشر العملي للبيانات الاصطناعية المولدة بالذكاء الاصطناعي يصطدم بتحديات بنوية مثل انهيار الأنماط (Collapse Mode)، وظهور الانحيازات التي قد تشوه التوزيع العادل للبيانات عبر المجموعات الديموغرافية [22]. بناءً على هذه التحديات، يظل تقييم النماذج الهيكلية التقليدية الحافظة للبنية الواقعية للبيانات محوراً أساسياً ومكماً لا غنى عنه في أدبيات حماية الخصوصية، وهو محور البحث في هذا المقال.

3-4- فائدة البيانات (Data Utility)

نظراً لعدم تحديد سيناريو استخدام نهائي للبيانات، سيتم اعتماد مجموعة من نماذج الفائدة ذات الأغراض العامة المدعومة في ARX من أجل تقييم فائدة بيانات الخرج بعد تطبيق إخفاء الهوية.

1) مقياس التفاصيل (Granularity):

يقيس هذا النموذج دقة البيانات مُخفاة الهوية اعتماداً على نموذج فقدان المعلومات (Loss Information)، ويُعرّف بشكل رسمي لمجموعة بيانات تحتوي على n سجل و m سمة كالتالي [10]:

$$1 - \frac{1}{m} \sum_{1 \leq x \leq m} \text{loss}(x), \quad [10]$$

حيث يمثل $\text{loss}(x) \in [0,1]$ مقدار فقدان المعلومات للسمة x ، الذي يتم تعريفه على أنه متوسط فقدان

المعلومات لجميع قيم هذه السمة في مجموعة البيانات:

$$\frac{1}{n} \sum_{1 \leq y \leq n} \text{loss}(x, y) \quad [10]$$

تُشير y إلى سجل البيانات، ويُحسب فقدان المعلومات لكل قيمة، المُشار إليه بـ $\text{loss}(x, y)$ ، وفقاً لنوع

السمة x والتحويل المطبق عليها. المزيد من التفاصيل في [10].

أما بالنسبة للقيم التي تم قمعها، يتم تعريف فقدان المعلومات على النحو التالي:

$$\text{loss}(x, y) = 1, \text{ if } \text{value}(x, y) \text{ is suppressed}, \quad [10]$$

2) الإنتروبي غير الموحدة (Non-uniform entropy):

تُعد الإنتروبي مقياساً معروفاً لمحتوى المعلومات، وقد تم التوصية بها كنموذج لقياس الفائدة لاسيما في البيانات الصحية. تعتمد فكرتها الأساسية على مقارنة تكرارات قيم السمات في مجموعة البيانات المحوَّلة مع

التكرارات المقابلة لها في مجموعة البيانات الأصلية [23]. تمثل إنتروبيا المتغير X_j مقياساً لكمية المعلومات التي يتم الكشف عنها عند إظهار قيمة عينة عشوائية من X_j ، أو بصورة مكافئة، مقياساً لمقدار عدم اليقين المتعلق بقيمة تلك العينة العشوائية قبل الكشف عنها [24].

(3) مقياس إمكانية التمييز (Discernibility):

يقيس مدى عدم قابلية تمييز السجل عن غيره من السجلات، من خلال إسناد عقوبة (Penalty) لكل سجل اعتماداً على عدد السجلات التي لا يمكن تمييزها عنه في مجموعة البيانات المُحوّلة. فإذا كان السجل الغير مقموع يقع ضمن فئة تكافؤ حجمها z ، فإنه يُمنح عقوبة مقدارها z . أما إذا كان السجل مقموعاً، فيُمنح عقوبة مقدارها $|D|$ ، أي حجم مجموعة البيانات الأصلية. وتعكس هذه العقوبة حقيقة أن السجل المقموع لا يمكن تمييزه عن أي سجل آخر في مجموعة البيانات [25].

(4) مقياس الخطأ التربيعي على مستوى السمات (Attribute-level squared error):

يُعد مجموع الأخطاء التربيعية (Sum of Squared Errors) مقياساً معروفاً على نطاق واسع لفقدان المعلومات، يُعرّف على أنه مجموع مربعات المسافات بين قيم السمات في السجلات ضمن مجموعة البيانات الأصلية X ونظيراتها في مجموعة البيانات المُخفاة الهوية. وبما أن هذا مقياس يوفّر قيماً مطلقة للخطأ، فقد تم تطبيقه للحصول على مقياس مستقل عن حجم مجموعة البيانات وكذلك عن نطاقات قيم السمات [26].
نقترح في هذا البحث مقياس فائدة موحد (Unified Utility Measure) UUM لتقييم أثر إخفاء الهوية بصورة شاملة، يجمع UUM بين المقاييس السابقة التي غالباً ما تعكس جوانب جزئية فقط من فائدة البيانات وقد تؤدي عند تقييمها بشكل مستقل إلى تفسيرات متعارضة، حيث يستند رياضياً على نموذج التجميع الخطي الموزون (Weighted Sum Model)، وهو أسلوب رياضي قياسي معتمد في نظرية القرار متعدد المعايير لدمج مؤشرات مستقلة تعكس أبعاداً جزئية مختلفة، حيث يمتاز بالبساطة والمرونة، ويضمن الحفاظ على خصائص الاستمرارية والتبادلية، ويسهل تفسيره في بيانات نشر البيانات الحقيقية [27]. تُصاغ العلاقة الرياضية للمقياس المقترح على النحو التالي:

$$UUM(D^*) = \sum_{i=1}^4 w_i \cdot U_i(D^*) ;$$

$$\sum w_i = 1 ,$$

$$U_i \in \{Granularity, Non - uniform entropy, Discernibility, Attribute - level squared error\}$$

تمثل D^* مجموعة البيانات المُخفاة الهوية، حيث يخضع المقياس للقيود الرياضية والمنهجية التالية:

- 1- جميع المقاييس المكونة لمقياس UUM والمتمثلة بـ U_i تم تطبيقها داخلياً في ARX ضمن المجال [0-1]، حيث تمثل القيمة 1 (أو 100%) مجموعة البيانات الأصلية غير المعدلة، بينما تمثل القيمة 0 (أو 0%) مجموعة بيانات أزيلت منها جميع المعلومات [10]. وبالتالي القيم الأعلى تكون دائماً أفضل من حيث الفائدة.
- 2- تمثل w_i الأوزان الممنوحة لكل مقياس بناءً على حساسيته الهيكلية. وتخضع لشرط الاستقرار الرياضي التالي:

$$\sum_{i=1}^4 w_i = 1, \quad \text{where } w_i > 0$$

- 3- الخصائص الحدودية وسلوك المقياس: استناداً إلى الشروط السابقة، فإن قيم UUM دائماً محصورة ضمن المجال:

$$0 \leq UUM(D^*) \leq 1$$

وتتميز هذه الصيغة الرياضية بخصائص الرتبة الطردية، وبالتالي أي تحسن في أي مقياس فرعي يؤدي بالضرورة إلى زيادة موازية في قيمة UUM الإجمالية، حيث تتوافق القيم الأعلى لمقياس UUM مع فائدة بيانات أعلى. في هذا البحث تم اختيار الأوزان بناءً على أساس منهجي وتحليلي يوازن بين خصائص المقاييس وطبيعة مجموعة بيانات المستخدمة في التجارب، بهدف قياس الفائدة التحليلية للبيانات مخفاة الهوية بدقة، وذلك بعد إجراء تحليل حساسية (في الفقرة 5-2- تحليل حساسية مقياس UUM لتغير الأوزان) أثبت الاستقرار الرياضي للمقياس المقترح عبر سيناريوهات مختلفة. وذلك وفق المبررات التالية فيما يتعلق بمجموعة بيانات Adult:

- مقياس الإنتروبيا غير الموحدة (الوزن 0.40): مُنح الوزن الأعلى نظراً لدوره الجوهري وحساسيته العالية في الحفاظ على التوزيع الاحتمالي الأصلي للقيم؛ حيث تعتمد فائدة التحليل الإحصائي طردياً على استقرار هذا التوزيع.
- مقياس التفاصيل (الوزن 0.30): حُصص له هذا الوزن المرتفع لكونه يعكس بدقة خسارة المعلومات الناتجة عن مستويات التعميم المطبقة، وهو مؤشر مباشر على مدى احتفاظ البيانات بخصائصها التفصيلية المفيدة.
- مقياس الخطأ التربيعة على مستوى السمات (الوزن 0.20): مُنح وزناً متوسطاً بالنظر إلى أن مجموعة البيانات تحتوي على سمة رقمية واحدة (سمة العمر) يتأثر بها هذا المقياس بشكل أساسي في حساب ضياع المعلومات.
- مقياس إمكانية التمييز (الوزن 0.10): حُصص له الوزن الأدنى لارتباطه الوثيق بنسبة السجلات المقموعة وحجم فئات التكافؤ، مما يجعله يعكس آثار عملية إخفاء الهوية بصورة غير مباشرة أكثر من تعبيره عن فائدة البيانات.

ونظراً لأن مجموعة البيانات الصحية جميع سماتها ذات طبيعة رقمية، تمت إعادة توزيع الأوزان منهجياً؛ حيث رُفع وزن مقياس الخطأ التربيعة على مستوى السمات ليصبح (0.40)، وتلاه مقياس الإنتروبيا غير الموحدة بوزن (0.30) لضمان مراقبة استقرار التوزيع الاحتمالي للبيانات، بينما حُصص وزن (0.20) لمقياس التفاصيل، وحافظ مقياس إمكانية التمييز على الوزن الأدنى (0.10).

وتجدر الإشارة إلى أن هذه الأوزان لا تمثل قيمةً معيارية ثابتة أو عالمية، بل هي إطار مرجعي أثبت استقراره تجريبياً، ويمكن لمسؤولي حماية البيانات تعديلها ديناميكياً وفقاً لحالة الاستخدام النهائي وأولويات المؤسسة وقد تضمن قسم التوصيات دليلاً إرشادياً يوضح المعايير المنهجية لاختيار الأوزان المناسبة وتعديلها وفقاً لهدف كل دراسة.

4- النتائج والمناقشة:

في مجموعة بيانات Adult، تم اعتماد 8 سمات كشبه معرفات في التجارب، وهي: العمر، فئة العمل، التعليم، الحالة الاجتماعية، المهنة، العرق، الجنس، والبلد الأصل. أما سمة الدخل تم اعتمادها كسمة حساسة لكونها تمثل بيانات مالية. نظراً لحساسية مجموعة البيانات الصحية العالية، تم اعتماد 6 سمات كشبه معرفات، وهي: العمر، الجنس، ضغط الدم الانقباضي أثناء الراحة (trestbps)، الكوليسترول (chol)، معدل سكر الدم الصيامي (fbs)، والحد الأقصى لمعدل ضربات القلب المحقق (thalach). في حين اعتبرت باقي السمات الثماني سمات حساسة نظراً لارتباطها المباشر بأمراض القلب، وهي: نوع الألم الصدري (cp)، نتائج تخطيط القلب أثناء الراحة (restecg)، الذبحة الصدرية المحفزة بالجهود (exang)، انخفاض مقطع ST الناتج عن الجهد (oldpeak)، ميل مقطع ST أثناء ذروة الجهد (slope)، عدد الأوعية الرئيسية الملونة بالفحص النقيظ (ca)، حالة الثلاثيميا (thal)، وتشخيص المرض القلبي (num).

بعد تصدير مجموعة البيانات إلى أداة ARX، تم بناء التسلسلات الهرمية للتعميم لكل سمة شبه معرفات وفق طبيعة بياناتها، حيث يمثل المستوى الأول القيم الأصلية للبيانات، أما المستوى الأخير يمثل القمع الكامل للقيم (استبدالها بالرمز *). تم تمثيل العلاقة بين معدل الخطر والفائدة بيانياً من أجل تحليل واضح للمقايضة وتحديد نقطة المقايضة المثلى. في جميع الأشكال، يمثل المحور الأفقي معدل خطر المدعي العام (Risk)، حيث تشير القيم المنخفضة إلى مستوى خصوصية أعلى، بينما يمثل المحور الشاقولي قيم مقياس UUM الذي يمثل فائدة البيانات (Data Utility)، كذلك يمثل السطر الأول في جميع الجداول معدل الخطر والفائدة للبيانات الأصلية، والذي لم يُمثل بيانياً.

4-1- تقييم نماذج الخصوصية على مجموعة بيانات Adult:

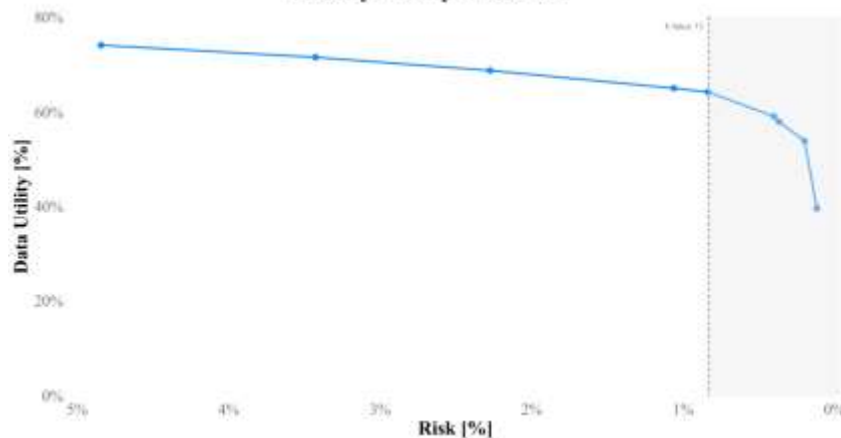
أولاً: تقييم k -anonymity:

يبين الجدول 3 نتائج تطبيق نموذج k -anonymity باستخدام قيم: $k=2,3,5,10,15$ ، وهي القيم الشائعة والمستخدم في الأدبيات ذات الصلة [8، 16]، ولغرض دراسة سلوك النموذج ضمن نطاق أوسع من مستويات الخصوصية، تم اختبار قيم إضافية أكبر هي: 40، 50، 100، 200.

الجدول 3. نتائج تطبيق نموذج k -anonymity

k	Average prosecutor risk	Granularity	N.-U. entropy	Discernibility	Attribute-level squared error	UUM
1	54.76538%	100%	100%	100%	100%	100%
2	4.83932%	84.89577%	53.95683%	93.11208%	88.75258%	74.11319%
3	3.42386%	82.35758%	51.49837%	90.26364%	86.15024%	71.56300%
5	2.27193%	79.23407%	48.77942%	87.57163%	83.62146%	68.76344%
10	1.05879%	74.96503%	42.72509%	87.53657%	83.34903%	65.00301%
15	0.83380%	72.40117%	42.50977%	89.37185%	82.85534%	64.23251%
40	0.39832%	67.32652%	36.48847%	83.43286%	79.60260%	59.05715%
50	0.36319%	66.19244%	35.62627%	81.90831%	78.04819%	57.90871%
100	0.19589%	60.03092%	31.00283%	81.05571%	76.68966%	53.85391%
200	0.11501%	55.42926%	27.59413%	77.43633%	21.09448%	39.62896%

Privacy-Utility Trade-off



الشكل 1. المقايضة بين الخصوصية وفائدة البيانات في نموذج k -anonymity

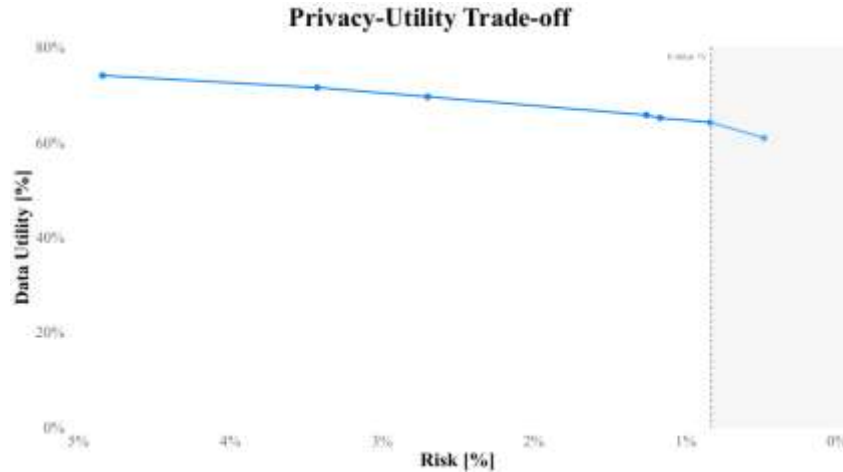
تظهر النتائج في الجدول 3 والشكل 1 وجود علاقة عكسية واضحة بين خصوصية البيانات (المتمثلة بانخفاض معدل خطر المدعي العام Average Prosecutor Risk) وفائدتها (المقاسة بمقياس UUM)، وبالتالي عند زيادة قيمة k ينخفض معدل الخطر، أي يرتفع مستوى الخصوصية وتنخفض فائدة البيانات. تطبيق نموذج k -anonymity عند $k=2$ ، يؤدي إلى تحسن كبير في مستوى الخصوصية، حيث انخفض معدل خطر المدعي العام بنسبة 49.92606% مقارنةً بالقيمة الأصلية، نتيجة التخلص من السجلات الفريدة الموجودة في البيانات الأصلية، التي تكون معرضة لخطر 100% من خلال دمجها مع سجلات أخرى في فئات تكافؤ تحتوي سجلين على الأقل، مقابل انخفاض في فائدة البيانات بنسبة 25.88681% مقارنةً بالفائدة الأصلية. نلاحظ انخفاض ملحوظ في فائدة البيانات عند الانتقال من $k=15$ إلى $k=40$ ، حيث تم فقدان ما يقارب 5% من الفائدة، مقابل تحسن محدود جداً في الخصوصية، إذ ينخفض معدل الخطر بنسبة 0.43548% فقط، بالتالي يمكن اعتبار $k=15$ نقطة المقايضة المثلى لتحقيق مستوى خصوصية مرتفع مع فقدان مقبول في فائدة البيانات. في التجربة الأخيرة عند $k=200$ ، انخفضت فائدة البيانات بشكل حاد بينما أصبح الانخفاض في معدل الخطر محدوداً للغاية، مما يدل على أن القيم الكبيرة لـ k تؤدي إلى انخفاض الفائدة بشكل كبير مقابل تحسن غير ملحوظ في الخصوصية.

ثانياً: تقييم k -map:

يبين الجدول 4 نتائج تطبيق نموذج k -map، باستخدام قيم k ذاتها المستخدمة في تجارب k -anonymity. تم اعتماد مقرر Zero-Truncated Poisson، لتقدير التكرار المتوقع لفئات التكافؤ على مستوى المجتمع الأصلي، حيث يتم تجاهل فئات التكافؤ التي لا تظهر في العينة، وهو المقرر الموصى به في أدبيات k -map [15]. كما تم تحديد مستوى الدلالة الإحصائية عند $\alpha=0.1$ بهدف الحصول على تقدير محافظ لاحتمال فردية الأنماط في المجتمع، مع افتراض حجم مجتمع إحصائي مقداره 452,220 سجلاً، بحيث تمثل قاعدة بيانات Adult نسبة 10% من إجمالي السكان المفترضين، وذلك لدراسة تأثير k -map بشكل واضح على فائدة البيانات والخصوصية، إذ إن زيادة حجم المجتمع الأصلي تقلل من حساسية النموذج لارتفاع قيم k .

الجدول 4. نتائج تطبيق نموذج k -map

k	Average prosecutor risk	Granularity	N.-U. entropy	Discernibility	Attribute-level squared error	UUM
1	54.76538%	100%	100%	100%	100%	100%
2	4.83932%	84.89577%	53.95683%	93.11208%	88.75258%	74.11319%
3	3.42386%	82.35758%	51.49837%	90.26364%	86.15024%	71.56300%
5	3.42386%	82.35758%	51.49837%	87.57163%	86.15024%	71.56300%
10	2.69698%	80.43306%	49.69637%	87.53657%	84.10755%	69.64072%
15	2.69698%	80.43306%	49.69637%	89.37185%	84.10755%	69.64072%
40	1.25710%	76.05174%	43.52282%	83.43286%	83.76056%	65.77317%
50	1.16491%	75.44547%	42.99926%	81.90831%	83.02764%	65.16276%
100	0.83380%	72.40117%	42.50977%	81.05571%	82.85534%	64.23251%
200	0.48004%	69.23214%	38.00994%	77.43633%	81.85633%	60.94653%



الشكل 2. المقايضة بين الخصوصية وفائدة البيانات في نموذج k -map

تظهر النتائج في الجدول 4 والشكل 2 أنه عند زيادة قيمة k ينخفض معدل الخطر، أي يرتفع مستوى الخصوصية وتنخفض فائدة البيانات. عند الانتقال من $k=3$ إلى $k=5$ حصلنا على نفس النتائج فيما يتعلق بمعدل الخطر وفائدة البيانات نتيجة تطبيق نفس التحويل على البيانات، وذلك لأن قيمة k الفعلية (أقل حجم لفئات التكافؤ ضمن العينة) كانت تساوي 3 في كلا التجريبتين، مما يدل على أن $k=3$ في العينة تحقق $k=3$ و $k=5$ ضمن المجتمع الأصلي.

تكرر ذلك عند الانتقال من $k=10$ إلى $k=15$ لنفس السبب.

نلاحظ انخفاض ملحوظ في فائدة البيانات عند الانتقال من $k=100$ إلى $k=200$ ، حيث تم فقدان ما يقارب 3% من الفائدة، مقابل تحسن محدود جداً في الخصوصية بانخفاض معدل الخطر بنسبة 0.35376%، بالتالي يمكن اعتبار $k=100$ نقطة المقايضة المثلى لتحقيق مستوى خصوصية مرتفع مع فقدان مقبول في فائدة البيانات. عند قيم k الكبيرة (أكبر من 40) يصبح الانخفاض في معدل الخطر وفي فائدة البيانات محدوداً جداً، وصولاً إلى $k=200$ ، تنخفض الفائدة إلى 60.94653% متأثرةً بازدياد كثافة التعميم والوصول إلى أعلى نسبة سجلات مقموعة ضمن التجارب (12.47623%).

حقق نموذج k -map عند $k=15$ فائدة بيانات أعلى بنسبة 5.40821% مقارنةً بالتالي حققها نموذج k -anonymity عند عتبة المقايضة المثلى ($k=15$)، وذلك نتيجة انخفاض كثافة التعميم المطبق، بينما حقق نموذج k -anonymity مستوى خصوصية أعلى بنسبة انخفاض 1.86318% في معدل الخطر، لأنه يفرض قيد الخصوصية ($k=15$) مباشرةً على العينة وليس على المجتمع الأصلي المُقدَّر.

يتمثل الاختلاف الجوهرى بين النموذجين في أن k -map يفترض أن المهاجم قد يمتلك وصولاً إلى المجتمع الأصلي الذي أخذت منه العينة، وهو السيناريو الأكثر واقعية، وبالتالي يطبق الحماية على مستوى المجتمع وليس على مستوى العينة، فإن السجلات التي تظهر بعدد قليل في العينة، قد تكون أكثر انتشاراً في المجتمع الأصلي، لذلك يمكن لنموذج k -map تحقيق نتائج مشابهة لنتائج k -anonymity عند العتبات المثلى للمقايضة.

ثالثاً: تقييم ℓ -diversity:

على الرغم من أن النماذج السابقة، تحقق مستوى مرتفع من الخصوصية والفائدة، إلا أنها لا تأخذ بعين الاعتبار السمة الحساسة (سمة الدخل). فعلى سبيل المثال، إذا احتوت إحدى فئات التكافؤ على سجلات تمتلك جميعها نفس القيمة للسمة الحساسة (مثلاً جميعها لها نفس الدخل $\leq 50k$)، فإن هذه الفئة تصبح عرضة لهجوم التجانس (Homogeneity Attack) [3]، فالمهاجم الذي يعلم أن سجل الضحية ينتمي إلى هذه الفئة، سيستنتج أن دخله منخفض. لذا يبرز نموذج ℓ -diversity كحل يضمن تنوع قيم السمة الحساسة ضمن فئات التكافؤ.

يبين الجدول 5 نتائج تطبيق نموذج ℓ -diversity، باستخدام $\ell=2$ ، نظراً لعدم تنوع قيم السمة الحساسة المستخدمة في الدراسة، وهي سمة الدخل الثنائية، التي تحتوي على قيمتين فريدتين فقط فإن أقصى قيمة ممكنة ل ℓ في هذه الحالة هي $\ell=2$ ، وهذا يعتبر أحد قيود النموذج عند التعامل مع سمات حساسة ذات تنوع محدود في القيم. إذ يعتبر هذا النموذج فعالاً بشكل خاص في الحد من المخاطر في مجموعات البيانات التي تحتوي على سمات حساسة فئوية [4].

الجدول 5. نتائج تطبيق نموذج ℓ -diversity

ℓ	Average prosecutor risk	Granularity	N.-U. entropy	Discernibility	Attribute-level squared error	UUM
1	54.76538%	100%	100%	100%	100%	100%
2	1.62623%	70.04410%	39.89278%	81.73791%	75.88680%	60.32149%

تتوزع قيم السمة الحساسة بشكل غير متوازن، حيث تمثل السجلات ذات الدخل المنخفض ($\leq 50k$) حوالي 75.2% من إجمالي السجلات، بينما تمثل السجلات ذات الدخل المرتفع ($> 50k$) حوالي 24.8% من إجمالي السجلات فقط، مما أدى إلى ارتفاع كثافة التعميم المطبق ونسبة السجلات المقموعة حيث بلغت حوالي 17% عند تطبيق نموذج ℓ -diversity عند $\ell=2$ ، وذلك لضمان وجود قيمتين فريدتين مختلفتين من السمة الحساسة ضمن كل فئة تكافؤ، وهذا يفسر انخفاض فائدة البيانات بشكل ملحوظ إلى 60.32149%.

حقق نموذج k -anonymity عند عتبة المقايضة المثلى ($k=15$)، فائدة بيانات أعلى بنسبة 3.91102% مقارنةً بنموذج ℓ -diversity عند $\ell=2$ ، نظراً لارتفاع كثافة التعميم المطبق وزيادة نسبة السجلات المقموعة في الأخير بنسبة 7.75508%. كما حقق نموذج k -anonymity عند $k=15$ مستوى خصوصية أعلى، حيث انخفض معدل الخطر بنسبة 0.79243%، مقارنةً بنموذج ℓ -diversity عند $\ell=2$ الذي يسمح بوجود فئات تكافؤ تحتوي سجلين على الأقل لضمان وجود قيمتين فريدتين من السمة الحساسة في كل فئة تكافؤ، بينما k -anonymity عند $k=15$ يضمن وجود فئات تكافؤ تحتوي على الأقل 15 سجلاً، وهذا يقلل معدل خطر المدعي العام بشكل واضح.

إلا أن نموذج ℓ -diversity، يوفر طبقة حماية إضافية لفئات التكافؤ التي من الممكن أن تحتوي على نفس القيمة للسمة الحساسة والتي ستكون عرضة لهجمات الكشف عن السمات.

رابعاً: تقييم ℓ -closeness:

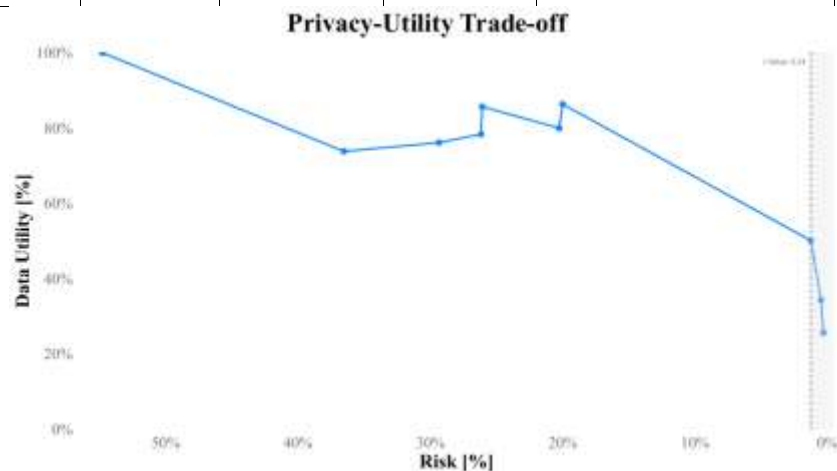
نظراً لعدم توازن توزيع قيم السمة الحساسة، تصبح البيانات عرضة لهجوم الانحراف (Skewness Attack) [17] حتى بعد تطبيق ℓ -diversity، وذلك لأن هذا النموذج يركز على عدد القيم المختلفة للسمة

الحساسية داخل فئات التكافؤ دون مراعاة توزيعها الاحتمالي. فعلى سبيل المثال، إذا احتوت فئة تكافؤ على 90% من السجلات ذات الدخل المنخفض، و10% من السجلات ذات الدخل المرتفع، فإن المهاجم الذي يستهدف هذه الفئة سيكون لديه احتمال بنسبة 90% لاستنتاج أن دخل الفرد المستهدف منخفض، وهذا يُعرف بهجوم الانحراف. ولمعالجة هذا القيد تم تطبيق نموذج f -closeness، الذي يضمن تقارب توزيع قيم السمة الحساسة داخل فئات التكافؤ من توزيعها الكلي.

يبين الجدول 6 نتائج تطبيق نموذج f -closeness، باستخدام عدة قيم للمعامل f ، وهي: $f=0.1, 0.2, 0.24, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 1$ ، وذلك لدراسة أشمل لتأثير قيمة f على كلٍّ من فائدة البيانات ومستوى الخصوصية. ولحساب المسافة f ، تعتمد أداة ARX على مقياس EMD.

الجدول 6. نتائج تطبيق نموذج f -closeness

f	Average prosecutor risk	Granularity	N.-U. entropy	Discernibility	Attribute-level squared error	UUM	Sample uniques
1	54.76538%	100%	100%	100%	100%	100%	41.17244%
0.9	54.76538%	100%	100%	100%	100%	100%	41.17244%
0.8	54.76538%	100%	100%	100%	100%	100%	41.17244%
0.7	19.98307%	93.46966%	73.63315%	96.56435%	95.89358%	86.32931%	11.88782%
0.6	26.06309%	90.50057%	76.35021%	92.61552%	93.52230%	85.65627%	16.01203%
0.5	20.24261%	86.56652%	68.72958%	90.28007%	87.41650%	79.97310%	12.22089%
0.4	26.15501%	82.47148%	70.00934%	84.40892%	85.84910%	78.35589%	16.53535%
0.3	29.33217%	79.52187%	69.05828%	80.82714%	82.77419%	76.11743%	19.14655%
0.25	36.50872%	75.35004%	69.34595%	75.89880%	79.27989%	73.78925%	25.94286%
0.24	1.22281%	55.65384%	32.27164%	70.11733%	67.23913%	50.06437%	0%
0.2	0.46663%	46.21690%	21.77320%	76.67038%	20.32140%	34.30567%	0%
0.1	0.28181%	34.48275%	14.81843%	58.46143%	17.39764%	25.59787%	0%



الشكل 3. المقايضة بين الخصوصية وفائدة البيانات في نموذج f -closeness

تظهر النتائج في الجدول 6 أن القيم المنخفضة للمعامل t توفر مستوى أعلى من الخصوصية، حيث بلغ معدل خطر المدعي العام 0.28181% عند $t=0.1$ ، مع عدم وجود سجلات فريدة في العينة، إلا أن فائدة البيانات كانت منخفضة نسبياً عند هذه القيم. مع زيادة قيمة t يزداد مستوى فائدة البيانات نتيجة انخفاض نسبة السجلات المقموعة وكثافة التعميم المطبق في معظم التجارب، مع ظهور سجلات فريدة في العينة.

يلاحظ وجود تذبذب في قيم معدل خطر المدعي العام عند بعض قيم المعامل t ، ويعكس ذلك طبيعة نموذج t -closeness، الذي يفرض قيوداً على توزيع السمة الحساسة داخل فئات التكافؤ دون فرض حد أدنى لحجم هذه الفئات، ونتيجة لذلك قد تتغير أحجام فئات التكافؤ عند تغيير قيمة t ، مما يؤدي إلى تغير معدل خطر المدعي العام.

عند الانتقال من $t=0.24$ إلى $t=0.25$ ، تحسنت فائدة البيانات بشكل كبير بنسبة 23.72488%، لكن بالمقابل ارتفع معدل الخطر بشكل كبير أيضاً، حيث وصل إلى 36.50872%، وهذه نسبة خطر مرتفعة غير مقبولة، كذلك أصبحت نسبة السجلات فريدة 25.94286%، بالتالي لا يمكن اعتبار $t=0.25$ نقطة مقايضة مثلى.

أما عند الانتقال من $t=0.2$ إلى $t=0.24$ ، تحسنت فائدة البيانات بشكل كبير بنسبة 15.75870%، بالمقابل ارتفع معدل الخطر بشكل بسيط، حيث وصل إلى 1.22281%، دون وجود سجلات فريدة في العينة مُخافة الهوية، بالتالي يمكن اعتبار $t=0.24$ نقطة المقايضة المثلى، كما هو موضح في الشكل 3.

يمكن تفسير التحسن الملحوظ في فائدة البيانات عند $t=0.25$ ، بكون هذه القيمة تساوي تقريباً نسبة السجلات التي دخلها $>50k$ ، والتي تمثل الفئة الأقل تمثيلاً في السمة الحساسة، فعند هذه العتبة يصبح نموذج t -closeness أكثر تساهلاً في قبول فئات تكافؤ ذات توزيع محلي منحرف عن التوزيع العام للسمة الحساسة.

وبشكل خاص، قد تصبح الفئات التي تهيمن عليها قيمة واحدة للسمة الحساسة مقبولة ضمن قيد الخصوصية، مما يؤدي إلى تقليل مستوى التعميم، وظهور سجلات فريدة في العينة، وبالتالي تحسن فائدة البيانات بشكل ملحوظ، وذلك على حساب انخفاض مستوى الخصوصية.

عند القيم الكبيرة للمعامل t ($t \geq 0.8$)، يتم الاحتفاظ بالبيانات الأصلية دون تعديل، نظراً لأن توزيع السمة الحساسة في البيانات الأصلية يحقق شرط t -closeness بشكل طبيعي عند هذه القيم، مما يؤدي إلى تحقيق فائدة كاملة (100%)، إلا أن ذلك يكون على حساب مستوى الخصوصية.

حقق نموذج k -anonymity عند نقطة المقايضة المثلى ($k=15$) فائدة بيانات أعلى من تلك التي حققها نموذج t -closeness عند نقطة المقايضة المثلى ($t=0.24$)، وذلك بنسبة 14.16815%.

كذلك حقق خصوصية أعلى، حيث بلغ معدل خطر المدعي العام 0.83380% عند تطبيق نموذج k -anonymity مقارنةً بـ 1.22281% في نموذج t -closeness، وذلك لأن نموذج k -anonymity يفرض حداً أدنى لعدد السجلات ضمن فئات التكافؤ، في حين يركز نموذج t -closeness على ضمان تقارب توزيع السمة الحساسة داخل فئات التكافؤ مع التوزيع العام، وهو قيد أكثر صرامة من حيث حماية السمة الحساسة.

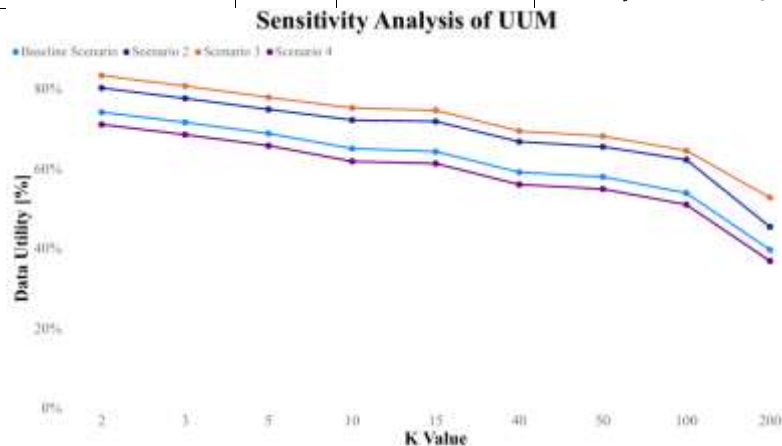
حقق نموذج ℓ -diversity عند $\ell=2$ فائدة بيانات أعلى بنسبة 10.25712%، مقارنةً بالفائدة التي حققها نموذج t -closeness عند نقطة المقايضة المثلى ($t=0.24$)، بسبب ارتفاع نسبة السجلات المقموعة عند تطبيق t -closeness، حيث بلغت الزيادة حوالي 11.69563%. بالمقابل، حقق نموذج t -closeness عند $t=0.24$ مستوى خصوصية أفضل، حيث انخفض معدل خطر المدعي العام بنسبة 0.40342% مقارنةً بنموذج ℓ -diversity.

4-2- تحليل حساسية مقياس UUM لتغير الأوزان:

وللتحقق من استقرار النتائج وموثوقية الأوزان المختارة، تم إجراء تحليل حساسية (Sensitivity Analysis) باستخدام عدة سيناريوهات مختلفة من الأوزان كما هو موضح في الجدول 7، حيث يمثل كل سيناريو توجهاً تحليلياً مختلفاً، وذلك بالاعتماد على نتائج تقييم نموذج k -anonymity (النموذج المرجعي لإخفاء الهوية). يظهر الشكل 4 منحنيات تحليل الحساسية، والتي نلاحظ أنها تسير بسلوك رتيب وتتحرك بشكل متواز تقريباً رُغم التباين في قيم الأوزان. حيث ثبتت نقطة المقايضة المثلى عند $k=15$ في جميع السيناريوهات. هذا السلوك الرياضي المستقر يقدم دليلاً تجريبياً على رصانة المقياس المقترح، ويؤكد أنه يعكس الخصائص الهيكلية الحقيقية للبيانات المخففة الهوية وضياح الفائدة الفعلي، بمعزل عن الانحياز الذاتي للأوزان، مع تأكيد أن السيناريو الأساسي يوفر التقييم الأكثر توازناً للفائدة التحليلية العامة.

الجدول 7. السيناريوهات المقترحة لتحليل حساسية مقياس UUM لتغير الأوزان

السيناريوهات المقترحة لتحليل الحساسية	الإنتروبيا غير الموحدة	التفاصيل	الخطأ التربيعي على مستوى السمات	إمكانية التمييز
السيناريو الأساسي (يركز على الفائدة التحليلية العامة)	0.40	0.30	0.20	0.10
السيناريو المحايد (أوزان متساوية لجميع المقاييس)	0.25	0.25	0.25	0.25
سيناريو يركز على مقدار التعميم والقمع المطبق	0.15	0.40	0.15	0.30
سيناريو يركز على الحفاظ على المحتوى المعلوماتي للبيانات	0.50	0.20	0.20	0.10



الشكل 4. منحنيات تحليل الحساسية لمقياس UUM

4-3 - التحليل التجريبي المقارن لنماذج إخفاء الهوية:

لتقييم النتائج وضمان أن الفروقات الملحوظة بين النماذج ذات دلالة إحصائية، تم إجراء اختبار دلالة إحصائية من نوع اختبار t للمجموعات المرتبطة (t-Test: Paired Two Sample) [28] عبر تسع تجارب متقابلة لتوحيد العوامل، وذلك للمقارنة بين النماذج k -anonymity، k -closeness و k -map، على مستوى مقياس الفائدة UUM. يبين الجدول 8 نتائج هذه الاختبارات، حيث أظهرت النتائج الإحصائية في الاختبار الأول بين k -anonymity و k -closeness تفوقاً حقيقياً لصالح نموذج k -anonymity، حيث بلغت قيمة الاختبار المحسوبة (-3.62)، وسجلت القيمة الاحتمالية (P -value = 0.0067)، وهي أقل بكثير من مستوى الدلالة ($\alpha = 0.05$). وفي الاختبار الثاني بين k -anonymity و k -map، أظهر التحليل الإحصائي تفوقاً حقيقياً لصالح نموذج k -map، حيث بلغت قيمة الاختبار (-3.0002)، وسجلت القيمة الاحتمالية (P -value = 0.0170). تثبت هذه

النتيجة رياضياً أن نموذج k -map يمتلك قدرة أعلى في الحفاظ على فائدة البيانات مقارنة بنموذج k -anonymity ضمن شروط التجربة نفسها. وهذا يتوافق مع النتائج التي تم التوصل إليها سابقاً في هذه الدراسة.

الجدول 8. نتائج اختبارات الدلالة الإحصائية (t-Test) بين النماذج على مستوى مقياس UUM

الاختبار	قيمة الاختبار t	القيمة الاحتمالية P-value	مستوى الدلالة α
الأول: بين نموذجي k -anonymity و t -closeness	-3.62	0.0067	0.05
الثاني: بين نموذجي k -map و k -anonymity	-3.0002	0.0170	0.05

4-4 - تقييم نماذج الخصوصية على مجموعة بيانات أمراض القلب:

أولاً: تقييم k -anonymity :

يبين الجدول 9 نتائج تطبيق نموذج k -anonymity باستخدام القيم: $k = 2, 3, 5, 10, 15, 40$.

الجدول 9. نتائج تطبيق نموذج k -anonymity

k	Average prosecutor risk	Granularity	N.-U. entropy	Discernibility	Attribute-level squared error	UUM
1	100%	100%	100%	100%	100%	100%
2	11.80556%	46.48610%	63.06615%	92.02839%	49.96336%	58.71553%
3	10.14493%	44.43184%	62.08047%	88.00164%	46.76653%	56.54868%
5	6.25%	39.79835%	58.93313%	82.98298%	45.53496%	53.49171%
10	2.86738%	30.51996%	55.32934%	77.76185%	48.18922%	50.46868%
15	2.82258%	29.69540%	54.97525%	71.32132%	0.63485%	35.25177%
40	1.45631%	21.58970%	50.40417%	49.17645%	0.43066%	29.52645%

تظهر النتائج في الجدول 9 وجود علاقة عكسية واضحة بين قيمة k وكل من معدل خطر المدعي العام وفائدة البيانات. تطبيق k -anonymity عند $k=2$ ، يؤدي إلى تحسن كبير في مستوى الخصوصية، نتيجة التخلص من السجلات الفريدة، والتي كانت تمثل نسبة 100% من السجلات (أي أن جميع سجلات مجموعة البيانات هي سجلات فريدة ضمن العينة)، مقابل انخفاض فائدة البيانات بنسبة 41.28% مقارنةً بالفائدة الأصلية، وذلك نتيجة ارتفاع كثافة التعميم المطبق وقمع نسبة 3% تقريباً من السجلات، لضمان وجود سجلين على الأقل ضمن كل فئة تكافؤ.

نلاحظ انخفاض ملحوظ في فائدة البيانات عند الانتقال من $k=10$ إلى $k=15$ ، نتيجة قمع ما يقارب 16.5% من السجلات، مقابل تحسن محدود جداً في الخصوصية، إذ ينخفض معدل الخطر بنسبة 0.045% فقط، بالتالي يمكن اعتبار $k=10$ نقطة المقايضة المثلى لتحقيق مستوى خصوصية مرتفع مع أقل فقدان في فائدة البيانات. حقق نموذج k -anonymity على مجموعة بيانات Adult نقطة مقايضة مثلى عند $k=15$ ، بينما على مجموعة بيانات أمراض القلب حققها عند $k=10$ ، متأثراً بحجم العينة الصغير، حيث يبلغ الفرق بين حجم العينتين ما يقارب 44,703 سجل، بالإضافة إلى طبيعة مجموعة بيانات أمراض القلب، حيث تبلغ نسبة السجلات الفريدة فيها 100%، بينما في مجموعة بيانات Adult بلغت هذه النسبة 41.17%، هذا التفرد المطلق وحجم العينة الصغير يفرض على خوارزمية الإخفاء تطبيق قيود قمع وتعميم قاسية لتشكيل فئات التكافؤ المطلوبة، مما يسرع من تدهور فائدة البيانات عند قيم أقل للمعامل k .

ثانياً: تقييم ℓ -diversity:

يبين الجدول 10 نتائج تطبيق نموذج ℓ -diversity، باستخدام قيم متدرجة للمعامل ℓ من 2 إلى 20. بدايةً تم تطبيق القيمة $\ell=2$ على جميع السمات الحساسة، ثم تم تطبيق هذا النموذج عند $\ell=3$ وذلك على جميع السمات الحساسة ماعدا السمات الثنائية (والتي تم الإبقاء عليها عند القيد $\ell=2$ نظراً لأن هذه القيمة تمثل التنوع المطلق والكامل الممكن لقيمتها الأصلية)، اتباعاً لنفس الاستراتيجية السابقة تم رفع القيد تدريجياً من $\ell=4$ وصولاً إلى $\ell=20$ ، الذي أدى إلى انهيار في فائدة البيانات، وذلك لضمان تنوعاً أكثر أماناً للقيم الحساسة داخل كل فئة تكافؤ.

الجدول 10. نتائج تطبيق نموذج ℓ -diversity

ℓ	Average prosecutor risk	Granularity	N.-U. entropy	Discernibility	Attribute-level squared error	UUM
1	100%	100%	100%	100%	100%	100%
2	5.79151%	37.74689%	57.97640%	78.68323%	40.33489%	50.70873%
3	1.18110%	19.17344%	51.36029%	56.72946%	44.83442%	43.50208%
4	1.18110%	19.17344%	51.36029%	56.72946%	44.83442%	43.50208%
20	0.83333%	17.19592%	49.14796%	43.91664%	0.53607%	27.65085%

تظهر النتائج في الجدول 10 وجود علاقة عكسية واضحة بين خصوصية البيانات وفائدتها. فعند زيادة قيمة ℓ ينخفض معدل الخطر، أي يرتفع مستوى الخصوصية وتتخفض فائدة البيانات. فإن تطبيق 2-diversity يؤدي إلى تحسن كبير جداً بالخصوصية، نتيجة التخلص من السجلات الفريدة، مقابل انخفاض في فائدة البيانات بمقدار 49.29127%، وذلك نتيجة ارتفاع كل من كثافة التعميم المطبق وبلوغ نسبة السجلات المقموعة حوالي 12.8%. عند الانتقال من $\ell=2$ على $\ell=3$ انخفضت فائدة البيانات بنسبة 7.20665%، مقابل تحسن الخصوصية بنسبة 4.61041%، وذلك نتيجة ارتفاع كثافة التعميم بشكل كبير وزيادة نسبة السجلات المقموعة التي وصلت إلى 14.47811%.

نلاحظ أن زيادة قيمة ℓ إلى 4 أدى إلى نفس النتائج السابقة، مما يشير إلى أن ضمان 3 قيم مختلفة لكل سمة حساسة أدى إلى مستويات تعميم مرتفعة ونسبة قمع مرتفعة كذلك تضمن وجود 4 قيم مختلفة للسمات الحساسة (عدا الثنائية) على الأقل في كل فئة تكافؤ، هذه الفئات المتولدة احتوت بالصدفة الإحصائية على توزيع متنوع طبيعي وغني للقيم الحساسة في القيود الأعلى تلقائياً دون توضيحية إضافية. فعلى سبيل المثال، على الرغم من أن السمة الحساسة المستمرة (oldpeak) تحتوي على 40 قيمة فريدة، إلا أن الهيكلية المتشكلة عند $\ell=3$ حققت لها تنوعاً داخلياً كافياً لاستيعاب قيود تصل إلى $\ell=19$ دون أي قمع أو تعميم إضافي. أما الانتقال لفرض القيد $\ell=20$ ، أدى لزيادة كثافة التعميم وتطبيق نسبة قمع سجلات بمقدار 19.19192%، مما أدى إلى انهيار الفائدة بشكل كبير، مقابل تحسن بسيط في الخصوصية. بالتالي يمكن اعتبار $\ell=3$ نقطة مقايضة مثلى تحقق مستوى خصوصية مرتفع مع أقل فقدان مقبول في فائدة البيانات.

حقق نموذج ℓ -diversity (عند $\ell=2$) عند تطبيقه على مجموعة بيانات Adult، فائدة بيانات أعلى ومعدل خطر أقل، مقارنة بنتائجه على مجموعة بيانات أمراض القلب، السبب يعود لحساسية البيانات الصحية العالية، فحماية 8 سمات حساسة يتطلب توضيحية أعلى بالفائدة مقارنة بحماية سمة حساسة واحدة (سمة الدخل في مجموعة بيانات Adult).

5- الاستنتاجات والتوصيات:

قدّم هذا البحث تقييماً تجريبياً لنماذج إخفاء الهوية الأكثر شيوعاً، بهدف تحليل المقايضة بين الخصوصية وفائدة البيانات وتحديد القيم المثلى لكل نموذج. وأظهرت النتائج أن تحقيق هذه المقايضة بين الخصوصية والفائدة يفرض قبول مستوى محدد من المخاطر، إذ لا يمكن الوصول إلى حماية كاملة، وذلك بهدف الحفاظ على مستوى مقبول أيضاً من الفائدة، وتتلخص أبرز الاستنتاجات والتوصيات المرتبطة بها فيما يلي:

حقق نموذج k -anonymity أعلى خصوصية وفائدة مقارنة بباقي النماذج المدروسة، وذلك لأنه يفرض قيوده على حجم فئات التكافؤ مباشرة دون أخذ السمات الحساسة بعين الاعتبار، مما يحمي من هجمات الكشف عن الهوية دون توفير حماية كافية ضد هجمات الكشف عن السمات [17]. كذلك تبين تأثيره الواضح بحجم مجموعة البيانات، ونسبة السجلات الفريدة حيث تسببت زيادتها في تسريع وتيرة انخفاض الفائدة. وبالمثل، حقق نموذج k -map عند عتبات المقايضة المثلى نتائج مشابهة لتلك التي حققها نموذج k -anonymity، لكن بقيد خصوصية أعلى وأكثر صرامة (قيم أكبر ل k).

وبالتالي يوصى بتطبيق نموذج k -anonymity بالاقتران مع نماذج خصوصية أخرى تحمي السمات الحساسة، (مثل ℓ -diversity)، واعتماده في سيناريوهات هجوم المدعي العام، بينما يُوصى باعتماد k -map في سيناريوهات هجوم الصحفي. وفي حال كان كلا السيناريوهين ممكناً، يفضل اعتماد k -anonymity لأنه يأخذ بعين الاعتبار الحالة الأسوأ للهجوم، وذلك يتوافق مع ماورد في الأدبيات السابقة [15].

أظهرت النتائج أيضاً أن نموذج ℓ -diversity يعاني من مجموعة من القيود عند عدم توازن توزيع قيم السمة الحساسة (تبقى البيانات عرضة لهجوم الانحراف). كذلك عند محدودية تنوع قيم السمة الحساسة (سيقتصر تطبيق ℓ -diversity على قيم منخفضة للمعامل ℓ وبالتالي مخاطر مرتفعة للمدعي العام). كذلك تبين أن كفاءة هذا النموذج ترتبط عكسياً بعدد السمات الحساسة المراد حمايتها، حيث تؤدي زيادتها كما في البيانات الصحية إلى تدهور متسارع في فائدة البيانات، نتيجة ارتفاع مستويات التعميم والقمع لتحقيق قيد الخصوصية.

أما في نموذج k -closeness، فقد تأثرت عتبة المقايضة المثلى بدرجة عدم توازن توزيع السمة الحساسة في مجموعة البيانات، حيث حقق مستويات خصوصية وفائدة أقل من k -anonymity، لتركيزه على الحماية من هجمات الكشف عن السمات، دون فرض حد أدنى مباشرة على حجم فئات التكافؤ، أي أنه لا يحمي من هجمات الكشف عن الهوية، وبالتالي يوصى بتطبيقه بالاقتران مع نموذج k -anonymity، لتوفير حماية متكاملة ضد هجمات الكشف عن الهوية وعن السمات، وهذا ما أكدته الدراسة [17].

أخيراً، يوفر مقياس الفائدة المقترح تقييماً أكثر شمولية ووضوحاً يدمج عدة جوانب من الفائدة بدلاً من الاعتماد على مقياس واحد فقط يعكس جانباً واحداً من جوانب الفائدة، أو الاعتماد على مجموعة منفصلة من المقاييس التي قد تبدي نتائج متعارضة متأثرة بالتحويلات التي تم تطبيقها، مما يشنت صنّاع القرار. وبالتالي يوصى باعتماده كأداة موثوقة لاتخاذ القرار عند تقييم فائدة البيانات مُخفاة الهوية، وبناءً على التحليل التجريبي للمقاييس الأربعة المكونة له، توصلنا إلى مجموعة من التوصيات فيما يتعلق بطريقة تحديد الأوزان المناسبة لهذه المقاييس، وفقاً لهدف عملية إخفاء الهوية:

حيث يمنح مقياس التفاصيل وزناً أعلى من باقي المقاييس عندما يكون الهدف تجنب تطبيق مستويات مرتفعة من التعميم، والحفاظ على أكبر قدر من تفاصيل البيانات. أما في حال كان التركيز على الحفاظ على التوزيع

الاحتمالي للبيانات يتم منح مقياس الإنتروبيا غير الموحدة وزناً أعلى. وفي المقابل يمكن منح مقياس إمكانية التمييز وزناً أعلى إذا كان الهدف التركيز على تقليل عدد السجلات المقموعة أثناء عملية إخفاء الهوية. وأخيراً، في حال كانت البيانات تحتوي على سمات رقمية بشكل أساسي، يمكن منح مقياس الخطأ التربيعي على مستوى السمات الوزن الأعلى، لتقدير ضياع المعلومات الناتج عن زيادة المسافة بين القيم الأصلية والقيم المخفاة الهوية بدقة.

References:

- [1] T. Carvalho, N. Moniz, P. Faria, and L. Antunes, "Survey on privacy-preserving techniques for data publishing," *arXiv preprint arXiv:2201.08120*, 2022.
- [2] J. F. Marques and J. Bernardino, "Analysis of Data Anonymization Techniques," in *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2020) - Volume 2: KEOD*, Budapest, Hungary, 2020, pp. 235–241.
- [3] P. R. M. Rao, S. M. Krishna, and A. P. S. Kumar, "Privacy preservation techniques in big data analytics: a survey," *Journal of Big Data*, vol. 5, no. 1, pp. 1–12, 2018.
- [4] Q. Razi, R. Piyush, A. Chakrabarti, A. Singh, V. Hassija, and G.S.S. Chalapathi, "Enhancing data privacy: A comprehensive survey of privacy-enabling technologies," *IEEE Access*, vol. 13, pp. 40354–40385, 2025.
- [5] A. Dalal, K. Kaushik, S.S. Polagani, L. Upadhyay, M. Soni, and Rohini, "Privacy-preserving techniques in data anonymization and masking: Trends and challenges," in *Smart Technologies and Intelligent Computing*. London: CRC Press, 2026, pp. 336-345.
- [6] A. Majeed and S. Lee, "Anonymization techniques for privacy preserving data publishing: A comprehensive survey," *IEEE Access*, vol. 9, pp. 8512–8545, 2021.
- [7] R. Bild, K. A. Kuhn, and F. Prasser, "SafePub: A truthful data anonymization algorithm with strong privacy guarantees," *Proceedings on Privacy Enhancing Technologies*, vol. 2018, no. 1, pp. 67–87, 2018.
- [8] R. Ratra, P. Gulia, and N. S. Gill, "Evaluation of Re-identification Risk using Anonymization and Differential Privacy in Healthcare," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 2, pp. 563–570, 2022.
- [9] B. Becker and R. Kohavi, "Adult Dataset," *UCI Machine Learning Repository*, 1996. [Online]. Available: <https://doi.org/10.24432/C5XW20>
- [10] F. Prasser, J. Eicher, H. Spengler, R. Bild, and K. A. Kuhn, "Flexible data anonymization using ARX—Current status and challenges ahead," *Software: Practice and Experience*, vol. 50, no. 7, pp. 1277–1304, 2020.

- [11] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano. "Heart Disease," *UCI Machine Learning Repository*, 1989. [Online]. Available: <https://doi.org/10.24432/C52P4X>.
- [12] I. E. Olatunji, J. Rauch, M. Katzensteiner, and M. Khosla, "A review of anonymization for healthcare data," *Big Data*, vol. 12, no. 6, pp. 538–555, 2024.
- [13] P. Samarati, "Protecting respondents' identities in microdata release," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010–1027, 2001.
- [14] F. Prasser, F. Kohlmayer, R. Lautenschläger, and K. A. Kuhn, "Arx-A comprehensive tool for anonymizing biomedical data," in *Proceedings of the AMIA 2014 Annual Symposium*, Washington D.C., USA, 2014, pp. 984–993.
- [15] K. El Emam and F. K. Dankar, "Protecting privacy using k-anonymity," *Journal of The American Medical Informatics Association*, vol. 15, no. 5, pp. 627–637, 2008.
- [16] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkatasubramanian, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 3–es, 2007.
- [17] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering (ICDE)*, Istanbul, Turkey, 2007, pp. 106–115.
- [18] W. Liu, Y. Zhang, H. Yang, and Q. Meng, "A survey on differential privacy for medical data analysis," *Annals of Data Science*, vol. 11, no. 2, pp. 733–747, 2024.
- [19] C. Dwork, "Differential privacy," in *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP 2006)*, Venice, Italy, 2006, pp. 1–12.
- [20] J. Domingo-Ferrer and D. Sánchez, "How to Get Actual Privacy and Utility from Privacy Models: the k-Anonymity and Differential Privacy Families," *arXiv preprint arXiv:2510.11299*, 2025.
- [21] A. Boudewijn and A. F. Ferraris, "Legal and regulatory perspectives on synthetic data as an anonymization strategy," *Journal of Personal Data Protection Law*, vol. 17, pp. 17–31, 2024.
- [22] R. Shikhaliyev, "Anonymization of personal medical data based on artificial intelligence," *Problems of Information Society*, vol. 17, no.1, pp. 35–43, 2026.
- [23] F. Prasser, R. Bild, and K. A. Kuhn, "A generic method for assessing the quality of de-identified health data," in *Proceedings of the Medical Informatics Europe Conference*, Munich, Germany, 2016, pp. 312–316.

- [24] A. Gionis and T. Tassa, "k-Anonymization with minimal loss of information," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 2, pp. 206–219, 2009.
- [25] R. J. Bayardo and R. Agrawal, "Data Privacy Through Optimal k-Anonymization," in *Proceedings of the 21st International Conference on Data Engineering (ICDE)*, Tokyo, Japan, 2005, pp. 217–228.
- [26] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez, "T-closeness through microaggregation: Strict privacy with enhanced utility preservation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 11, pp. 3098–3110, 2015.
- [27] S. Malefaki, D. Markatos, A. Filippatos, and S. Pantelakis. "A Comparative analysis of Multi-Criteria Decision-Making methods and normalization techniques in holistic sustainability assessment for engineering applications," *Aerospace*, vol. 12, no. 2, Art. no. 100, 2025.
- [28] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*, 7th ed. Hoboken, NJ, USA: Wiley, 2018.