Avaluació de tècniques per a la privacitat en la publicació de dades

Alba Antas Hernandez

Resum– Aquest Treball de Final d'Estudis (TFE) tracta sobre l'estudi de diferents tècniques d'anonimització de dades, amb l'objectiu de protegir la informació confidencial sense comprometre la utilitat del conjunt de dades. S'analitzaran els conceptes teòrics de cada tècnica per posteriorment realitzar la seva implementació pràctica. Finalment, es compararan els diferents mètodes valorant el seu nivell de privacitat i utilitat en diferents contextos mitjançant mètriques especifiques que permetran mesurar la distorsió introduïda en les dades i la capacitat de protegir els diferents registres.

Paraules clau– Privacitat, utilitat, dataset, anonimització de dades, k-anonimitat, privacitat diferencial, generalització

Abstract– This Final Degree Project (TFE) focuses on the study of different data anonymization techniques, with the goal of protecting confidential information without compromising the utility of the dataset. The theoretical concepts of each technique will be analyzed, followed by their practical implementation. Finally, the different methods will be compared, evaluating their level of privacy and utility in various contexts through specific metrics that will measure the distortion introduced in the data and the ability to protect the different data.

Keywords– Privacy, utility, dataset, data anonymization, k-anonimity, differential privacy, generalitzation

1 Introducció

Durant els últims anys, la digitalització ha contribuït en l'augment de la quantitat de dades que es poden recopilar i processar. Aquestes són utilitzades amb diferents finalitats i en molts casos la publicació d'aquestes és necessària per impulsar la recerca en diferents àrees. Per exemple, en l'àmbit de la medicina la publicació de dades mèdiques pot ser beneficiosa per identificar patrons, tendències i noves àrees d'estudi. No obstant això, en la informació que es publica existeixen dades confidencials que sense aplicar els mètodes de protecció adequats pot suposar un risc per a la privacitat i seguretat de la informació personal. [3]

Aquest conflicte ha donat lloc a trobar un equilibri entre la utilitat de les dades i la protecció de la privacitat. Per tal de preservar la privacitat d'aquestes dades es poden aplicar diferents mètodes que facilitin la publicació de la informació assegurant la seva confidencialitat. Una possible solució seria aplicar mètodes de protecció com per exemple l'addició de soroll la qual consisteix a modificar certs valors dins d'un conjunt de dades de manera que es conservi la utilitat estadística, però es dificulti la identificació d'individus.

Cal destacar que no totes les tècniques de protecció proporcionen els mateixos resultats, ja que poden oferir diferents nivells de protecció i poden impactar de diferent manera sobre la qualitat de les dades. Per tant, poden alterar la precisió i la utilitat dels resultats, això fa que l'elecció del mètode de protecció sigui un factor important a tenir en compte.

Així doncs, aquest projecte s'enfocarà en comparar els diferents mètodes de protecció de dades, avaluar les seves característiques, avantatges i els possibles inconvenients. D'altra banda, es buscarà identificar quin d'aquests mètodes aconsegueix un millor equilibri entre la seguretat i confidencialitat de les dades i permetre un anàlisi significatiu de les dades.

• E-mail de contacte: alba.antas@autonoma.cat
• Menció realitzada: Tecnologies de la Informació
• Treball tutoritzat per: Guillermo Navarro Arribas (Àrea de Ciències de la Computació i Intel·ligència Artificial)
• Curs 2024/25

1.1 Objectius

A continuació es mostren els objectius principals del projecte:

- Realitzar un estudi de l'art i conèixer i entendre els diferents models i mètodes de privacitat que existeixen en l'actualitat. Aquest pas és essencial per tal de obtenir un coneixement previ per poder assolir els altres objectius.
- Aplicar i seleccionar els diferents mètodes i tècniques en un conjunt de dades (dataset) [5] per aconseguir preservar la privacitat d'aquestes, veure com funcionen els diferents mètodes i obtenir una sèrie de resultats per poder fer el posterior anàlisi compartiu. Aquest és l'objectiu principal del projecte i la seva realització és essencial.
- Comparar l'efectivitat de cada mètode avaluant els avantatges i els inconvenients en aspectes com la preservació de la utilitat del conjunt de dades i la dificultat per identificar els individus.
- Un cop obtinguts els resultats, extreure unes conclusions sobre quins mètodes són més adients segons els resultats que es vulguin aconseguir.

A continuació es plantegen també objectius secundaris i opcionals que es realitzarien en el cas de tenir el temps suficient amb la finalitat de donar una visió més àmplia del comportament dels diferents mètodes i tècniques que s'apliquin.

- Mesurar i comparar l'impacte computacional de cada mètode per tal de veure no només quin ens ofereix més privacitat i utilitat a les dades sinó també quin ens proporciona un millor rendiment.
- Combinar diferents mètodes entre ells per veure quins resultats s'obtenen de forma conjunta i si aquests son millors que quan s'apliquen de manera independent.

1.2 Metodologia

La metodologia que s'aplicarà en aquest projecte és el model de desenvolupament en cascada, ja que per aconseguir realitzar les taques correctament aquestes s'han de fer de manera seqüencial, començant per la investigació dels diferents mètodes de privacitat i finalitzant amb la implementació i avaluació d'aquests.

Aquest projecte s'iniciarà amb la recerca dels diferents mètodes de privacitat i l'estudi de l'estat de l'art utilitzant diferents eines de documentació. Una vegada analitzada aquesta informació s'escollirà un conjunt de dades adequat per tal d'aplicar els diferents mètodes i tècniques [4]. Un cop seleccionat el conjunt de dades, es procedirà a la implementació dels mètodes seleccionats. S'utilitzarà Python com a llenguatge de programació i Visual Studio Code com a entorn per a realitzar el projecte. Finalment, es realitzarà un estudi dels resultats obtinguts comparant-los i analitzant les característiques de cada tècnica de privacitat.

1.3 Planificació

La planificació del projecte s'ha fet seguint un diagrama de Gantt per tal de veure de forma gràfica i clara les diferents tasques que s'han de realitzar conjuntament amb la seva duració. Les tasques s'organitzen de manera seqüencial de manera que no es comença una fins que hagi finalitzat l'anterior seguint així el model de desenvolupament en cascada. [6]

En aquest projecte es realitzaran les següents tasques:

- Realitzar una recerca dels diferents models i tècniques de privacitat
- Buscar i generar un conjunt de dades per aplicar les diferents tècniques de privacitat
- Implementar les diferents tècniques de privacitat i realitzar una comparació d'aquestes avaluant el nivell de privacitat i utilitat que ens proporcionen.

En les figures 2 i 3 del apèndix podem veure les diferents tasques que s'han realitzat i el temps previst per dur-les a terme juntament amb el diagrama de Gantt generat.

2 Estat de l'art

Un dels aspectes més importants a tenir en compte a l'hora de treballar amb un conjunt de dades és garantir la privacitat d'aquestes. Això es pot aconseguir aplicant diferents mètodes de protecció a les dades. Suposem que tenim un conjunt de dades X i apliquem un mètode de protecció p sobre aquest conjunt de dades. Aquest mètode transforma les dades originals en una versió protegida X' de manera que X' = p(X). D'aquesta manera, aconseguim que les dades del conjunt original X segueixin sent privades encara que es facin públiques les dades X', sempre que el mètode de protecció p que s'apliqui asseguri unes propietats concretes que evitin la publicació d'informació sensible o privada. [9]

Un cop s'aplica un mètode de protecció al conjunt de dades, aquestes poden ser utilitzades per a diferents finalitats, com per exemple impulsar la investigació en diversos camps.

La finalitat principal d'aplicar un mètode de protecció a les dades que volem fer públiques és reduir el risc de revelació, és a dir, el risc que les dades proporcionin informació confidencial. No obstant això, reduir el risc de revelació implica modificar les dades originals. Per tant, es produeix una pèrdua d'informació que pot produir una pèrdua de la utilitat de les dades.

L'objectiu que se'ns planteja és aplicar un mètode de protecció que aconsegueixi minimitzar el risc de revelació, reduir la pèrdua d'informació i maximitzar la utilitat del conjunt de dades.Per tant, es considera com un bon mètode de protecció si aconsegueix mantenir un bon equilibri entre privacitat i utilitat.

Quan un atacant pot obtenir informació privada o pot identificar un individu a partir del conjunt de dades es produeix revelació (disclosure). Existeixen dos tipus de revelació:

- Revelació d'atribut: l'atacant pot obtenir informació sobre el valor d'un atribut, ja sigui de manera directa o indirecta, a partir del conjunt de dades o la relació entre diversos atributs.

- Revelació d'identitat: l'atacant pot identificar correctament una entitat particular o pot relacionar alguna dada protegida amb una entitat especifica, és a dir, quan es produeix una identificació.

Podem classificar els atributs o variables d'un conjunt de dades en diferents categories segons el risc de revelació que presentin: [9]

- Identificadors: identifiquen de manera inequívoca l'entitat o individu. Alguns exemples poden ser el nom o DNI. Normalment s'eliminen o s'encripten per a la publicació del conjunt de dades.

- Quasi-identificadors: no identifiquen a l'individu de manera única però en combinació es poden vincular amb informació externa per identificar una entitat o individu. Alguns exemples són: edat, codi postal, ciutat, etc. Normalment s'aplica un mecanisme de protecció a aquests atributs.

- Confidencials: contenen informació confidencial sobre l'entitat (salari, estat de salut, religió, etc). Normalment no es modifiquen per tal que puguin ser objecte d'estudi.

- No confidencials: no contenen informació confidencial sobre l'individu.

No obstant això, per tal de protegir el conjunt de dades, no és suficient amb eliminar els atributs identificadors, ja que un atacant podria obtenir informació sobre els individus a partir dels atributs quasi-identificadors.

2.1 Mesures de privacitat

2.1.1 K-anonimitat

Un conjunt de dades X satisfà k-anonimitat si com a mínim hi ha k elements indistingibles entre ells respecte a un conjunt de quasi-identificadors, és a dir, un element no es pot distingir d'altres k-1 elements del conjunt de dades. En les taules 1 i 2 podem veure un exemple. [7] [8] [9]

City	Age	Salary
Madrid	23	600
Barcelona	48	500
Valencia	20	734
Madrid	32	543
Sevilla	38	890
Barcelona	41	678
Zaragoza	21	399
Málaga	44	943
Bilbao	31	632

TAULA 1: TAULA EXEMPLE ABANS D'APLICAR K-ANONIMITAT

Country	Age	Salary
(Spain)	[20, 29]	600
(Spain)	[40, 49]	500
(Spain)	[20, 29]	734
(Spain)	[30, 39]	543
(Spain)	[30, 39]	890
(Spain)	[40, 49]	678
(Spain)	[20, 29]	399
(Spain)	[40, 49]	943
(Spain)	[30, 39]	632

TAULA 2: TABULA EXEMPLE DESPRÉS D'APLICAR K-ANONIMITAT

Com podem veure en l'exemple de la figura 2, sempre hi ha tres registres indistingibles, per tant k=3. En aquest cas s'aplica k-anonimitat als quasi-identificadors de la ciutat i l'edat. En aquest context, la probabilitat de revelació és com a màxim 1/3.

2.1.2 Privacitat diferencial

Existeix privacitat diferencial si en considerar dues bases de dades idèntiques, excepte per la inclusió o l'exclusió d'un sol registre, es produeixen resultats casi indistingibles, és a dir, el resultat de l'anàlisi que es fa sobre aquestes dades no es veu afectat. [1]

City	Age	Salary
Madrid	23	600
Barcelona	48	500
Valencia	20	734
Madrid	32	543
Sevilla	38	890
Barcelona	41	678
Zaragoza	21	399
Málaga	44	943
Bilbao	31	632
Madrid	18	545

D1

City	Age	Salary
Madrid	23	600
Barcelona	48	500
Valencia	20	734
Madrid	32	543
Sevilla	38	890
Barcelona	41	678
Zaragoza	21	399
Málaga	44	943
Bilbao	31	632

D2

Fig. 1: Dues versions d'una mateixa BD

En la figura 1 podem veure dues versions d'una base de dades en la que difereixen en un sol registre. En la primera versió tenim el registre amb l'atribut edat = 18, en canvi, en la segona versió aquest registre no existeix.

Considerem que f es la mitjana aritmètica de tots els registres, si fem una consulta f a les dues versions obtenim els resultats que apareixen a la taula 3.

Dataset	f(D)
f(D1)	33.1
f(D2)	31.6

TAULA 3: RESULTATS CONSULTA MITJANA ARITMÈTICA

Si comparem f(D1) i f(D2) podem obtenir informació sobre el valor eliminat.

L'objectiu principal de la privacitat diferencial és que les diferències entre les consultes siguin el menys diferent possible. Això s'aconsegueix aplicant aleatorització a la resposta, com per exemple afegint soroll. [9]

2.2 Mètodes de protecció

Podem classificar els mètodes de protecció de la següent manera:

- Pertorbatius: les dades originals X es distorsionen, per tant, les dades X' contenen informació incorrecta.

- No Pertorbatius: es substitueixen els valors originals per altres menys detallats o precisos.

2.2.1 Mètodes pertorbatius

Soroll Additiu - L'addició de soroll és una tècnica de protecció de les dades que consisteix en afegir soroll a les dades de manera que siguin menys precises i sigui més difícil reconstruir les dades originals. Aquesta tècnica es basa en sumar soroll a les dades originals X per obtenir unes dades protegides X'. Per tal de modelar el soroll, normalment s'utilitza una distribució normal.[9]

Soroll multiplicatiu - Aquesta tècnica de protecció de les dades a diferència de l'addició de soroll, consisteix en modificar les dades multiplicant-les per un valor aleatori. El principal avantatge respecte a l'addició de soroll és que l'error que s'introdueix a partir del soroll és proporcional al valor que s'aplica. [9]

Microagregació - La microagregació consisteix en agrupar registres similars en clústers d'almenys k registres. Construeix petits microclústers i substitueix cada dada original per el seu representant o centroide. Com a resultat, tots el registres del mateix microclúster passen a tenir el mateix valor. [9]

2.2.2 Mètodes no pertorbatius

Generalització - La generalització és un mètode de protecció no pertorbatiu que consisteix en substituir un valor original per un més general obtenint com a resultat dades menys concretes [9]. Podem aplicar generalització de diverses maneres depenent de l'atribut. En el cas del exemple de les taules 4 i 5 i , podem veure com a l'atribut edat es substitueix per un rang en lloc de mostrar el valor spec, podem veure com a l'atribut edat es substitueix per un rang en lloc de mostrar el valor específic o en el cas del codi postal es substitueixen les dues últimes xifres per un asterisc.

Age	Zip	Profession
30	08193	Nurse
36	08176	Paramedic
32	08191	Veterinarian
25	08034	Mathematician
29	08022	Physics
27	08010	Engineer

TAULA 4: DATASET ORIGINAL ABANS D'APLICAR GENERALITZACIÓ

Age Range	Zip Code	Profession
[30,40)	081**	Health prof.
[30,40)	081**	Health prof.
[30,40)	081**	Health prof.
[20,30)	080**	Science prof.
[20,30)	080**	Science prof.
[20,30)	080**	Science prof.

TAULA 5: DATASET DESPRÉS D'APLICAR GENERALITZACIÓ

3 CREACIÓ DEL DATASET

El dataset escollit per tal de realitzar aquest estudi és una base de dades que inclou informació dels pacients d'un hospital. Aquest dataset s'ha obtingut de la plataforma Kaggle[4]. S'ha escollit aquest dataset perquè conté mes de 50.000 registres amb diferents tipus d'atributs, aquests són: Name, Age, Gender, Blood Type, Medical Condition, Doctor, Hospital, Insurance Provider, Room Number, Admission Type, Discharge Date, Medication, Test Results. Per a la generació del dataset s'ha implementat una funció Python que llegeix el dataset i elimina els atributs identificadors Name i Doctor. A continuació, crea una nova mostra amb el número n de registres passats per paràmetre. Finalment es genera un nou fitxer csv amb el dataset modificat.

4 IMPLEMENTACIÓ SOROLL ADDITIU I MULTIPLICATIU

Un cop s'ha generat el dataset, s'han implementat els mètodes aleatoris. Aquesta implementació inclou la realització de dues funcions, la primera per aplicar soroll additiu a les dades i la segona per aplicar soroll multiplicatiu. A les dues funcions s'itera sobre cada columna especificada per la llista columnes que es passa per paràmetre, les columnes escollides en aquest cas són: 'Age', 'Billing Amount', 'Room Number'. Seguidament, per a cada columna es calcula la desviació estàndard (sigma) i es multiplica per un valor p que ens indica el nivell de distorsió de les dades. Com més gran sigui p més privacitat. A continuació, es genera soroll a partir d'una distribució normal amb mitja 0 i una desviació estàndard igual a l'arrel quadrada del valor calculat anteriorment. Finalment es suma o es multiplica el soroll, segons el mètode, a les dades originals.

Per a la implementació de les dues funcions s'han utilitzat les llibreries pandas i numpy que ens permeten manipular les dades de manera fàcil.

En la taula 6 podem veure un exemple dels 10 primers registres per a la columna Age després d'aplicar soroll additiu

Original	Soroll Additiu
85	85
84	102
30	42
69	74
57	63
78	68
44	52
67	76
32	35
66	64

TAULA 6: TAULA VALOR ORIGINAL I DESPRÉS D'ALICAR SOROLL ADDITIU PER A LA COLUMNA AGE

Age	Blood T	Date of Admission	Room Num
77	AB-	2021-07-10	294
65	AB+	2019-12-05	171
65	AB-	2021-02-15	353
26	A-	2023-10-23	313
63	A-	2019-10-05	336
64	A-	2023-11-04	152
73	A-	2023-01-01	152
75	O+	2020-02-09	182
57	O+	2022-02-08	425
41	AB-	2019-08-23	113

TAULA 7: VALORS DELS ATRIBUTS AGE, BLOOD TYPE, DATE OF ADMISSION I ROOM NUMBER EN EL DATASET ORIGINAL

5 IMPLEMENTACIÓ DEL MÈTODE DE GENERALITZACIÓ

En aquesta fase del projecte s'ha implementat la tècnica de generalització de les dades. El principal objectiu d'aquest mètode és mantenir la privacitat de les dades originals i aconseguir satisfer la propietat de k-anonimitat.

La generalització s'ha aplicat en diversos atributs amb l'objectiu de fer-los menys concrets, es a dir, generalitzarlos. Per exemple, en el cas de l'edat podem agrupar aquest valor en grups de 10 anys ,de tal manera que el valor d'aquest camp serien rangs com per exemple, de 0-10 anys , de 11-20 anys, etc. Amb això aconseguim reduir el risc de reidentificació ja que es substitueixen valors específics per intervals o categories més amplies.

Per tal d'aconseguir aplicar aquests mètodes al dataset, s'ha implementat una funció utilitzant les llibreries pandas i numpy, com en les funcions anteriors, per facilitar la manipulació de les dades.

Per a la generalització, s'han escollit els atributs Age, Room Number, Blood Type , Date Admission i Discharge Date. En el cas dels atributs Age i Room Number, aquests s'han agrupat en rangs de 10 i 100 respectivament, transformant una edat especifica com 31 en intervals com 30-39. Per el cas de les dates, que corresponen als atributs Date Admission i Discharge Date, s'han generalitzat eliminant el dia i substituint-lo per un asterisc deixant només el mes i l'any. Per últim, per generalitzar l'atribut Blood Type s'han substituït els símbols "+" i "-", que identifiquen el tipus de sang, per un asterisc, agrupant així els tipus de sang positius i negatius en una sola categoria.

Per tal de veure un exemple en les taules 7 i 8 podem veure els resultats d'aplicar el mètode en el dataset.

Age	Blood T	Date of Admission	Room Num
70-79	AB*	2021-07	200-299
60-69	AB*	2019-12	100-199
60-69	AB*	2021-02	300-399
20-29	A*	2023-10	300-399
60-69	A*	2019-10	300-399
60-69	A*	2023-11	100-199
70-79	A*	2023-01	100-199
70-79	O*	2020-02	100-199
50-59	O*	2022-02	400-499
40-49	AB*	2019-08	100-199

TAULA 8: VALOR DELS ATRIBUTS AGE, BLOOD TYPE, DATE OF ADMISSION I ROOM NUMBER DESPRÉS D'APLICAR GENERALITZACIÓ

6 IMPLEMENTACIÓ DEL MÈTODE DE MICROAGREGACIÓ

La microagregació consisteix en crear grups o clústers de registres similars on les dades es substitueixen per valors agregats. L'objectiu principal és reduir el risc de reidentificació mantenint la utilitat del conjunt de dades.

En aquest cas, s'ha aplicat microagregació univariable, ja que es tracta cada columna de manera individual. Aquesta microagregació, s'ha aplicat a columnes amb valors numèrics, com l'edat i el número d'habitació, i s'ha utilitzat un algoritme de partició en grups consecutius, on els valors de cada columna s'ordenen de major a menor i es divideixen en grups de k elements. A continuació, es calcula la mitjana aritmètica de cada grup i es substitueixen els valors originals per la mitjana calculada.

Per aconseguir aplicar aquesta microagregació al conjunt de dades, s'ha implementat una funció utilitzant les llibreries pandes i numpy per poder manipular les dades fàcilment. Els atributs on s'ha aplicat microagregació són Age i Room Number, aquest es passen per paràmetre en la variable columnes.

En la taula 9 podem veure un exemple del resultat d'aplicar microagregació en l'atribut Age amb una k = 3

Original	Microagregació
85	76.6
44	41.3
51	54
79	76.6
43	41.3
66	76.6
37	41.3
51	54
60	54

TAULA 9: VALOR DEL ATRIBUTS AGE EN LA VERSIÓ ORIGINAL I DESPRÉS D'APLICAR MICROAGREGACIÓ

Age	Room Number
77	294
65	171
65	353
63	336
64	152
73	152
75	182
57	425
41	113
66	358

TAULA 10: VERSIÓ ORIGINAL ATRIBUTS: AGE I ROOM NUMBER

7 IMPLEMENTACIÓ TÈCNIQUES DE PRIVACITAT DIFERENCIAL

En aquesta fase del projecte, l'objectiu principal és aconseguir anonimitzar un conjunt de dades i que si afegim o modifiquem algun registre del conjunt no afecti al resultat de l'anàlisi de les dades. Per tal de aconseguir-ho, s'aplica algun tipus d'aleatorietat a la resposta, en aquest cas s'afegeix soroll laplacià. És important destacar que aquest mètode afegeix privacitat a les respostes de consultes que es fan al conjunt de dades, per tant aquest no queda exposat.

Per aconseguir privacitat diferencial, el soroll que s'afegeix a la consulta ha de complir uns paràmetres concrets que depenen del tipus de consulta i de la sensibilitat dels atributs que es volen protegir. Per una banda, hem de determinar el tipus de sensibilitat que determina quina és la màxima influencia que pot tenir un registre en el resultat de la consulta i posteriorment serà utilitzada per definir la distribució de soroll. Per determinar aquest paràmetre, calcularem la diferencia entre el màxim i el mínim valor entre el nombre de registres.

Per tal d'aplicar això al conjunt de dades, s'han implementat dues funcions. La primera consisteix en aplicar soroll laplacià, aquest es genera seguint una distribució Laplaciana centrada en zero on la dispersió del soroll ve determinada per el paràmetre èpsilon i la sensibilitat, quan més gran és aquest valor més petita és la quantitat de soroll, per tant, s'aconsegueix menys privacitat. Aquesta distribució ve determinada per la següent funció, on μ és la mitjana de la distribució que pren el valor de 0 i b és la dispersió del soroll que es determina a partir de la sensibilitat i el valor d'èpsilon:

$$f(x \mid \mu, b) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}$$

La segona funció, rep la consulta realitzada per paràmetre i aplica soroll mitjançant la funció anterior. Cal destacar que aquest mètode només es podrà aplicar als atributs numèrics.

Per comprovar el funcionament d'aquest mètode s'ha realitzat una consulta als atributs Age i Room Number amb una epsilon de 0.1. En la taula 10 i 11 podem observar les diferències entre el conjunt original i el resultat de la consulta.

Age	Room Number
87.31	299.04
53.46	187.15
60.98	354.43
60.44	320.67
56.03	181.02
76.92	179.04
58.71	181.20
67.61	456.73
39.16	95.37
64.28	357.37

TAULA 11: ATRIBUTS AGE I ROOM NUMBER DESPRÉS D'APLICAR SOROLL A LA RESPOSTA DE LA CONSULTA DE PRIVACITAT DIFERENCIAL

8 ANÀLISI COMPARATIU DELS RESULTATS

L'objectiu principal d'aquesta fase del projecte és avaluar els diferents mètodes i tècniques de privacitat que s'han aplicat al conjunt de dades original. Per realitzar aquesta comparació s'ha analitzat la privacitat i la utilitat que ens proporciona cada tècnica i identificant les avantatges i desavantatges de cada una. Per poder mesurar la privacitat i la utilitat s'han considerat mètriques com el percentatge de reidentificacions com indicador de privacitat i l'error introduït com indicador de utilitat.

En el cas de la utilitat, s'ha analitzat el impacte que generen els diferents mètodes en el conjunt de dades originals. Per tal d'aconseguir-ho, s'ha calculat l'error introduït en cada columna. Per a les columnes categòriques, com per exemple el tipus de sang, s'ha mesurat el número de coincidències entre el conjunt de dades original i el modificat i calculem l'error com el complement del percentatge de coincidències. Per a les columnes numèriques, s'ha calculat el error relatiu (MRE) [12]per observar la desviació introduïda. S'utilitza el MRE, ja que ens permet comparar dades de diferents escales i magnituds. A més a més, en el cas d'atributs no numèrics es fa servir una metodologia similar al MRE seguint la mateixa lògica. Per exemple, per a les dates, es calcula la distancia temporal entre el valor original i el protegit en dies i es calcula l'error relatiu a partir d'aquesta distancia.

Per mesurar la privacitat s'ha calculat el percentatge de reidentificacions, aquesta tècnica es coneix com a record

linkage o enllaç de registres. El record linkage, ens indica quants registres del conjunt de dades anonimitzat es poden associar correctament amb registres del dataset original. Aquesta associació es fa mitjançant mesures de distancia, com la distancia euclidiana[10][11], la qual es fa servir en aquest cas. S'utilitza aquesta distancia ja que, és fàcil d'implementar i interpretar, i es tracta cada atribut com una dimensió d'un espai multidimensional, es a dir, si hi han 8 atributs (columnes), cada registre és un punt en un espai de 8 dimensions.

Aquesta distancia ve determinada per la següent fórmula on X i Y son els dos registres que es comparen i *i* el nombre d'atributs de cada registre:

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Per tant, per a mesurar la privacitat, es calculen les distancies entre els registres del conjunt de dades protegit i tots els registres del conjunt de dades original. Seguidament, per a cada registre del dataset protegit, s'identifica el registre del dataset original amb la mínima distancia, si aquest coincideix amb el registre original es produeix una reidentificació.

Per comparar totes les tècniques, s'han aplicat dues funcions, una per mesurar la privacitat i una per la utilitat. A més a més, s'analitza com varien aquest resultats depenent de la mida del dataset per poder observar patrons i tendències que podrien influir en escollir una tècnica especifica per a diferents situacions.

Per visualitzar els resultats d'una manera més clara, s'ha realitzat una taula per a els resultats de cada mètode. La primera columna ens indica la mida del conjunt de dades utilitzat. En la següent columna de la taula es mostra el percentatge d'encerts, es a dir, el percentatge de registres que s'han reidentificat mitjançant record linkage. Si aquest percentatge és elevat ens indica que molts registres poden ser identificats fàcilment, per tant, hi ha menys privacitat. La columna d'error, ens mostra el percentatge d'error calculat amb el MRE. Aquesta columna ens indica la distorsió introduïda en les dades després d'aplicar les tècniques de protecció. Per últim, la columna d'utilitat representa el percentatge del conjunt de dades protegit per mantenir les seves propietats originals. Es defineix com 100% - Error%, de manera que un percentatge d'error elevat implica una utilitat baixa.

8.1 Resultats soroll additiu

El mètode de soroll additiu consisteix en agregar valors aleatoris, generats a partir d'una distribució calculada, a les variables numèriques del dataset original, amb l'objectiu de dificultar la identificació dels registres sense afectar significativament a la utilitat. S'ha utilitzat un soroll de 0.3, ja que introdueix una quantitat de soroll moderada sense alterar excessivament les dades i aconseguir un equilibri entre privacitat i precisió en el conjunt de dades. El soroll additiu s'afegeix a cada valor del dataset seguint una distribució normal on es pren μ=0 i σ ve determinada per el soroll que es vol afegir, en aquest cas si el soroll es 0.3 la distribució normal utilitzada es podria expressar com:

$$X \sim N(0, 0.3^2)$$

En la taula 12 podem veure el resultats obtinguts amb diferents mides del dataset.

Mida	Encerts(%)	Error (%)	Utilitat (%)
50	16	4.64	95.35
125	6.4	5.94	94.05
250	6	6.08	93.91
375	4.8	6.89	93.10
500	4.39	6.65	93.34

TAULA 12: TAULA RESULTATS COMPARACIÓ SOROLL ADDITIU

Com podem veure en els resultats, el percentatge de reidentificacions varia entre un 4,39% i un 16%, per tant, aquest mètode ens ofereix un bon nivell de privacitat ja que no és un percentatge molt elevat. També podem observar com aquest percentatge disminueix a mesura que augmenta la mida del dataset, això podria ser perquè a mesura que augmenta el conjunt de dades, els valors alterats son menys identificables per l'increment de diversitat en les dades. Pel que fa a la utilitat, l'error associat es troba en un rang de 4,64% a 6,89%, això ens indica que es produeix més distorsió en les dades en datasets més grans. No obstant això, la utilitat de les dades es manté elevada per diferents mides del conjunt de dades, entre un 95% i un 93%, això en mostra que, encara que es perdi utilitat, no es produeix un distorsió significativa.

8.2 Resultats soroll multiplicatiu

La tècnica de soroll multiplicatiu es un bon mètode per ocultar les relacions directes entre variables numèriques. S'ha utilitzat un soroll de 0.3 igual que en el cas anterior. Com podem veure en la taula 13, que ens mostra els resultats obtinguts, s'observa un percentatge de reidentificacions elevat. Això es degut a que, el soroll multiplicatiu modifica el valors originals però es mantenen les proporcions, per tant, si existeixen patrons o relacions entre els valors facilita la deducció de les dades originals. També es pot observar, com la privacitat augmenta a mesura que creix la mida del dataset, ja que, existeix més diversitat de valors, per tant, augmenta la dificultat d'identificar cada dada original. Aquest augment de la privacitat també podria ser per la probabilitat de que hi hagin més registres similars, aconseguint així més privacitat.

Mida	Encerts(%)	Error(%)	Utilitat(%)
50	86	16.35	83.64
125	64	15.82	84.17
250	57.59	15.62	84.37
375	50.66	15.93	84.06
500	54.2	16.06	83.93

TAULA 13: TAULA RESULTATS COMPARACIÓ SOROLL MULTIPLICATIU

Respecte a la utilitat, aquesta es manté constant, sense importar la mida del dataset, per tant, aquest mètode introdueix una distorsió uniforme. El percentatge d'error associat a aquest mètode és més elevat en comparació amb el

soroll additiu, per tant, existeix més distorsió en els valors numèrics.

8.3 Resultats generalització

L'objectiu principal de la generalització és fer menys concrets els valors originals agrupant-los en categories o en rangs. Aquest mètode destaca per proporcionar un nivell elevat de privacitat, i això es pot veure en els resultats obtinguts que es mostren en la taula 14.

Mida	Encerts(%)	Error(%)	Utilitat(%)
50	2	8.83	91.16
125	0.8	8.92	91.07
250	0.4	8.93	91.06
375	0.26	8.93	91.06
500	0.2	8.92	91.07

TAULA 14: TAULA RESULTATS COMPARACIÓ GENERALITZACIÓ

El resultats ens mostren uns percentatges de reidentificacions baixos, per tant, podem dir que aquest mètode en proporciona una privacitat alta. En reemplaçar valors específics amb representacions més generals augmenta la dificultat de reidentificar els registres de manera directa. A més a més, podem observar com la privacitat augmenta amb la mida del dataset, ja que augmenta la probabilitat de que els intervals o categories comprenguin més registres.

L'error associat a la generalització es constant, al voltant del 8%, per tant, la utilitat de les dades protegides és alta, un 91%, el que ens indica que aquesta tècnica és adequada per anàlisis estadístics i tendències generals.

8.4 Resultats microagregació

L' objectiu principal de la microagregació és agrupar registres similars en grups de k elements. En aquest cas s'ha utilitzat una k = 3 per assegurar un nivell moderat de privacitat sense alterar significativament les dades. Valors més grans de k ens proporcionarien més privacitat però podrien afectar a la utilitat del conjunt de dades. En els resultats obtinguts, que es poden veure en la taula 15, el percentatge de reidentificacions és elevat, entre un 38% i un 88%. També podem veure com aquest percentatge va augmentant conforme augmenta la mida del conjunt de dades. Amb l'existència de més registres, es poden formar grups amb més dades que comparteixen característiques similars, per tant, el valor central que representa al grup, s'apropa més als valors originals fent que les dades siguin més distingibles i disminueixi el nivell de privacitat.

Mida	Encerts(%)	Error (%)	Utilitat (%)
50	38	0.87	99.12
125	49.6	0.34	99.65
250	68.4	0.25	99.74
375	81.6	0.14	99.84
500	88.4	0.37	99.62

TAULA 15: TAULA RESULTATS COMPARACIÓ MICROAGREGACIÓ

A diferencia del nivell de privacitat, el percentatge d'utilitat que aconseguim amb la microagregació és molt elevat, al voltant del 99%. Aquest percentatge ens indica que les dades protegides mantenen un bon nivell de precisió.

8.5 Resultats privacitat diferencial

Per realitzar l'anàlisi del conjunt de dades aplicant privacitat diferencial, s'han generat diferents conjunts de dades amb les consultes realitzades. El valor d'èpsilon utilitzat és de 0.3 per garantir un nivell alt de privacitat i sense distorsionar significativament la informació. Aquest valor d'èpsilon ens proporciona un equilibri entre privacitat i precisió de les dades. S'ha aplicat soroll a les columnes a les quals s'ha realitzat la consulta, aquestes son: Age i Room Number. Els resultats obtinguts es poden observar en la taula 16.

Mida	Encerts(%)	Error (%)	Utilitat (%)
50	42	14.62	85.37
125	30.4	13.91	86.08
250	16.4	13.19	86.80
375	12.53	13.6	86.39
500	9.6	13.68	86.31

TAULA 16: TAULA RESULTATS COMPARACIÓ PRIVACITAT DIFERENCIAL

En els resultats podem veure com la privacitat augmenta considerablement amb la mida del dataset. La privacitat diferencial introdueix soroll en les respostes per protegir les dades. Quan el conjunt de dades és més gran, el impacte del soroll disminueix proporcionalment per l'existència de més registres, per tant, augmenta la dificultat de identificar registres específics.

Per últim, en relació a la utilitat, obtenim un bon percentatge d'error i aquest es manté constant per a les diferents mides del conjunt de dades, per tant, es conserva la precisió de les dades.

9 CONCLUSIONS

9.1 Conclusions dels resultats

Un cop aplicats els mètodes al conjunt de dades, es pot observar que cada tècnica afecta de manera diferent a les dades originals obtenint diferents resultats que les fa més adequades depenent del context i objectius per a les quals s'utilitzin.

Per una banda, el soroll additiu mostra un bon equilibri entre privacitat i utilitat. No obstant això el percentatge de reidentificacions és mes elevat en comparació amb altres mètodes, especialment en conjunts de dades més petits. Amb el soroll multiplicatiu, també obtenim un bon percentatge d'utilitat, però es un dels mètodes que menys privacitat ens proporciona amb percentatges de reidentificacions especialment elevats. Tot i que no es proporcioni un bon nivell de privacitat, aquest mètode segueix sent útil per a escenaris on es necessiti precisió en les dades.

Pel que fa a la generalització, s'ha vist que és la tècnica que més privacitat ens ha aportat amb percentatges de reidentificacions molt baixos. A més a més, ens proporciona

un bon nivell d'utilitat, per tant, és una bona tècnica per a escenaris on es necessiti un equilibri en privacitat i utilitat.

Respecte a la microagregació, mostra percentatges molt elevats d'utilitat, amb valors propers al 99%. No obstant, a mesura que augmenta la mida del dataset, el percentatge de reidentificacions també augmenta considerablement, això fa que la microagregació no sigui un bon mètode per mantenir la privacitat en grans conjunts de dades.

En relació amb la privacitat diferencial, mostra un rendiment constant en la utilitat de les dades, amb percentatges al voltant del 86%. Pel que fa a la privacitat, s'ha observat com augmenta considerablement quan augmenta la mida del conjunt de dades, passant de un 42% a un 9%, per tant, podríem dir que aquesta tècnica és adequada per treballar amb grans conjunts de dades.

En conclusió, tot i que hi ha mètodes que ens proporcionen un millor equilibri entre privacitat i utilitat, la elecció del mètode depèn de diferents factors, com la mida del dataset o el context de l'anàlisi. Per exemple, la generalització seria adecuada per protegir dades molt sensibles, en canvi, el soroll additiu i la microagregació són preferibles per anàlisis que requereixen precisió en els resultats. En grans conjunts de dades, els mètodes com la privacitat diferencial i el soroll multiplicatiu són més adequats per reduir el risc de reidentificació sense comprometre significativament la utilitat de les dades

9.2 Conclusions finals

Com a conclusió final, durant aquest projecte s'ha analitzat i comparat l'eficàcia de diferents tècniques d'anonimització per observar quin grau de privacitat i utilitat ens proporciona cada una. S'han analitzat el conceptes teòrics de cada mètode i la seva implementació pràctica, avaluant el seu impacte en diferents contextos. Aquest anàlisis, ha contribuït en entendre com cada tècnica s'adapta de manera diferent segons el context o els requisits que es plantegin. En general, la elecció de cada tècnica depèn de l'equilibri que es busqui entre privacitat i utilitat i la qualitat de les dades que es volen protegir.

REFERÈNCIES

[1] IBM, "What is differential privacy?,ÏBM Data Privacy. https://www.ibm.com/topics/differential-privacy.

[2] V. Torra, Guide to data privacy: Models, technologies, solutions, 1a ed. Springer International Publishing, 2022.

[3] National Institute of Standards and Technology (NIST), An introduction to privacy engineering and risk management in federal systems, NIST, 2018. https://csrc.nist.gov/publications/detail/nistir/8062/final.

[4] "Find Open Datasets and machine learning Projects,"Kaggle.com. https://www.kaggle.com/datasets.

[5] D. C. Solis, "Datasets: Qué son y cómo acceder a ellos,"Openwebinars.net, 1 maig 2023. https://openwebinars.net/blog/datasets-que-son-y-como-acceder-a-ellos/.

[6] Y. Arrabal, "Metodología Waterfall: cómo funciona,"Randstad, 5 desembre 2023. https://www.randstad.es/contenidos360/productividad/metodologia-waterfall/.

[7] Agencia Española Protección de Datos, LA K-ANONIMIDAD COMO MEDIDA DE LA PRIVACIDAD. https://www.aepd.es/guias/nota-tecnica-kanonimidad.pdf

[8] Alexandre Viejo.(2020). Privadesa en la publicació de dades. Universitat Oberta de Catalunya https://openaccess.uoc.edu/bitstream/10609/150223/8/PrivadesaEnLaPublicacioDeDades.pdf

[9] Guillermo Navarro Arribas. Introducció a la privadesa en la publicació de dades. Universitat Oberta de Catalunya.

[10] Understanding Euclidean distance analysis. (s/f). ArcGIS Pro. [Online]. https://pro.arcgis.com/es/pro-app/latest/tool-reference/spatial-analyst/understanding-euclidean-distance-analysis.htm

[11] Distancia euclidiana. (s/f). Lifeder. https://www.lifeder.com/distancia-euclidiana/

[12] MRE documentation. (s/f). Permetrics. https://permetrics.readthedocs.io/en/latest/pages/regression/MRE.html

[13] González Martínez, P. (2023). Trabajo Final de Máster. https://openaccess.uoc.edu/bitstream/10609/148163/1/pedrogonzalezmartinezTFM0623memoria.pdf

APÈNDIX

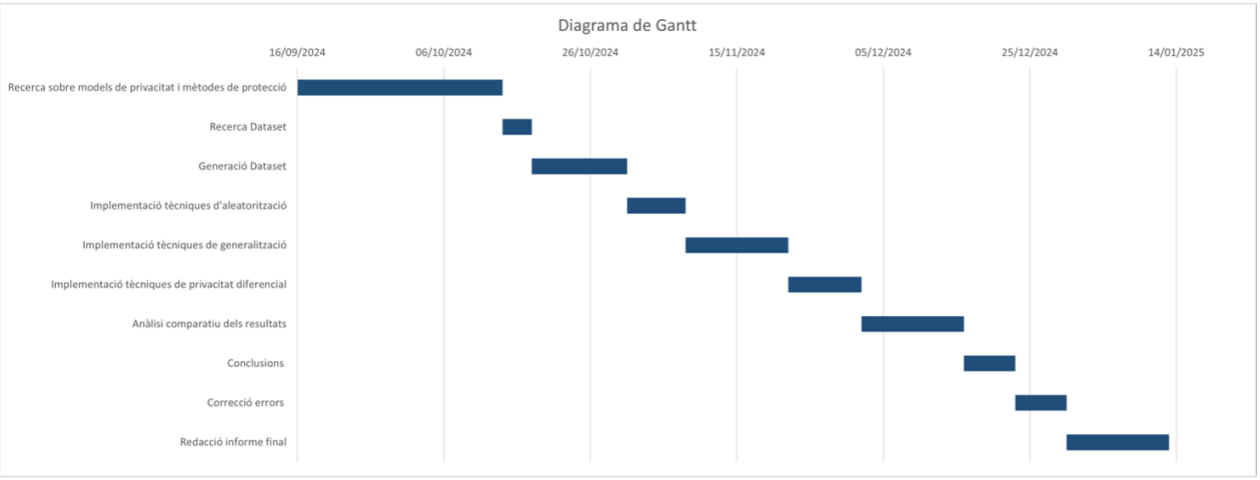

Fig. 2: Diagrama de Gantt

Tasques	Data Inici	Duració	Data Final
Recerca sobre models de privacitat i mètodes de protecció	16/09/2024	28	14/10/2024
Recerca Dataset	14/10/2024	4	18/10/2024
Generació Dataset	18/10/2024	13	31/10/2024
Implementació tècniques d'aleatorització	31/10/2024	8	08/11/2024
Implementació tècniques de generalització	08/11/2024	14	22/11/2024
Implementació tècniques de privacitat diferencial	22/11/2024	10	02/12/2024
Anàlisi comparatiu dels resultats	02/12/2024	14	16/12/2024
Conclusions	16/12/2024	7	23/12/2024
Correcció errors	23/12/2024	7	30/12/2024
Redacció informe final	30/12/2024	14	13/01/2025

Fig. 3: Taula planificació del projecte