مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (6) العدد(12) 2022
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (6) No. (12) 2022

تطوير تقنية هجينة للحفاظ على الخصوصية في عملية التنقيب في البيانات

د. راغب طعمة *
تاله نبيه عمران **

(تاريخ الإيداع 2022/6/23 . قُبِل للنشر في 2022/12/22)

□ ملخّص□

في العديد من المنظمات يتم جمع كمية كبيرة من البيانات التي يتم استخدامها في بعض الأحيان من قبل المؤسسات للقيام بمهام التنقيب في البيانات. ومع ذلك ، قد تحتوي البيانات التي تم جمعها على معلومات خاصة أو حساسة يجب حمايتها. تعد حماية الخصوصية مشكلة مهمة إذا أصدرنا تلك البيانات لغرض التنقيب أو المشاركة. تسمح تقنيات الحفاظ على خصوصية بنشر البيانات بينما تحتفظ في الوقت نفسه بالمعلومات الخاصة للأفراد. تم اقتراح العديد من التقنيات للحفاظ على الخصوصية في التنقيب في البيانات ولكنها تعاني من أنواع مختلفة من الهجمات وفقدان المعلومات. في هذا البحث تم تطبيق طريقة هجينة للحفاظ على الخصوصية في استخراج البيانات حيث تحمي البيانات الحساسة مع فقد معلومات أقل و تحقيق مستوى خصوصية أعلى لها مما يزيد من قابلية استخدام هذه البيانات ويمنع أيضاً تعرض البيانات الحساسة لأنواع مختلفة من الهجمات.

الكلمات المفتاحية: تنقيب في البيانات، الحفاظ على الخصوصية، فقدان المعلومات ، مستوى الخصوصية

*مدرس في قسم هندسة تكنولوجيا المعلومات -كلية هندسة تكنولوجيا المعلومات والاتصالات -جامعة طرطوس -سوريا.
**طالبة ماجستير في قسم هندسة تكنولوجيا المعلومات -كلية هندسة تكنولوجيا المعلومات والاتصالات -جامعة طرطوس -سوريا.

مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (6) العدد(12) 2022
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (6) No. (12) 2022

Develop a hybrid method to privacy preservation in the data mining process

Dr.Ragheb Toama *
Eng.Tala Omran **

(Received 23/6/ 2022 . Accepted 22/12/ 2022)

□ ABSTRACT

In many organizations a large amount of data is collected which is sometimes used by organizations to perform data mining tasks. However, the data collected may contain private or sensitive information that must be protected. Privacy protection is an important issue if we release that data for the purpose of mining or sharing. Privacy technologies allow data to be released while at the same time preserving private information for individuals. Many techniques have been proposed to maintain privacy in data mining but suffer from different types of attacks and information loss. In this research, a hybrid privacy-preserving method has been applied in extracting data, where it protects sensitive data with less information loss and a higher level of privacy, which increases the usability of this data and also prevents sensitive data from being exposed to different types of attacks.

Key Words: Data mining, Privacy preserving, information loss , Privacy level

*Lecturer, InformationTechnology Engineering Department, Information and communication Technology Engineering, Tartous University, Syria.

** Master student, Information Technology Engineering Department, Information and communication Technology Engineering, Tartous University, Syria

1- المقدمة :

التنقيب في البيانات(data miming) هو عملية طرح استفسارات واستخراج أنماط و اتجاهات مفيدة من كمية كبيرة من البيانات حيث يتم تنفيذ هذه العملية باستخدام تقنيات مختلفة مثل التعرف على الأنماط والتعلم الآلي، حالياً أصبح هناك كميات ضخمة من البيانات المتاحة في كل منظمة و مؤسسة مع التطور التكنولوجي المتزايد التي تحتاج لاستخراج المعلومات المخفية القّيمة من خلال عملية التنقيب في البيانات. يمكن استخدام هذه البيانات التي تم جمعها في بعض المجالات المختلفة مثل الأعمال التجارية ، والرعاية الصحية، والأمن السيبراني ، في الخطوة الثانية ، والعملية المهمة هي أنه عند جمع هذه البيانات ، يجب أن يتم استخراج المعرفة المفيدة من المعلومات الأولية [1].
تحتوي مجموعات البيانات على معلومات فردية حساسة و مهمة للأفراد والتي قد تتعرض لهجومات من قبل أطراف خارجية أثناء القيام بعملية التنقيب ،الأمر الذي قد يخلق حاجزاً لإنجاز عملية التقيب بنجاح ، أصبحت حماية خصوصية الأفراد و بياناتهم الحساسة قضية مهمة إذا تم إصدار هذه البيانات لغرض التنقيب أو المشاركة، مما استدعى لظهور تقنيات التي تسمح بنشر ومشاركة بيانات الأفراد للقيام بعملية التنقيب واستخراج الأنماط والمعارف منها بينما تحتفظ في الوقت نفسه بسرية المعلومات الخاصة للأفراد والتي تدعى بتقنيات الحفاظ على خصوصية في عملية التنقيب عن البيانات PPDM(Privacy Preservation Data Mining) هي تقنيات تقوم بتعديل البيانات مع المحافظة على تنفيذ خوارزميات التنقيب في البيانات بفعالية دون المساس بأمن المعلومات الحساسة الموجودة فيها مع المحافظة على الهدف الأساسي من عملية التنقيب .

من منظور التنقيب في البيانات ، يتم تصنيف الخصوصية إلى أربعة وجهات نظر. هذا التصنيف على النحو التالي[1]:

1. الخصوصية في وقت جمع البيانات قبل التنقيب عن البيانات.
2. الخصوصية في وقت نشر البيانات
3. بعد الانتهاء من عملية خوارزميات التنقيب عن البيانات.
4. الخصوصية أثناء توزيع البيانات

تم الاعتماد في هذا البحث على مجموعة من الدراسات السابقة التي قامت كل منها باقتراح طريقة للحفاظ على الخصوصية لتحقق الغاية المرجوة وهي مستوى خصوصية عالي و فقدان بيانات أقل :

في البحث [2] تم تطبيق خرائط التنظيم الذاتي بالاشتراك مع خوارزميات إخفاء الهوية المستندة إلى المجموعات للحصول على المزيد من البيانات، بالإضافة للمقارنة بين تقنيات إخفاء الهوية حيث تم تطبيق كل من تقنية القمع لسمتين ثم تطبيقها على ثلاث سمات لمجموعة بيانات للأفراد البالغين و قياس فعالية كل منهما لمهام التنقيب عن البيانات ، بالإضافة تحليل شامل لتأثيرات معلمات الخصوصية وبعض جوانب مجموعات البيانات على عملية إخفاء الهوية.

في البحث[3] تم وصف المشكلات الخاصة بمجموعة البيانات و تم اقتراح طريقة هجينة للحفاظ على خصوصية البيانات أثناء عملية التنقيب ،في البداية تم تقسيم سمات قاعدة البيانات إلى سمات حساسة و سمات شبه حساسة ثم تطبيق تقنية Randomization على السمات شبه الحساسة ، ليتم بعدها تطبيق تقنية Perturbation على البيانات الجديدة ، حقق هذا النهج نتائج أفضل من ناحية زمن التنفيذ و الخصوصية من خلال مقارنته مع عمل سابق.

في البحث[4] تم تطبيق نهج يتم من خلاله إرسال البيانات في شكل مضطرب بحيث لا يمكن كشف هوية المستخدمين أو المعلومات الحساسة. في النظام المقترح، تم تطبيق تقنية الاضطراب Perturbation Technique على البيانات الحساسة للمستخدمين من خلال استخدام تقنية الاضطراب الإضافي حيث يتم استبدال البيانات الأصلية X بـ Y = X + R بالإضافة تم استخدام بروتوكول المصادقة بناءً على MAC الخاص بجهاز الاستقبال بحيث لا يتمكن سوى المستخدم المصدق عليه من الوصول إلى البيانات. إنها العملية التي يتم فيها تشويش البيانات الموجودة في المستند وتأمينها بواسطة مفتاح السر. في هذه التقنية، سيتم استرداد البيانات الأصلية بعد المصادقة فقط عندما يتطابق المفتاح والملف وعنوان MAC المصدق. تظهر نتيجة التقييم تقليل فقدان المعلومات أثناء إعادة بناء البيانات الأصلية. قدمت الدراسات السابقة تقنيات مختلفة لتحسين التقنيات المستخدمة في الحفاظ على الخصوصية أثناء عملية التنقيب عن البيانات(PPDM) لم تحقق أي من الطرق المذكورة سابقاً نتيجة مثالية، مما يجعل باب التحسين على التقنيات المستخدمة في PPDM مفتوحاً للمزيد من التحسينات .

2- أهمية البحث و أهدافه :

مع التطور التكنولوجي أصبح هناك وفرة متزايدة من البيانات المتاحة في كل منظمة و مؤسسة التي تحتاج غالباً لاستخراج المعلومات المخفية القيّمة من خلال عملية التنقيب في البيانات (data mining) مع تزايد شعبية وتطور تقنيات التنقيب في البيانات التي باتت تجلب تهديداً خطيراً لأمن البيانات الحساسة للفرد. ظهرت الحاجة لتحسين تقنيات الحفاظ على خصوصية وسرية البيانات أثناء عملية التنقيب فيها بشكل يراعي عدة مقاييس. يقدم هذا البحث مساهمة جديدة في مجال الحفاظ على الخصوصية في التنقيب عن البيانات، من حيث اقتراح طريقة هجينة قادرة على تحسين عملية الحفاظ على خصوصية و سرية البيانات أثناء عملية التنقيب فيها، حيث سوف يتم دمج عدة تقنيات و قياس فعاليتها من الخصوصية (Privacy) وزمن التنفيذ (Execution time) .

3- طرائق البحث ومواده:

مجموعة البيانات هي جزء أساسي يتم استخدمه في أبحاث الحفاظ على خصوصية البيانات حيث يعتمد البحث على:

3-1قاعدة البيانات المستخدمة و سماتها:

مجموعة البيانات المستخدمة في هذا البحث و التي سيتم تطبيق تقنيات الحفاظ على الخصوصية المقترحة عليها هي مجموعة بيانات استهلاك الكحول للطلاب من أرشيف التعلم الآلي لـ(Fabio Pagnotta 2016) UCI (UC Irvineهو عبارة عن مستودع لقواعد بيانات يتم استخدمها في مجتمع التعلم الألي). تأتي البيانات الأصلية من استطلاع أجراه أستاذ في هولندا. كان السبب الرئيسي لهذه البيانات هو رؤية آثار الشرب على درجات الطلاب في مادة الرياضيات. تتكون مجموعة البيانات هذه من 33 سمة كما هو موضح في الجدول:

الجدول(1): سمات مجموعة البيانات

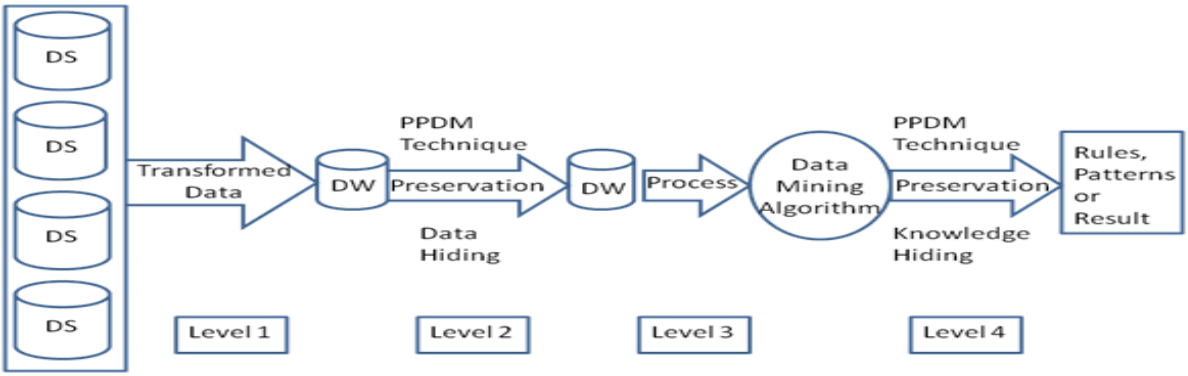
Description	Attribute
student's school	School
student's sex	Sex
student's age	Age
student's home address	Address
family size	Famsize
parent's cohabitation status	Pstatus
mother's education	Medu
father's education	Fedu
mother's job	Mjob
father's job	Fjob
reason to choose this school	reason
student's guardian	guardian
home to school travel	traveltime

Description	Attribute
home to school travel	traveltime
number of past class failures	failures
extra educational support	schoolsup
family educational support	famsup
extra paid classes within the course subject	paid
extra-curricular activities	activities
attended nursery school	nursery
wants to take higher education	higher
Internet access at home	internet
with a romantic relationship	romantic
quality of family relationships	famrel

Description	Attribute
free time after school	freetime
going out with friends	goout
workday alcohol consumption	Dalc
weekend alcohol consumption	Walc
current health status	health
number of school absences	absences
Grade1	G1
Grade2	G2
Grade3	G3

2-3 التقنيات المستخدمة في البحث :

يوضّح الشكل (1) إطار عمل PPDM، في البداية يتم تجميع البيانات من مصادر مختلفة ومعالجتها مسبقاً. يتم تخزين البيانات المعالجة في مستودع البيانات. في المستوى (2) يتم تطبيق تقنيات إخفاء البيانات لتوفير الخصوصية للبيانات، ثم يتم استخدام خوارزميات التنقيب في البيانات للعثور على الأنماط واكتشاف المعرفة من البيانات في المستوى(3). بعد عملية التنقيب في البيانات ، في المستوى 4 ، يتم تطبيق تقنيات الحفاظ على الخصوصية على نتائج التنقيب عن البيانات لحمايتها من الوصول غير المصرح به[5].



الشكل (1) : إطار عمل PPDM

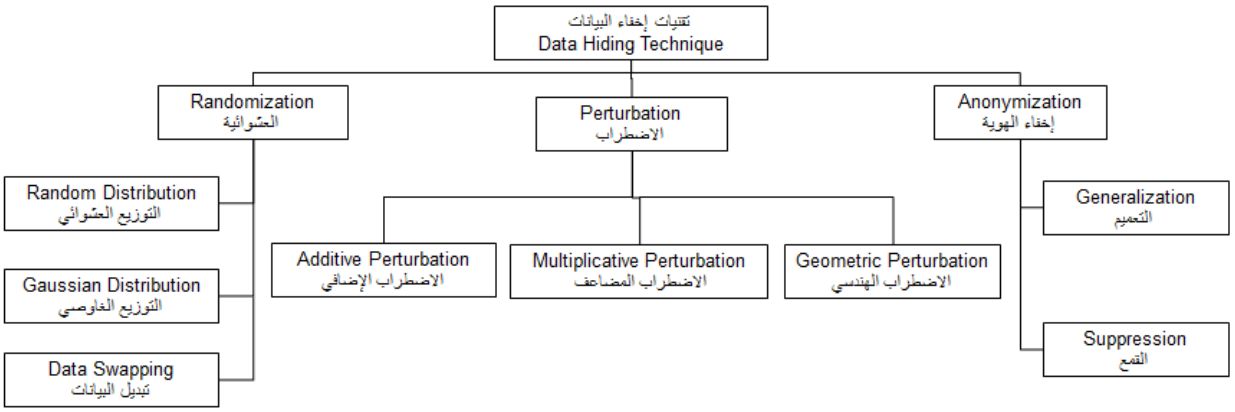
DS(Data set) : قاعدة البيانات – DW(Data Warehouse): مستودع البيانات .

يمكن تقسيم تقنيات PPDM إلى مجموعتين :

أ.تقنيات إخفاء البيانات (Data Hiding Techniques) : في هذه التقنية، يتم تغيير البيانات المقدمة لمهمة التنقيب عن البيانات قصها أو حظرها بطريقة لا تعرض المعلومات الحساسة الموجودة فيها لأطراف أخرى من خلال تقنيات متعددة(Perturbation Technique Randomization Technique– Anonymization Technique).

ب. تقنيات إخفاء المعرفة (Knowledge Hiding Techniques): في هذه التقنية، يتم استبعاد نتائج عملية التنقيب (تسمى المعرفة) من الاستخدام والتي تعد حساسة للأفراد كونها تشير إلى بيانات و معلومات سرية مثل الحساب البنكي المصرفي للأفراد . هذه التقنيات مهمة للغاية لأن المعرفة الحساسة المستخرجة من عملية التنقيب في البيانات يمكن استخدامها لاستخلاص معلومات سرية. توجد طرق مختلفة لتنفيذ هذه التقنيات.

في هذا البحث تم الاعتماد على تقنيات إخفاء البيانات بما أن حماية البيانات الشخصية و الحساسة للأفراد من الاختراق و السرقة من قبل أطراف خارجية هي الهدف الأساس لبحثنا. يوضّح الشكل (2) الطرق المختلفة لتنفيذ تقنيات إخفاء البيانات:



الشكل (2) : تقنيات PPDM

1- تقنية إخفاء الهوية Anonymization Techniques: يشير إخفاء هوية البيانات(و يقصد بها إخفاء معلومات خاصة بالأفراد مثل الاسم و الجنس...) إلى طريقة الحفاظ على المعلومات الخاصة أو السرية عن طريق حذف أو تشفير المعرفات التي تربط الأفراد بالبيانات المخزنة. يتم ذلك لحماية النشاط الخاص لفرد أو شركة مع الحفاظ على مصداقية البيانات التي يتم جمعها وتبادلها. تعتمد تقنيات إخفاء الهوية على تقنية القمع (Suppression) وتقنية التعميم (Generalization). في طريقة إخفاء الهوية القائمة على مبدأ القمع ، يتم تكوين مجموعات فرعية من سجلات البيانات الأصلية عن طريق إخفاء قيم بعض السمات المختارة جيدًا (بناءً على غرض التنقيب)، ويتم هذا الإخفاء باستخدام بعض القيم الخاصة مثلاً يتم استبدال كل حرف بقيمة *. وفي طريقة إخفاء الهوية القائمة على مبدأ التعميم، يتم تكوين مجموعات فرعية من سجلات البيانات الأصلية عن طريق استبدال القيم الأصلية بقيم أكثر عمومية في قاعدة البيانات ، مثل استبدال قيم (طرطوس- دمشق) بالقيمة (سوريا)

2- اضطراب البيانات Data Perturbation : في هذه التقنية، يتم تعديل البيانات المتاحة قبل أن يتم تمريرها إلى خوارزمية التنقيب في البيانات . هناك عدد من الطرق لتعديل البيانات مثل تقنية إضافة الضوضاء، حيث يضيف مالك البيانات بعض الأرقام العشوائية (الضوضاء) للبيانات . يُستمد هذا الرقم العشوائي عمومًا من التوزيع الطبيعي بمتوسط صفر وانحراف معياري صغير متعلق بالبيانات التي يتم العمل عليها ، مما يحافظ على إحصائيات البيانات الأصلية. ثم يشارك مالك البيانات هذه البيانات الضوضائية لمهمة التنقيب في البيانات. يقوم منقب البيانات بإعادة توزيع مجموعة البيانات الأصلية. لكن منقب البيانات لا يمكنه استرداد قيم البيانات الفعلية. يمكّن هذا خوارزمية التنقيب في البيانات من بناء نتيجة أكثر دقة دون الكشف عن البيانات الفعلية.

3-العشوائية (Randomization) :تعتبر تقنية التوزيع العشوائي طريقة جيدة للحفاظ على خصوصية البيانات. يجب تنفيذ عملية التوزيع العشوائي للبيانات لضمان أداء خوارزمية التنقيب في البيانات بالإضافة إلى الحفاظ على الخصوصية. يحمي هذا النهج بيانات العملاء من خلال السماح لهم بتعديل سجلاتهم تعتمد هذه التقنية على التوزيع العشوائي(Random Distribution) أو توزيع الغاوصي (Gaussian Distribution) أو تبادل البيانات (Data Swapping) وهي التقنية المستخدمة في بحثنا كونها تؤمن فقدان بيانات أقل و خصوصية جيدة بناءً على الدراسات السابقة [4].

4.الخوارزمية المقترحة :

تم اقتراح منهجية تعتمد على دمج تقنية تبادل البيانات Data Swapping وتقنية إضافة الضجيج Adding noise وتقنيات إخفاء الهوية Anonmyization . الفكرة الأساسية هي استخدام هذه التقنيات بشكل هجين مع بعضها للمرة الأولى و دراسة تأثير دمجها على كل من خصوصية البيانات و الزمن اللازم لتنفيذها و مقارنتها مع دراسة سابقة ،تم تنفيذ منهجية PPDM بالخطوات التالية:

4-1 تقسيم سمات قاعدة البيانات إلى أربع مجموعات(Categorize the attributes into Four categories):

يُفترض أن مراقبي البيانات الذين يحتاجون إلى اتخاذ احتياطات الخصوصية من أجل منع انتهاكات البيانات موثوق بهم ولديهم التزامات قانونية ، يخزن ويستخدم البيانات التي تم جمعها من التطبيقات الرقمية

باستخدام الأساليب المناسبة ، ومشاركتها عن طريق إخفاء الهوية عند الضرورة. في أنظمة PPDM يتم تصنيف البيانات المجمعة إلى أربع مجموعات[6]:

a) المعرفات (ID:Identifier): تحتوي على معلومات تحدد هوية الأفراد بشكل فريد ومباشر مثل الاسم الكامل ورقم الضمان.

b) سمات شبه حساسة (Quasi Attribute): السمات التي تؤدي إلى التعريف غير المباشر للفرد. هذه السمات هي بيانات غير فريدة مثل الجنس والعمر والرمز البريدي.

c) السمات الحساسة (Sensitive Attribute): هي السمات التي تحتوي على بيانات خاصة وحساسة للأفراد، كالمرض والراتب.

d) السمات غير الحساسة (Non Sensitive Attribute): هي السمات التي تحتوي على بيانات عامة وغير سرية لا تغطيها سمات أخرى مثل الدرجة العلمية للأم .....

يوضّح الجدول(2) تصنيف السمات الموجودة في مجموعة البيانات :

الجدول(2):تصنيف السمات

السمة	المجموعة التي تنتمي لها السمة
الاسم	المعرفات
العنوان، المدرسة، الجنس	سمات شبه حساسة
حجم العائلة، حالة الوالدين، تعليم الأم، تعليم الأب، عمل الأم، عمل الأب، سبب اخيار المدرسة، ولي أمر الطالب، مدة السفر من المنزل إلى المدرسة عدد مرات الفشل، دعم تعليمي إضافي، دعم الأسرة التربوي، فصول مدفوعة، الأنشطة، حضانة ، يريد الالتحاق بالتعليم العالي، الانترنت في المنزل، في علاقة رومانسية، جودة العلاقة الأسرية، وقت الفراغ، الخروج مع الأصدقاء، استهلاك الكحول في أيام العمل، استهلاك الكحول في العطلة، الصحة، عدد الغيابات،	سمات غير حساسة
درجة الفصل الأول، درجة الفصل الثاني، درجة الفصل الصيفي	سمات حساسة

بما أن هدف تقنيات PPDM هو إخفاء هوية الأفراد، فإن النهج المقترح سوف يتم تطبيقه على السمات شبه الحساسة (Quasi Attribute) كونها تمثل البيانات التي تحدد هوية الفرد حسب التصنيف السابق و التي سوف يُنفذ تباعاً في الخطوات اللاحقة .

4-2 تطبيق تقنية Randomization على سمات العنوان و المدرسة و الجنس (Randomize the school address, sex fields) :

من تقنيات العشوائية Randomization المستخدمة في البحث هي مبادلة البيانات Data swapping ،في تقنية مبادلة البيانات، يمكن تحقيق حماية السرية من خلال التبادل الانتقائي لمجموعة فرعية من قيم السمات بين أزواج السجلات المختارة مثل مبادلة قيم سمة المدرسة بين الأسطر بأسلوب عشوائي . تحافظ مبادلة البيانات على خصوصية المعلومات الحساسة الأصلية المتاحة على مستوى السجل. إذا تم اختيار السجلات بشكل عشوائي لكل عملية مبادلة، فإنها تسمى مقايضات عشوائية. يصعب على الدخيل التعرف على شخص أو كيان معين في قاعدة البيانات ، لأن جميع السجلات يتم تعديلها إلى الحد الأقصى. من نقاط القوة لتقنية المبادلة هي أنها بسيطة. يوضّح الشكل(3) و الشكل (4) الفرق في مجموعة البيانات قبل و بعد تطبيق تقنية تبادل البيانات على السمات المختارة

address	age	sex	school	
Ubran	18	Female	Mousinho da Silveira	1
Rural	17	Male	Gabriel Pereira	2
Ubran	15	Female	Mousinho da Silveira	3
				4

الشكل(3):مجموعة البيانات بعد تطبيق data swapping

address	age	sex	school	
Urban	18	Female	Gabriel Pereira	1
Urban	17	Female	Gabriel Pereira	2
Urban	15	Female	Gabriel Pereira	3
				4

الشكل(4):مجموعة البيانات قبل تطبيق data swapping

4-3 تطبيق الاضطراب على سمة العمر من خلال إضافة قيمة مقدرة للضجيج (Perturb the age by adding the appropriate noise):

الاضطراب هو إضافة ضجيج بطريقة مشروطة على البيانات الأصلية. في هذا البحث تم تطبيق تقنية الضجيج الإضافي Additive Noise حيث يتم استبدل البيانات الأصلية X بـ Y = X + R. حيث R عبارة عن قيمة ضجيج يتم إنشاؤها بشكل مستقل عن متجه عشوائي، وتم تطبيقها على سمة العمر بحيث أن R=5(هي الفرق بين أعلى قيمة للعمر و أقل قيمة) وتم اختيارها على أنها قيمة وسطية مناسبة لإخفاء قيم سمة العمر مع المحافظة على العمر المنطقي لطلاب المدارس و بالتالي تحقيق مقدار ثقة عالي للمستخدم[7]. يوضّح الشكل (5) و الشكل (6) يوضح الفرق في مجموعة البيانات قبل و بعد تطبيق تقنية الاضطراب الإضافي على السمة المختارة.

address	age	sex	school	
Ubran	23	Female	Mousinho da Silveira	1
Rural	22	Male	Gabriel Pereira	2
Ubran	20	Female	Mousinho da Silveira	3
				4

الشكل(5): مجموعة البيانات بعد تطبيق تقنية الضجيج

address	age	sex	school	
Ubran	18	Female	Mousinho da Silveira	1
Rural	17	Male	Gabriel Pereira	2
Ubran	15	Female	Mousinho da Silveira	3
				4

الشكل(6): مجموعة البيانات قبل تطبيق تقنية الضجيج

4-4إخفاء هوية سمة الجنس بتقنية التعميم(Anonymize the sex field with generalization technique) :

generalization على سمة الجنس

address	age	Sex	school	
Ubran	23	Person	Mousinho da Silveira	1
Rural	22	Person	Gabriel Pereira	2
Ubran	20	Person	Mousinho da Silveira	3
				4

الشكل(7):مجموعة البيانات بعد تطبيق تقنية generalization

address	age	sex	school	
Ubran	23	Female	Mousinho da Silveira	1
Rural	22	Male	Gabriel Pereira	2
Ubran	20	Female	Mousinho da Silveira	3
				4

الشكل(8):مجموعة البيانات قبل تطبيق تقنيةgeneralization

4-5 إخفاء هوية سمة المدرسة بتقنية القمع(Anonymize the school field with suppression technique): الطريقة المستخدمة لإخفاء هوية سمة المدرسة هي القمع (Suppression) عن طريق استخدام احتمالية الحدوث، حيث سوف يتم استبدال قيم school بقيمة عددية تعبر عن احتمالية ظهور كل قيمة من قيمها:

47 Mousinho da Silveira ← 349 Gabriel Pereira

يوضّح الشكل (9) و الشكل (10) الفرق في مجموعة البيانات قبل و بعد تطبيق تقنيةSuppression على سمة المدرسة:

address	age	Sex	school	
Holand	23	Person	Mousinho de Siliveira	2
Holand	22	Person	Gabriel Pereira	3
Holand	20	Person	Mousinho de Siliveira	4

address	age	Sex	school	
Holand	23	Person	47	2
Holand	22	Person	349	3
Holand	20	Person	47	4

الشكل(9): مجموعة البيانات قبل تطبيق تقنية Suppression **الشكل(10):مجموعة البيانات بعد تطبيق تقنية Suppression**

4-6 إخفاء هوية حقل العنوان بتقنية القمع والتعميم Anonymize the address field with suppression& generalization technique:

لا يمكن تطبيق تقنية القمع (Suppression) على سمة العنوان لعدم وجود جذر مشترك بين بيانات سمة العنوان حيث قيم سمة العنوان ضمن مجموعة البيانات هي (Urban، Rural) لا يوجد جذر مشترك بين القيميتين لذلك تم القيام بدمج كل من تقنية القمع (Suppression) و تقنية التعميم (generalization) على السمة المختارة لتحقيق عملية حفاظ على خصوصية عالية من خلال تطبيق تقنية التعميم (generalization) على قيم السمة لنحصل على قيمة جديدة عمومية تمثلهما لتكن (Holand) و ثم تطبيق تقنية القمع (Suppression) لنحصل على (*****). يوضح الشكل (11) والشكل (12) الفرق في مجموعة البيانات قبل و بعد تطبيق تقنية القمع (Suppression) و تقنية التعميم (generalization) على سمة العنوان

address	age	Sex	school	
*****	23	Person	47	2
*****	22	Person	349	3
*****	20	Person	47	4

address	age	Sex	school	
Holand	23	Person	47	2
Holand	22	Person	349	3
Holand	20	Person	47	4

الشكل(11): مجموعة البيانات قبل تطبيق تقنية (Sup&Gen) **الشكل(12): مجموعة البيانات بعد تطبيق تقنية (Sup&Gen)**

4-7التحقق من الخصوصية المحفوظة و فقدان البيانات و زمن التنفيذ Check Privacy Preserved ,Information loss and execution time:

4-7-1 الخصوصية المحفوظة Privacy Preserved هو نطاق القيم المستخلص من إحصائيات العينة، والذي من المحتمل أن يحتوي على قيمة عينة غير معروفة. يتطلب حساب مقياس الخصوصية معرفة ثلاث بارامترات للعينة وهي : القيمة المتوسطة، الانحراف المعياري وحجم العينة المتمثلة بالقانون :

$$P_{ri} = \bar{x} + z\frac{s}{\sqrt{n}}$$

حيث أن:

$\bar{x}$ –تعبر عن المتوسط الحسابات للعينات

S– تعبر عن الانحراف المعياري

N – تعبر عن حجم العينة

Z – تعبر عن مستوى الثقة (يشير إلى احتمالية تقدير السمة الإحصائية في مسح العينة صحيحاً) ، يحدد الباحث قيمة وسطية لمستوى الثقة (90% أو 95% أو 99%)، تقدّر في هذا النهج *Z=95%* و ذلك لأن البيانات التي يتم العمل عليها إحصائية و إمكانية الخطأ في عملية الإدخال ممكنة .

4-7-2 زمن التنفيذ Execution time: هو زمن تنفيذ الخوارزميات المستخدمة في هذا النهج، ويقدّر بالثواني. يتم تنفيذ النهج للمرة الأولى ثم تكرار هذا الإجراء تسع مرات إضافية و أخذ متوسط زمن تنفيذ المرات (التكرارات) العشرة في الاعتبار هنا للمقارنة بناءً على الدراسة التي تم مقارنة النتائج معها[2].

4-7-3 فقدان البيانات Information Loss (IL): يقيس الاختلاف بين البيانات الأصلية و البيانات المخفية (البيانات بعد تطبيق تقنيات PPDM). مجموعة البيانات التي تم العمل عليها تتضمن نوعين من البيانات :البيانات المستمرة (العددية) و البيانات الفئوية (البيانات النصية) حيث كل منها تشكل حالة لقياس فقدان البيانات[7] .

- قياس IL للبيانات المستمرة من خلال تطبيق القانون :

$$IL = \frac{\left(\sum Orginal\ values - \sum New\ Values\right)^2}{Orginal\ values + New\ values}$$

Original values(البيانات الأصلية قبل التعديل)، New values(البيانات المعدلة).

- قياس IL للبيانات الفئوية :

1- بناء الشجرة الهرمية للبيانات الفئوية التي تم تطبيق تقنيات Anonymization عليها .

2- قياس IL للبيانات الفئوية (Categorical data) التي تم تطبيق تقنيات Anonymization من خلال القانون :

$$IL_{qi} = \frac{lqi}{n} \qquad 0 \le i \le n$$

حيث :

n عدد المستويات الكلي في شجرة الهرمية

lqi هو المستوى الذي أصبحت فيه العينة غير قابلة للتمييز

3- بناء الجدول لبيان القياسات المطبقة على البيانات المخفية Anonymized data

4- حساب المتوسط الحسابي لنسب فقدان البيانات لكل عينة من البيانات المختلفة إلى عدد مرات الحدوث (مجموع البيانات):

$$IL_{\text{categorical data}} = \text{Avg } IL/QI$$

حيث أن QI إجمالي عدد مرات التكرار

5- حساب فقدان البيانات الكلي وهو مجموع فقدان البيانات للبيانات المستمرة و فقدان البيانات للبيانات الفئوية

$$IL_{\text{total}} = IL_{\text{continuous data}} + IL_{\text{categorical data}}$$

5- النتائج و المناقشة:

تتكون مجموعة البيانات المستخدمة في الاختبار من *396* سجل. تم تطبيق التقنيات المقترحة باستخدام لغة Java ضمن بيئة عمل Netbeans و بالإضافة لاستخدام بيئة عمل Matlab. سوف يتم دراسة تأثير النهج المقترح على مجموعة البيانات بإجراء القياسات التالية :

1- خصوصية البيانات Privacy 2- زمن التنفيذ Execution time 3- فقدان البيانات Information Loss

5-1 دراسة تأثير النهج المقترح على خصوصية البيانات Privacy :

تم تطبيق قانون الخصوصية $P_{ri} = \bar{x} + z\frac{s}{\sqrt{n}}$ ، على كل سمة من السمات شبه الحساسة التي تم تطبيق تقنيات النهج المقترح عليها

من خلال حساب المتوسط الحسابي لقيم كل سمة (العمر(Age)، العنوان(Address)، الجنس(Sex)، المدرسة(School))و الانحراف المعياري لها ثم تم أحساب الجذر التربيعي لحجم العينة مع الأخذ بعين الاعتبار القيمة المقدّرة لمستوى الثقة Z لنحصل على النتائج الموضحة في الجدول(3) :

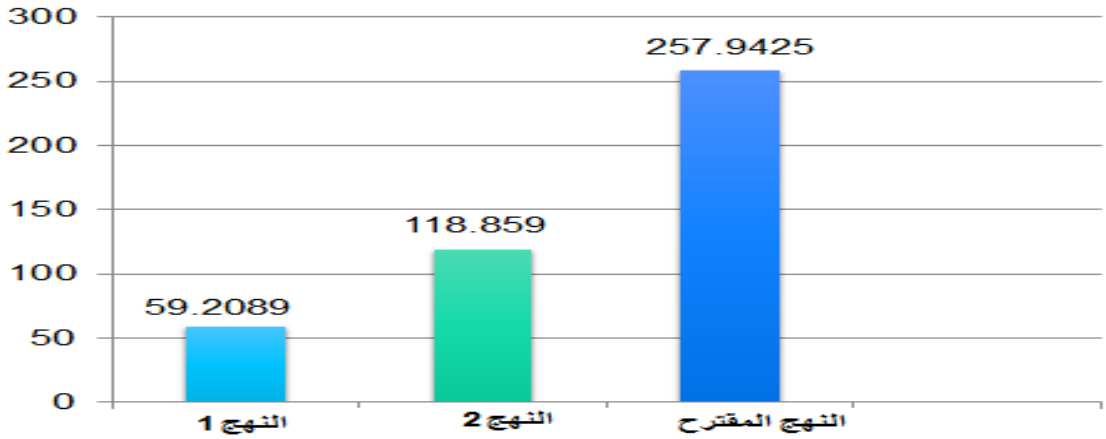
الجدول(3): مستوى الخصوصية (privacy) لكل سمة

P_{School}	P_{Sex}	$P_{Address}$	P_{Age}
287.2	471	252	21.57

لقياس P_{Total} نطبق القانون : $$P_{Total} = \frac{P_{Age} + P_{Address} + P_{School} + P_{Sex}}{4}$$

بالحساب تبيّن أن $P_{Total} = 257.9425$ ويعود ذلك لاستخدام أربع تقنيات و بالتالي رفع مستوى الخصوصية.

يبين الشكل (13) مقارنة النتائج التي حصلنا عليها مع النتائج السابقة في الدراسة[2]:



الشكل (13) : مقارنة بين النهج المقترح و نُهج سابقة من حيث خصوصية البيانات

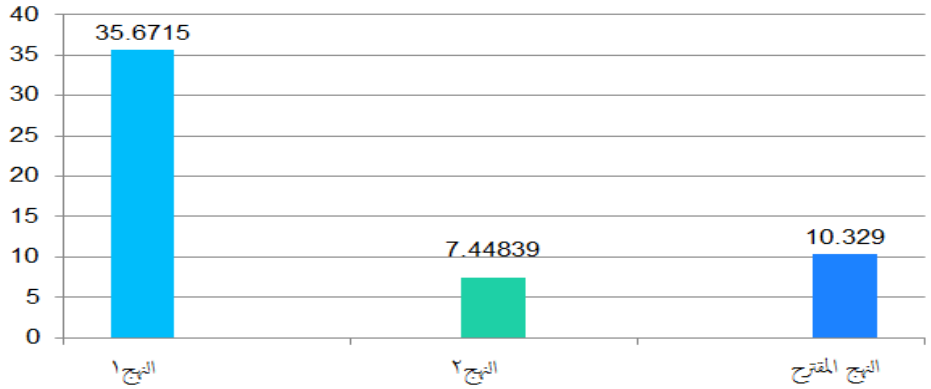
5-2 دراسة تأثير النهج المقترح على زمن التنفيذ *Execution time*:

سوف يتم إجراء 10 تكرارات و أخذ متوسط هذه التكرارات العشرة في الاعتبار هنا للمقارنة، يبين الجدول(4) زمن التنفيذ في كل تكرار [8] :

الجدول(4): الزمن في كل تكرار

التكرار	1	2	3	4	5	6	7	8	9	10	$Time_{total}$
الزمن	10.35	10.33	10.36	10.32	10.34	10.32	10.29	10.31	10.34	10.33	10.329

يبين الشكل (14) مقارنة النتائج التي حصلنا عليها مع النتائج السابقة في الدراسة[2]:



الشكل (14) : مقارنة بين النهج المقترح و نُهج سابقة من حيث زمن التنفيذ

من خلال هذه النتائج تبين أن النهج المقترح المستخدم في هذا البحث ، كان له تأثير في زيادة زمن التنفيذ بالمقارنة مع نُهج سابقة و يُعزى ذلك لعدد التقنيات التي تم تطبيقها في هذا النهج و التي تبلغ أربع تقنيات .

5-3 دراسة تأثير النهج المقترح على فقدان البيانات Information Loss:

5-3-1 قياس *IL* للبيانات المستمرة (سمة العمر *Age*) من خلال تطبيق القانون [9] :

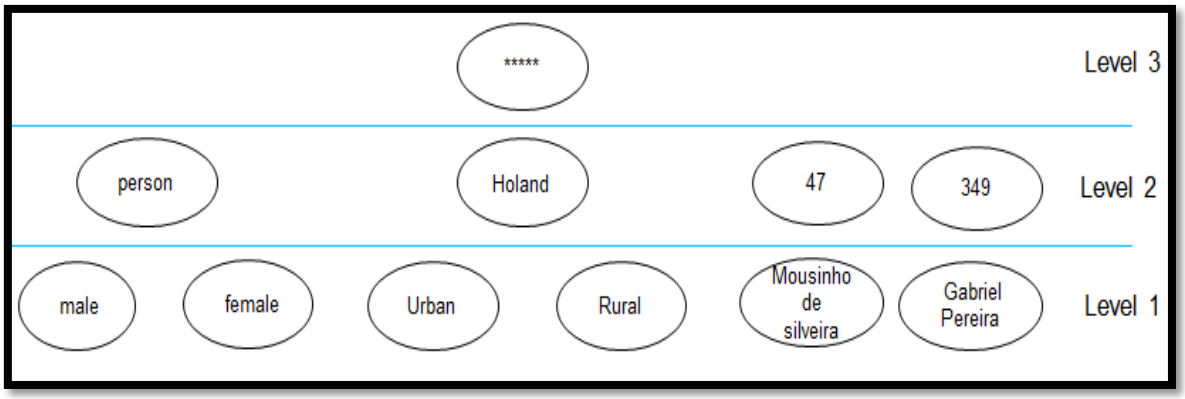
$$IL = \frac{(\sum Orginal\ values - \sum New\ Values)^2}{Orginal\ values + New\ values}$$

$$IL_{Age} = \frac{((15+16+17+18+19+20+21+22)-(20+21+22+23+24+25+26+27))^2}{(15+16+17+18+19+20+21+22)+(20+21+22+23+24+25+26+27)}$$

$$IL_{Age} = \frac{(148-188)^2}{148+188} = 0.47$$

5-3-2 قياس IL للبيانات الفئوية من خلال تطبيق الخطوات التي تم ذكرها سابقاً :

1- يبين الشكل (15) الشجرة الهرمية للبيانات الفئوية التي تم بناؤها بناء على المستوى الذي تم إخفاء هوية السمة فيه [10]:



الشكل (15):الشجرة الهرمية للبيانات الفئوية

2- قياس IL للبيانات الفئوية في كل حالة (IL Per Occur) من خلال القانون[10] :

$IL_{qi}=\frac{lqi}{n}$ ثم حساب IL الكلية لكل حالة في مجموعة البيانات من خلال تطبيق القانون Total IL=Number of occur*ILPer Occur، يبين الجدول (3) القياسات التي تم حسابها لقياس IL :

الجدول(3):جدول توضيحي لقياس IL

Original value	Anonymized value	Number of occur	IL per occur	Total IL
Male	Person	185	2/2=1	185
Female	Person	207	2/2=1	207
Urban	******	307	2/3=0.33	101.31
Rural	******	86	2/3=0.33	28.38
Mousinho de silveira	47	47	2/2=1	47
Gabriel Pereira	348	348	2/2=1	348
		Total=1180		Total=916.69

3- حساب المتوسط الحسابي لنسب فقدان البيانات لكل عينة من البيانات المختلفة إلى عدد مرات الحدوث (مجموع البيانات):

Avg IL/QI=916.69/1180=0.776

4-حساب فقدان البيانات الكلي وهو مجموع فقدان البيانات للبيانات المستمرة و فقدان البيانات للبيانات الفئوية :

Total IL=0.47+0.776=1.246

تقييم النتائج :

مما سبق نجد أنه :

- ساهمت البنية المقترحة في الحفاظ على خصوصية البيانات حيث:

✓ بعد تطبيق تقنية Randomization و Perturbation، يتغير شكل البيانات الأصلي بسبب التبديل الذي تم على سجلات الأفراد في مجموعة البيانات و أيضاً بسبب الاضطراب الإضافي الذي تم تطبيقه على البيانات المبدلة، و بالتالي إن إعادة تحديد الهوية عن طريق ربط السجل أو خوارزميات المطابقة يكون أصعب وغير مؤكد، حتى عندما يكون الدخيل قادراً على إعادة تحديد الهوية ،لا يمكنه أن يكون واثقاً من أن البيانات التي تم الكشف عنها متوافقة مع الأصل.

✓ بعد تطبيق تقنية Anonymization، أصبحت البيانات المعدّلة سابقاً مخفية مما يمكّن من نقل المعلومات بشكل آمن مع تقليل مخاطر تسريب المعلومات و البيانات الشخصية للأطراف الخارجية.

- يعتبر زمن تنفيذ النهج المقترح ليس صغيراً، و يعود ذلك لعدد التنقيات المستخدمة و التي تبلغ أربع تقنيات بالمقارنة مع تقنيتين في الدراسة التي تمت مقارنة النتائج معها .

التوصيات المستقبلية :

يعد الحفاظ على الخصوصية في التنقيب في البيانات حالياً موضوعاً ذو أهمية للبحث. توضح مراجعة الأبحاث السابقة أن هناك العديد من تقنيات الحفاظ على الخصوصية المتاحة ولكن لا تزال بها أوجه قصور.

في الأبحاث القادمة يمكن أن يتم:

✓ تطوير الطريقة المستخدمة من خلال دمج تقنيات إضافية بهدف تحقيق مستوى خصوصية أعلى و سرية أكبر لبيانات الأفراد مما يجعل البيانات الحساسة الخاصة بهم في مأمن عن الهجمات الخارجية.

✓ تطوير الطريقة المستخدمة بهدف تقليل زمن التنفيذ و ذلك للحصول على نهج يحقق خصوصية أعلى في زمن صغير نسبياً.

المراجع:

[1] Nasiri, N., & Keyvanpour, M. (2020, December). Classification and Evaluation of Privacy Preserving Data Mining Methods. In 2020 11th International Conference on Information and Knowledge Technology (IKT) (pp. 17-22). IEEE.

[2] Mohammed, K., Ayesh, A., & Boiten, E. (2021). Complementing Privacy and Utility Trade-Off with Self-Organising Maps. *Cryptography*, *5*(3), 20.

[3] Namdev, P., & Kumar, M. (2016). Hybrid Approach for Privacy Preservation data Mining Using Random and Mod Techniques. International Journal of Computer Science Engineering (IJCSE, 5(3), 162-168.

[4] Gunjan S. Bonde Akash D. Waghmare, "Privacy Preservation of Data using Hybrid Approach", International Journal of Management, Technology And Engineering, India, (2019).

[5] Shelke, S., & Bhagat, B. (2015). Techniques for privacy preservation in data mining. *International Journal of Engineering Research*, *4*(10).

[6] Patel, M. K., Patel, M. T., & Patel, M. D. (2016). Privacy Preservation of Data in Data mining using K-anonymity and Randomization Method. International Journal for Innovative Research in Science and Technology.

[7] Srivastava, A., & Srivastava, G. (2015). Privacy Preserving Data Mining in Electronic Health Record using K-anonymity and Decision Tree. International Journal of Computer Science & Engineering Technology, 6(7), 416-426.

[8] Neha, P., Lade, S., & Gupta, R. (2015). Quasi and Sensitive Attribute Based Perturbation Technique for Privacy Preservation.

[9] Kaur, A. (2017, February). A hybrid approach of privacy preserving data mining using suppression and perturbation techniques. In 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) (pp. 306-311). IEEE.

[10] Fletcher, S., & Islam, M. Z. (2015). Measuring information quality for privacy preserving data mining. International Journal of Computer Theory and Engineering, 7(1), 21.