Estudi de la privacitat en processos de publicació de dades

Guillem Valero Palomares

Resum – Aquest Treball de Final de Estudis (TFE) tracta de l'estudi dels mètodes o accions a realitzar per poder anonimitzar un conjunt de dades. Es durà a terme un seguit de mètodes de preservació de la privacitat sobre un conjunt de dades per extreure dos datasets, un enfocat en preservar l'utilitat de les dades i el segon en mantenir la privacitat. Tot seguit es comparan els dos conjunts per extreure conclusions sobre quins mètodes porten a cap una millor anonimització.

Paraules clau – Dataset, anonimització, privacitat, clusterització, k-means, R

Abstract – This Final Degree Project (FDP) deals with the study of the methods or actions to be performed in order to anonymize a data set. A set of privacy methods will be carried out on the raw data set to extract two datasets, one focused on maintaining the usefulness of the data and the second on achieving privacy. They will be compared to draw conclusions about which methods lead to better anonymization.

Keywords – Dataset, anonymization, privacy, clustering, k-means, R

1 Introducció

Amb l'augment de la intel·ligència artificial (IA) en camps com la medicina, ha sorgit la necessitat d'obtenir grans volums de dades per poder entrenar xarxes neuronals [1] [2]. El principal problema és que les dades que es necessiten poden ser dades sensibles, les quals no es poden compartir lliurement segons la legislació vigent [3]. A Europa regeix el Reglament General de Protecció de Dades (RGPD) que va sorgir el 2016 i va ser actualitzada el 2018 [4]. Per aquest motiu, s'han hagut d'introduir mètodes de preservació de la privacitat que fa anys que es treballen per aconseguir aquest propòsit. Aquests mètodes tenen el problema de l'introducció de soroll per poder anonimitzar i privatitzar les dades [5]. Un factor molt important a l'hora de anonimitzar qualsevol dada és la quantitat de soroll que s'ha d'introduir en el conjunt de dades, ja que si el valor del soroll és molt elevat, aquestes dades queden privatitzades però són inservibles. És a dir, ha d'haverhi un balanç entre el soroll, anonimitat i utilitat de les dades [6] [7].

• E-mail de contacte: 1492309@uab.cat
• Menció realitzada: Tecnologies de la Informació
• Treball tutoritzat per: Jordi Casas Roma (Àrea de Ciències de la Computació i Intel·ligència Artificial)
• Curs 2021/22

1.1 Objectius

Aquest treball vol donar resposta als següents objectius:

- O1: Realitzar un estudi de l'estat de l'art sobre els diferents mètodes i tècniques d'anonimització de dades que existeixen actualment [8].
- O2: Implementar algunes de les diferents tècniques i mètodes de anonimització de dades sobre un conjunt de dades [9] [10].
- O3: Elaborar un anàlisi comparatiu dels conjunts de dades (datasets, en anglès) generats, per així poder estudiar la privacitat i pèrdua d'informació dels diferents mètodes.

  A partir dels resultats obtinguts, extreure unes conclusions sobre quins mètodes són els més adients i quina quantitat de soroll és la més adequada per als diferents tipus de dades que constitueixen els conjunts de dades o datasets.

1.2 Planificació del projecte

Aquest projecte vol dur a terme les següents tasques:

- Realitzar una recerca d'informació sobre la privatització, anonimització de dades i del dataset amb què treballar. Per poder complir, el primer objectiu de fer un estudi de l'estat de l'art.

- Generar els dos tipus de dataset amb l'implementació dels mètodes sobre el dataset trobat, complint el segon objectiu del treball.
- Anàlisis comparatiu sobre el conjunt de dades, per obtenir mètriques amb les quals realitzar la comparativa, complint el tercer objectiu del treball.

El detall de la planificació de les tasques, es pot trobar el diagrama de Gantt de la planificació en la taula 4 de l'apèndix.

2 ESTAT DE L' ART

La publicació d'un conjunt de dades, segueix un escenari. Aquest es composa de dues parts: una privada i una publica. La part privada es composa de la recopilació de dades, per generar un conjunt da dades. Aquest conjunt de dades esta format per: identificadors, quasi identificadors, informació sensible i no sensible. La part publica esta formada per les dades del conjunt sense els identificadors, per tant no es trivial identificar a un individu.

Encara que s'han eliminat tots els identificador, es pot identificar de forma univoca a un individu amb una conbinació dels quasi identifcador. Per aquest motiu el propietari de les dades ha de realitzar una fase d'anonimització de les dades del conjunt per evitar aquesta identificació.

Els diferents atributs que conté una base de dades (BBDD) es divideixen en quatre classes, depenen del tipus d'informació que contenen: Els identificadors són atributs que poden identificar a un individu de forma unívoca, com el DNI o un compte corrent. Els quasi identificadors, són atributs que per si sols no identifica a un individu, però que si es combinen amb altres, llavors sí que poden identificar de forma única a un individu. Els atributs sensibles presenten informació sensible de l'individu, com ara, la religió, el salari, les malalties, etc. Per últim, els atributs no sensibles són els atributs que no es poden classificar en cap dels anteriors. Per més informació sobre els tipus d'atributs [11].

A l'hora d'anonimitzar un conjunt de dades s'han de tenir en compte uns riscos. Aquests són, la singularitat, la vincularitat i la inferència.

- **Singularitat:** És la possibilitat d'extreure d'un conjunt de dades, un registre o tots els registres que referencien a un individu[8].
- **Vincularitat:** Es tracta de la capacitat per vincular, com a mínim, dos registres d'un únic interessat o grup d'interessats d'una o més bases de dades[8].
- **Inferència:** És la possibilitat de deduir el valor d'un atribut a partir dels valors d'altres atributs, amb una probabilitat significativa[8].

Per poder mitigar aquests riscos existeixen dues formes. La primera fa referència a l'aleatorització de les dades, què és una família de tècniques que modifiquen lleugerament el valor de les dades per eliminar el vincle amb la persona. Alguns exemples de tècniques com ara: l'addició de soroll, les permutacions, la privacitat diferencial. Gràcies a aquestes tècniques es pot mitigar el risc de la inferència. La segona forma és la generalització. Aquestes tècniques busquen generalitzar o fer menys específics els atributs dels interessats modificant les respectives escales o ordres de magnitud (per exemple, substituir una ciutat per una regió). S'utilitzen tècniques com: les agregacions, k-anonimat, la l-diversitat (l-diversity, en anglès), la t-proximitat (t-closeness). Gràcies a aquestes tècniques, es pot mitigar el risc de la singularitat. Però encara queda el risc de la vincularitat. Per mitigar la vincularitat, s'han d'aplicar tècniques de pseudonimització. Consisteixen en substituir dades privades per identificadors o pseudònims. S'han fet servir tècniques com: la criptografia de clau simètrica, les funcions Hash, la descomposició per token, etc. Cal aclarir el fet que la pseudonimització no és un mètode d'anonimització, és una tècnica que mitiga el risc de la vincularitat[8].

A l'hora de publicar les dades del conjunt, s'ha de tindre en compte el risc de divulgació. Existeixen dos enfocament per limitar el risc de divulgació:

- Protecció interactiva: Consisteix a realitzar una consulta de dades sobre el conjunt de dades original i, a continuació, es retorna una versió protegida dels resultats.
- Protecció no interactiva: Es aquella que genera i allibera una versió protegida del conjunt de dades originals.

Si el tipus d'analisis de dades és desconeix en el moment de la publicació de les dades, la protecció no interactiva és l'única opció viable.

Un exemple real dels problemes esmentats anteriorment, és el que va patir l'empresa Netflix l'any 2006, quan va llençar el premi Netflix. Aquest esdevenimen posava a disposició de 500.000 usuaris cent milions de registres de pel·lícules, amb l'objectiu de recompensar a qui aconseguis millorar el seu servei de recomanació. Les dades facilitades estaven anonimitzades, però no van tenir en consideració tots els riscos.

Un grup d'investigadors va utilitzar com a font de dades externa la BBDD pública *Internet Movie Dataset* (IMDb), una BBDD online sense restricció d'access que conté dades sobre pel·lícules. Es va realitzar un experiment vinculant les dues BBDD, on es preguntava: Quant ha de saber una persona sobre un subscriptor de Netflix per identificar les seves dades a la BBDD? L'experiment va demostrar que sí que hi havia una relació entre les BBDD, concretament en les valoracions dels usuaris a les pel·lícules. Entre d'altres, va ser possible identificar a una usuaria. En concret, una mare de família homosexual, que mantenia la seva orientació sexual en secret, resident en una regió molt conservadora dels Estats Units (EEUU). Quan es va assabentar de la noticia, va demandar a l'empresa i com a resultat de l'escàndol que va formar aquesta notícia, un equip d'investigació de l'universitat de Texas, va demostrar que era possible identificar a un usuari si aquest habia qualificat 6 pel·lícules, que es trobaben a la BBDD IMDb en un 84% dels casos. Aquesta xifra augmentava fins a un 99% si havia qualificat 8 pel·lícules i si es sabia la data que es va fer la qualificació, amb un marge d'error de catorze dies. Fins i tot, dues valoracions amb un error de tres dies en realitzar les valoracións podien identificar al 68% dels usuaris de la plataforma Netflix [12].

2.1 Tècniques d'Aleatorització

2.1.1 Adició de Soroll

És una tècnica molt útil quan els atributs del conjunt causen un efecte advers sobre la persona. Consisteix en modificar el conjunt de dades de tal forma que siguin menys exactes, però sense afectar a la distribució general del conjunt. Un exemple senzill seria l'alçada, que normalment se sol donar amb precisió de centímetre (cm). El conjunt de dades anonimitzades tindria una precisió de ±10 cm. Amb la característica, de que si s'aplica de manera correcta, un tercer seria incapaç de restaurar les dades o esbrinar com s'han modificat.

2.1.2 Privacitat Diferencial

És una tècnica d'aleatorització que adopta un enfocament diferent de les altres tècniques. Les tècniques convencionals realitzen l'inserció del soroll abans del moment de la difusió de les dades, en canvi, la privacitat diferencial genera vistes anonimitzades del conjunt de dades. Les vistes anonimitzades, es duen a terme gràcies a les consultes de tercers. Les respostes de les consultes tenen soroll inserit amb posterioritat. Per aquest motiu, és important una supervisió en cadascuna de les noves consultes, per garantir que la possibilitat d'identificar a una persona és nul·la. Gracies a aquest mecanisme, es permet conservar les dades originals.

La principal avantatge d'aquest mètode, és que el conjunt de dades que s'entreguen a tercers autoritzats és una resposta a una consulta concreta, la qual no deixa exposada tota la informació de la BBDD[10].

2.2 Tècniques de Generalització

2.2.1 Agregació o Microagregació

És una tècnica que s'utilitza principalment per dades numèriques. Consisteix a generar grups d'un rang i substituir les dades pel valor mitjà del rang. Proporcionen d'aquesta forma una pertorbació en les dades. De forma que, amb un rang acord a les dades originals no afecta la distribució, però si a un tercer que no sàpiga el valor original.

2.2.2 *k*-anonimitat

Té com a objectiu impedir que una persona sigui singularitzada sobre un conjunt d'almenys *K* persones. Per aconseguir aquest objectiu, generalitza els valors dels atributs de les *K* persones al mateix valor. L'agregació i *k*-anonimitat són aplicables quan la correlació dels valors puntuals d'alguns registres, pot generar quasi identificadors[6].

2.2.3 l-diversitat

Aquest mètode exten la *k*-anonimitat, garantint que no es poden realitzar atacs per inferència determinista. Consisteix en assegurar que per cada classe d'equivalència, existeixen com a mínim L valors diferents. Això permet limitar l'ocurrència de les classes d'equivalència que tenen un nombre limitat d'atributs. Permeten que un atacant amb coneixements previs tingui cert grau d'incertesa[8].

2.2.4 t-proximitat

És el perfeccionament de la l-diversitat, que crea classes equivalents que mantenen la distribució inicial dels atributs de la taula. La tècnica permet mantenir les dades més pròximes a les originals. A més d'existir L valors diferents per cada classe, cada valor s'ha de reflectir tantes vegades com sigui necessari per mantenir la distribució inicial de cada atribut[8].

3 METODOLOGIA

La metodologia escollida és una metodologia de cascada[13], ja que les tasqués a realitzar són seqüencials i aquesta metodologia és la que millor s'adapta a aquesta situació. Aquest projecte s'iniciarà amb la recerca d'informació on es situa l'estat de l'art referent a articles, papers per revisar l'estat de l'art de l'anonimització de dades de caràcter sensible. També s'analitzarà quins mètodes existeixen i com poder implementar-los per solucionar els diferents problemes relacionats amb la preservació de la privacitat de les dades. Una vegada analitzada aquesta informació, s'escollirà un conjunt de dades (dataset) amb el qual s'implementaran alguns mètodes i tècniques trobades per generar dos tipus de dataset, un enfocat a preservar la utilitat de les dades i l'altre en augmentar la privacitat de les dades. Finalment, s'elaborarà un estudí empíric comparatiu entre els diferents datasets per saber la quantitat de soroll idònia a aplicar a cada atribut. La figura 1 recull l'esquema de procediments..

4 CREACIÓ DEL DATASET

El dataset que s'analitzarà en aquest estudi és una BBDD sobre clients d'assegurances mèdiques d'una empresa dels Estats Units (EEUU). S'ha escollit aquest dataset perquè conté valors de caràcter sensible a mes de tipus de dades habituals en les empreses. Disposa d'un total de 1.020 registres amb dades de clients, les quals contenen valors literals i numèrics. Aquesta BBDD conté els següents atributs: DNI, USBankNumber, edat, sexe, alçada, pes, índex de massa corporal (bmi), estat, ciutat, número de fills, fumador, religió, raça i despeses mèdiques individuals facturades per l'assegurança mèdica. D'aquests atributs, el DNI, USBankNumber, pes, alçada, ciutat, religió i raça són dades sintètiques afegides al dataset. La creació de les dades sintètiques forma part del desenvolupament d'aquest projecte, i s'ha implementat a partir d'una aplicació en Java.

Aquesta aplicació genera els valors sintètics de les dades de la següent forma:

- DNI: Es genera amb una funció que retorna un string format per vuit números aleatoris i una lletra de l'abecedari.
- USBankNumber: Generat en agafar un registre aleatori d'una llista de USBankNumbers vàlids als EEUU.
- Pes i Alçada: Es generen amb una funció que, amb el valor del bmi, busca una parella de valors que proporcioni el valor de bmi del client.

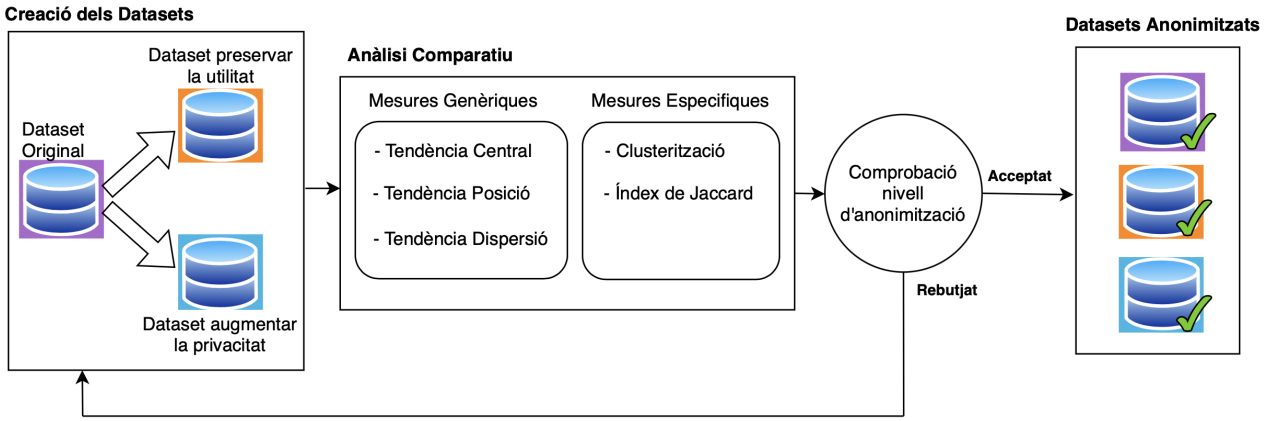

Fig. 1: Procediments de creació dels datasets

- Ciutat, Religió i Raça: Generat en agafar un registre aleatori d'unes llistes de ciutats, religions i races dels EEUU.

Tot seguit, per generar l'arxiu CSV, afegeix a aquest arxiu les dades sintètiques per cada client del dataset. A continuació, s'importen els arxius CSV a un document Excel per comprovar la correcta generació de l'arxiu CSV. Finalment, amb les dades de l'Excel es genera l'arxiu CSV del dataset01.

Per generar els datasets de preservar la utilitat i augmentar la privacitat, s'ha de classificar els atributs en identificadors, quasi-identificadors, atributs sensibles o atributs no sensibles. Una vegada classificat, per cada tipus es realitzarà una tècnica o altra. La taula 6 recull per cada atribut el tipus, tècnica i quantitat de soroll aplicada.

4.1 Dataset preservar la Utilitat

L'objectiu de generar aquesta versió del dataset original és obtenir una dataset que estigui anonimitzat, però que la quantitat d'anonimització sigui la suficient per emmascarar les dades, sense afectar a la utilitat dels valors. D'aquesta forma tenir un dataset anonimitzat, que poden utilitzar tots els treballadors de l'empresa o per entrenar amb més veracitat els models predictius, mantenint la utilitat de l'original amb la seguretat que les dades tenen una capa d'anonimització.

Per generar el dataset de preservar la utilitat, s'han fet servir les dades proporcionades per la funció *dUtility()* de la llibreria sdcMicro del llenguatge de programació R per estimar la quantitat de soroll en les dades quan s'aplica una quantitat determininada de soroll als valors de cada atribut [14]. La funció *dUtility()* ens permet obtenir un valor numèric perfecte per poder comparar amb els valors de les diferents quantitats de soroll. La taula1 recull els valors dels rangs de soroll emprats al dataset i subratllat de color groc el valor que millor resultat a donat a la funció *dUtility()*.

La funció *dUtility()* calcula la utilitat de les dades comparen les dades originals amb les dades pertorbades, per exemple, mitjançant les distàncies agregades des dels punts originals fins als valors corresponents a les dades pertorbades dividides per la desviació estàndard de cada variable [15]. El valor que proporciona la funció és un valor entre 0 i 1 on la utilitat és major si és més pròxim a 1.

Atributs	Rang de soroll
USBankNumber	5% , 10%, 15%, 20%, 30%
Edat	5% , 10%, 15%, 20%, 25%
Sexe	5% , 10%, 15%, 20%, 30%
Alcada	5% , 10%, 15%, 20%, 25%
Pes	5% , 10%, 15%, 20%, 25%
bmi	5% , 10%, 15%, 20%, 25%
Ciutat	5% , 10%, 15%, 20%, 30%
Num Fills	5% , 10%, 15%, 20%, 30%
Fumador	5% , 10%, 15%, 20%, 30%
Religio	5% , 10%, 15%, 20%, 30%
Raza	5% , 10%, 15%, 20%, 30%
Despeses	5% , 10%, 15%, 20%, 25%

TAULA 1: VALOR DEL RANG DE SOROLL APLICAT ALS ATRIBUTS DEL DATASET 1 AL DATASET DE PRESERVAR LA UTILITAT

La tècnica escollida pels atributs dependrà de si el valor és literal o numèric.

En cas de ser literal, s'ha escollit la tècnica de rankswap, pel fet que proporciona privacitat sense editar la dada. S'ha aplicat aquesta tècnica amb diferents valors de soroll per realitzar permutacions amb els valors en diferents rangs. Aquests rangs varien segons la quantitat de soroll. Si el valor de l'atribut és numèric s'han aplicat dues tècniques: la tècnica d'addició de soroll i la tècnica de microagregació.

Una vegada aplicat el resultat de la funció *dUtility()*, s'observa que els resultats han sigut molt inferiors respecte a la tècnica de microagregació, que no pas en la tècnica addició de soroll. Això és degut al fet que l'addició de soroll modifica cada valor individualment afegint una quantitat de soroll dins d'un rang preestablert. En canvi, la tècnica de microagregació, el que fa és, per un rang de valors suma aquests i realitza la mitjana i aplica el mateix valor a tot el rang. Per tant, la quantitat de soroll aplicada amb la tècnica addició de soroll és superior a la microagregació, perquè, els resultats de la funció *dUtility()* són majors en el cas d'addició de soroll que en microagregació. Aleshores, per generar el dataset de preservació de la utilitat s'han utilitzat els valors obtinguts en aplicar la tècnica microaggregation per

atributs numèrics i rankSwap per valors literals.

4.2 Dataset augmentar la Privacitat

L'objectiu del dataset de privacitat és molt diferent del centrat en la utilitat. Aquest busca aplicar la quantitat de soroll necessària per anonimitzar les dades de tal forma que un tercer no sigui capaç de trobar les dades originals. A més, aquest seria el dataset que es podria exportar cap a tercers, mantenint l'essència del dataset original però amb la seguretat de les dades gràcies a l'anonimització.

Per generar el dataset centrat en la privacitat s'han fet servir les dades proporcionades per la funció *dRisk()* de la llibreria sdcMicro del llenguatge de programació R, per saber la privacitat que aporta a les dades quan s'aplica una quantitat determininada de soroll als valors de cada atribut [14]. Gràcies a la funció, s'ha obtingut un valor numèric perfecte per poder comparar-lo amb els valors de les diferents quantitats de soroll. La taula2 recull els valors dels rangs de soroll utilitzats al dataset i subratllat de color groc el valor que millor resultat a donat a la funció *dRisk()*.

Atributs	Rang de soroll
USBankNumber	5%, 10%, 15%, 20%, 30%
Edat	1%, 2%, 3%, 4%, 5%
Sexe	5%, 10%, 15%, 20%, 30%
Alcada	10%, 20%, 30%, 40%, 50%
Pes	10%, 20%, 30%, 40%, 50%
bmi	5%, 10%, 15%, 20%, 25%
Ciutat	5%, 10%, 15%, 20%, 30%
Num Fills	5%, 10%, 15%, 20%, 30%
Fumador	5%, 10%, 15%, 20%, 30%
Religio	5%, 10%, 15%, 20%, 30%
Raza	5%, 10%, 15%, 20%, 30%
Despeses	50%, 100%, 250%, 500%, 1.000%

TAULA 2: VALOR DEL RANG DE SOROLL APLICAT ALS ATRIBUTS DEL DATASET 1 AL DATASET D'AUGMENTAR LA PRIVACITAT.

La funció *dRisk()*, ens proporciona una estimació del risc de divulgació basada en la distància, a través, dels intervals de les dades originals amb les pertorbades, basades en la desviació estàndard al voltant de les observacions [15]. El resultat de la funció, és un valor entre 0 i 1 on, si el valor és més pròxim a zero, menor risc de divulgació existeix.

La tècnica escollida són les mateixes que les utilitzades en el dataset de preservació de la utilitat. En cas de ser literal, s'ha escollit la tècnica de rankswap i en cas de ser numèric s'han aplicat dues tècniques: la tècnica d'addició de soroll i la tècnica de microagregació.

Una vegada aplicat el resultat de la funció *dRisk()*, s'observa que els resultats han sigut molt inferiors respecte a la tècnica d'addició de soroll, que no pas en la tècnica de microagregació. L'addició de soroll és superior a la microagregació pel fet que els resultats de la funció *dRisk()* són més pròxims a 0 que els resultats de la tècnica de microagregació. Aleshores, per generar el dataset privacitat s'han utilitzat els valors obtinguts en aplicar la tècnica addNoise per atributs numèrics i rankSwap per valors literals.

5 ANÀLISI COMPARATIU

L'anàlisi comparatiu, te com a objectiu comparar els diferents datasets generats (utilitat i privacitat), per comparar amb el dataset original si la quantitat de soroll aplicada a cada atribut del dataset ha sigut la idònia. Per comparar-los, s'han realitzat dos tipus de mesures; unes **genèriques**, que contenen dades més estàndard dels datasets, com són les mesures de tendència central, mesures de tendència posició i mesures de tendència dispersió. També s'han dut a terme unes mesures **específiques**, com és el cas de la clusterització i l'índex de Jaccard. Gràcies a aquestes mesures, es poden comparar els datasets i donar resposta a quin és el percentatge de soroll idoni per cada atribut.

5.1 Mesures Genèriques

5.1.1 Mesures de tendència central

Les mesures de tendència central, són mesures estadístiques que busquen resumir un conjunt de valors a un únic valor. Les mesures de tendència central més utilitzades són: la mitja, la mitjana i la moda. S'han calculat aquests valors per cada atribut de cada dataset, per comparar el dataset d'utilitat i privacitat amb les dades del dataset original.

5.1.2 Mesures de tendència posició

Les mesures de tendència de posició, permeten dividir els registres en parts iguals, per poder situar una dada en el dataset. Les mesures de tendència de posició que s'han utilitzat són: la freqüència de valors, la freqüència absoluta acumulada, la freqüència relativa acumulada i els quartils. Amb aquestes mesures, s'ha obtingut una divisió de les dades juntament amb la freqüència d'aparició de les dades del dataset.

La freqüència de valors, és una mesura que ens proporciona la quantitat de valors que hi ha a més del nombre de repeticions. La freqüència absoluta acumulada és el número de vegades que ha aparegut en el conjunt de dades un valor menor o igual que el de la variable. Com la freqüència absoluta, és una mesura que està influenciada per la mida del conjunt. Això descarta que sigui una mesura útil per poder comparar. Per aquest motiu és necessari calcular la freqüència relativa, que és el quocient entre la freqüència absoluta i la mida del conjunt de dades. Finalment, els quartils són una mesura que divideix el conjunt en quatre parts iguals. Els quartils són útils per calcular ràpidament la dispersió i la tendència central d'un conjunt de dades.

5.1.3 Mesures de tendència dispersió

Les mesures de tendència dispersió, mesuren el grau de dispersió que hi ha entre els valors de cada atribut del dataset. Les mesures que s'han calculat són: el rang, el mínim, el màxim, la variància, la desviació estàndard i el coeficient de variació. Gràcies a aquestes mesures s'ha calculat quan s'han dispersat les dades en el dataset de preservar la utilitat i augmentar la privacitat de les dades del dataset original.

5.2 Mesures Especifiques

5.2.1 Clusterització

La clusterització, és un mètode de quantificació vectorial, que té com a objectiu dividir n observacions en k clústers, on cada observació pertany al clúster amb la mitja més pròxima, al centroide del clúster, servint com prototip del clúster. El que resulta en una partició de l'espai de dades en cel·les. Una de les tècniques d'agrupació de clúster és el k-means. El k-means és un algoritme no supervisat que té una estreta relació amb el classificador de veïns k-nearest, una popular tècnica d'aprenentatge automàtic supervisat per a la classificació. Aplicar el classificador k-nearest als centroides dels clústers, obtinguts amb k-means, classifiquen noves dades als clústers existents.

Per poder dur a terme la clusterització, un dels mètodes més utilitzats en la literatura és el k-means. L'algoritme estàndard k-means es compon dels següents passos:

1. **Inicialització**: On s'escull la localització dels centroides dels k grups aleatòriament.

2. **Assignació**: Es calcula per cada dada del conjunt la distància euclidiana a cada centroide i s'assigna al centroide que tingui la distància més petita.

3. **Actualització**: Es torna a calcular la posició de cada centroide, amb la mitja aritmètica de les posicions de les dades de cada clúster.

Es repeteixen els passos 2 i 3 iterativament fins que no succeeixin més canvis. En aquest moment es pararia la iteració aconseguint la classificació de cada dada del conjunt amb els k clústers.

L'algoritme k-means té aplicacions com ara: la quantització vectorial, anàlisis de clústers o aprenentatge de característiques. En aquest estudi, s'ha centrat en l'anàlisi de clústers on l'algoritme k-means s'utilitza per particionar conjunts de dades, en aquest cas el Dataset01 [16].

S'ha de calcular la matriu de distàncies amb la funció *get_dist()* amb el mètode "euclidian". Seguidament, s'han de generar un seguit de gràfiques per comprovar el valor de clústers (k). Existeix una funció en el llenguatge R anomenada *NbClust()* que ho realitza de forma automàtica. Es caracteritza per comprovar 30 tècniques diferents que calculen el valor idoni de k. Una vegada calculada la funció, s'ha generat una gràfica 2 amb els resultats pel dataset original.

Una vegada se sap el valor del número de k, s'ha aplicat el mètode de clusterització kmeans amb la funció *kmeans()* indicant el dataset i el número de clústers. Per visualitzar el clúster s'ha generat una gràfica de núvol de punts indicant l'objecte generat per la funció kmeans i el dataset que s'ha treballat. A més, si es vol veure amb més detall quins registres s'agrupen s'ha generat el densiograma o dendograma a la figura 4.

Per analitzar millor els diferents clústers generats, s'ha generat una taula per fer una gràfica on es visualitza el valor mitjà de cada clúster per cada atribut del dataset. Per generar la taula s'ha utilitzat la funció *gather()* i per la gràfica la funció *ggplot()* [17] [18] .

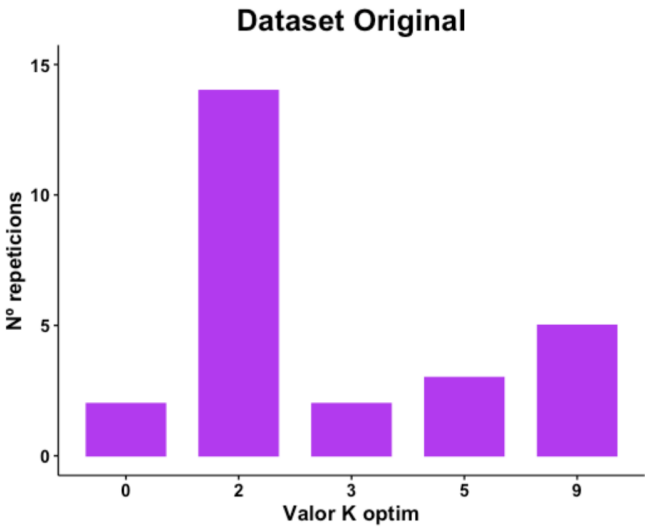

Fig. 2: Valor de k òptim pel dataset original.

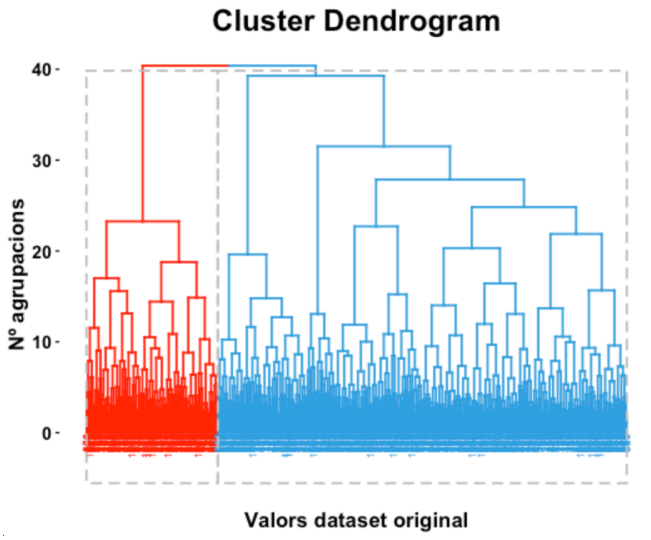

Fig. 3: Desiograma del dataset original amb valor $k = 2$.

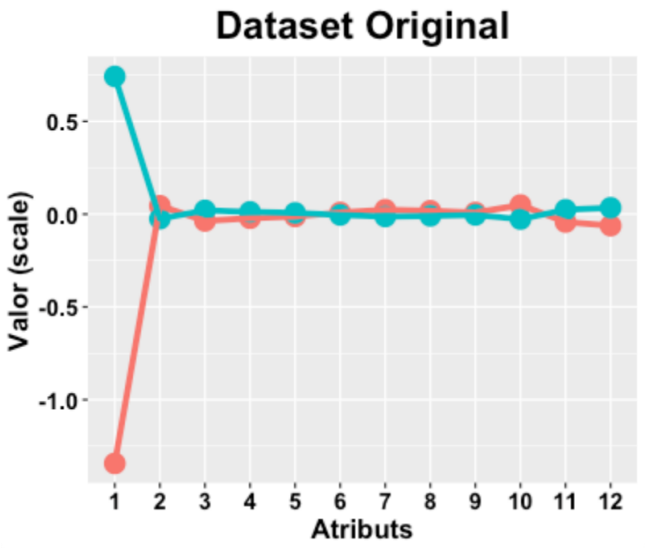

Fig. 4: Gràfica clusterització k-means $k = 2$.

5.2.2 Índex de Jaccard

L'Índex de Jaccard, és un valor entre 0 i 1 que ens indica el grau de similitud entre dos conjunts, on 1 indica una coincidència absoluta i 0 indica que no hi ha coincidència. Utilitza la fórmula:

$$\mathcal{J}(X,Y) = \frac{|X \bigcap Y|}{|X \bigcup Y|}$$

És a dir, la cardinalitat de la intersecció dels dos conjunts dividida per la cardinalitat de la seva unió.

6 RESULTATS

6.1 Mesures Genèriques

6.1.1 Mesures de tendència central

Pel que fa a l'atribut USBankNumber, el valor de la mitja, mitjana i moda no han patit cap canvi, pel fet que la tècnica utilitzada no afegeix soroll. L'atribut edat ha patit canvis: el dataset utilitat presenta una mitja més petita, el mateix valor de mitjana i la moda està igualada amb els valors 18 i 19. Pel que fa al dataset de privacitat, mostra una mitja encara més petita, el mateix valor de mitjana i el valor moda és 19 en comptes de 18. Per tant, el nombre de clients amb 18 i 19 anys destaquen sobre la resta de clients. L'atribut sexe no pateix cap canvi, tenint un valor de mitja de 0,49. Com a resultat, a la població del dataset predominen, lleugerament, les dones. Pel que fa a l'alçada, la mitja és d'1,75 metres, la mateixa en el dataset d'utilitat però més petita que la del dataset de privacitat. La mitjana és la mateixa per als tres datasets i la moda del dataset original i d'utilitat és de 2,05 metres, tanmateix al dataset de privacitat és 1,76 metres. Per tant, l'augment de soroll ha fet que les dades es centrin en el valor de la mitja. L'atribut pes ha patit canvis en la mitja entre els datasets original i d'utilitat amb el de privacitat, en el qual el valor és menor. La mitjana ha anat disminuint en relació amb la quantitat de soroll aplicada. La moda ha desaparegut en el dataset d'utilitat i de privacitat. L'atribut bmi ha patit canvis en les tres mesures; la mitja del dataset original i utilitat és la mateixa, no obstant el dataset de privacitat ha disminuint una mica de 31,08 a 31,06. La mitjana ha anat disminuïnt amb l'augment del soroll. La moda ha canviat, al dataset de privacitat té 6 elements, mentre que a l'original i al d'utilitat tenen 1. Els atributs ciutat, número de fills, fumador, religió i raça no han patit cap canvi, gràcies a la tècnica utilitzada per generar els valors, la qual no ha afegit cap soroll. Finalment, l'atribut despeses ha patit un canvi en la mitja, entre el dataset original i d'utilitat amb el dataset de privacitat. La mitjana és molt similar als datasets original i d'utilitat però, completament diferent del dataset de privacitat, passant de 97.000 a 510.000, cinc vegades més. Finalment, la moda ha desaparegut, tenint tots els valors amb el màxim de repeticions a causa de la tècnica aplicada.

6.1.2 Mesures de tendència posició

L'atribut USBankNumber, presenta els mateixos resultats pel que fa a les diferents mesures de tendència de posició a causa de la tècnica utilitzada, perquè permuta els valors i no afegeix soroll. Les dades mostren que dels tres diferents bancs que conté el dataset, un té el 35% dels clients, l'altre un 34% i l'últim un 31%. Per tant, la distribució dels bancs és bastant equitativa i la contractació dels bancs són molt similars. L'atribut Edat ha patit canvis entre els datasets, que es reflecteixen en la freqüència de valors i les freqüències acumulades, però els quartils no han patit cap canvi, en conseqüència, no són uns canvis molt significatius. Amb l'atribut sexe succeeix el mateix que amb l'atribut USBankNumber, el qual no ha variat en als diferents datasets. Per altra banda, l'atribut alçada ha patit canvis. Comparant el dataset original amb el d'utilitat, aquests són mínims, tanmateix, comparant amb el de privacitat, són més rellevants. Es nota significativament en les freqüències, així i tot, pel que fa als quartils varia solament en el 75%. Per tant, les dades dels últims registres del dataset són les que han patit més canvis. L'atribut pes ha patit una gran diferència entre el dataset de privacitat i els altres, perquè, els valors dels quartils han variat tots. Això indica que la quantitat de soroll aplicada al dataset de privacitat ha sigut suficient per canviar les dades, sobretot als extrems, fet que es reflectirà en les mesures de tendència de dispersió. Passa el mateix fenomen amb l'atribut bmi, per tant, la quantitat de soroll aplicada en el bmi indica que les dades han sigut emmascarades. L'atribut ciutat no varia entre els datasets a causa de la tècnica d'anonimització utilitzada. L'atribut nombre de fills pateix canvis, ja que la tècnica emprada ha generalitzat les dades eliminant el valor 5. Això ha resultat que els números de registres 3, 4 i 5 hagin variat al fusionar-se el 4 i el 5. A més d'agafar població del valor 3, que ha anat cap al valor 4. Encara que s'ha eliminat un valor, els quartils són els mateixos. Per tant, els registres afectats han estat una minoria perquè els quartils no han variat. Els següents atributs: fumador, religió i raça no han variat, igual que USBankNumber i sexe. Finalment, l'atribut despeses també ha patit un gran canvi, sobretot al dataset de privacitat, degut principalment a la quantitat excessiva de soroll aplicada, cosa que es nota per la generació de valors negatius.

6.1.3 Mesures de tendència dispersió

L'atribut USBankNumber no ha patit cap canvi, tenint el mateix valor de rang, mínim, màxim, variància, desviació estàndard i coeficient de variació. Pel que fa a l'atribut edat, ha patit variació al rang i mínim al dataset de privacitat. La variància ha anat disminuint amb l'augment del soroll i la desviació estàndard ha variat mínimament. L'atribut sexe no ha patit cap canvi. Per altra banda, l'atribut Alçada ha patit variació en totes les mesures al dataset de privacitat com indicaven les mesures de tendència anteriors. Destacar sobretot el canvi del mínim d'1,4 a 1,1 metres i màxim de 2,1 a 2,3 metres. L'atribut pes ha patit una lleu variació entre el dataset original i el d'utilitat, però, al dataset de privacitat el mínim i el màxim han variat ±20 kg. Això ha repercutit en què els extrems no tinguin consistència, generant un mínim de 9,91 kg per una persona adulta, el qual és impossible. Pel que fa al màxim passa de 210 kg en el dataset original a 220 kg en el de privacitat. No és impossible, així i tot, amb el mínim del dataset sí que ha generat un conflicte. Per tant, la quantitat de soroll aplicada ha sigut molt elevada. L'atribut bmi ha patit una petita variació, sobretot al dataset de privacitat, pel que fa al rang, mínim i màxim. L'atribut ciutat no ha patit cap canvi. L'atribut nombre de

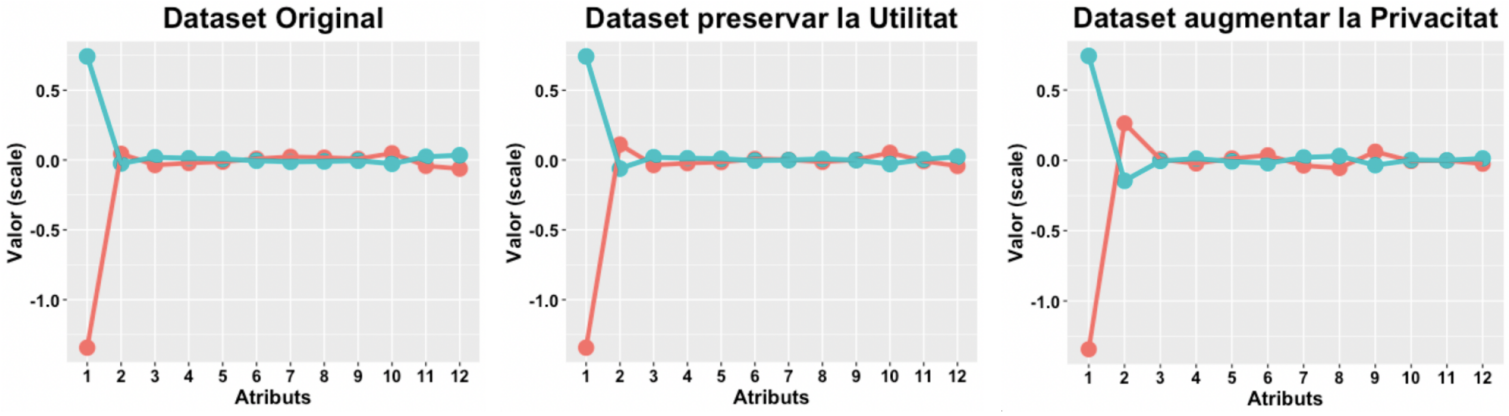

Fig. 5: Gràfiques de la clusterització $k = 2$ als dataset original, utilitat i privacitat

fills ha patit un canvi, a causa de l'eliminació del valor 5, el qual ha canviat el màxim, variància, desviació estàndard i coeficient de variació, aixi i tot de manera poc significativa. Els atributs fumador, religió i raça no han variat res, perquè la tècnica utilitzada no ha aplicat soroll. Finalment, l'atribut despeses ha canviat completament en totes les mesures. Al dataset original, al d'utilitat i amb més variació, al dataset de privacitat, a causa de l'excés de soroll aplicat.

6.2 Mesures Especifiques

6.2.1 Clusterització K = 2

Pel que fa a l'atribut USBankNumber (1), ha patit canvis mínims els quals no es poden apreciar a la gràfica 5. El següent atribut, edat (2), sí que ha patit una gran variació entre els valors, com mostra la gràfica 5 on l'augment de soroll ha generat una distanciació entre els valors dels clústers. Amb l'atribut sexe (3) el dataset original i utilitat, no han patit canvis significatius, però, al dataset privacitat sí que ha tingut canvis. S'aprecia, ja que els clusters estan l'un sobre de l'altre, pel fet que el soroll ha unificat els valors. Com s'aprecia a la gràfica 5.L'alçada (4) i el pes (5) han patit una distanciació entre els clústers, sobretot el pes. Aquest fet destaca, ja que segons les mesures genèriques, el pes ha tingut un major canvi.El bmi (6) al dataset de privacitat ha patit una distanciació a causa del soroll aplicat, el que indica una quantitat de soroll encertada. L'atribut ciutat (7) ha patit una generalització al dataset d'utilitat, perquè els clústers tenen la mateixa mitjana de valors. El dataset de privacitat ha tingut una permutació dels valors del clúster. Els valors del clúster vermell estan per sobre dels valors del clúster verd. Per tant, els valors dels clústers del dataset original i el dataset de privacitat estan intercanviats. L'atribut nombre de fills (8) tant als clústers del dataset d'utilitat com privacitat, estan permutats amb el dataset original, pero al dataset d'utilitat, els clústers estan més unificats i la mitja dels valors és més semblant. En canvi, al dataset privacitat als clústers hi ha més distanciació. La permutació ha sorgit degut a l'eliminació del valor 5 a l'atribut nombre de fills al utilitzar les tècniques d'aleatorització i generalització per anonimitzar el dataset. L'atribut fumador (9) ha patit canvis, sobretot en el dataset de privacitat, on els valors dels clústers s'han distanciat per la quantitat de soroll aplicada. Pel que fa a l'atribut religió (10), ha patit canvis, on l'augment de soroll ha fet que les dades es generalitzin més entre els clústers, com es mostra a la gràfica 5. L'atribut raça (11) ha patit canvis en el dataset de privacitat, generant distanciació dels clústers, el qual indica que la quantitat de soroll ha sigut encertada. Per últim, l'atribut despeses (12) ha patit una permutació dels clústers al dataset de privacitat, mentre que el d'utilitat es manté igual que l'original amb una generalització mínima dels valors dels clústers.

6.2.2 Clusterització K = 9

Pel que fa a la clusterització de $k = 9$ primerament esmentar que les dades que mostren les gràfiques, de la figura 6 es veu una similitud entre el dataset original i el d'utilitat, mentre que el dataset de privacitat és diferents dels altres datasets. Al USBankNumber (1), els nou clústers es divideixen en dos grups, on a la gràfica del dataset de privacitat la distanciació és més pronunciada per la quantitat de soroll aplicada. Per l'atribut Edat(2), els diferents clústers estan molt agrupats, amb valors molt pròxims entre ells. Al dataset de privacitat, aquesta agrupació es pot apreciar millor, com s'observa a la figura 6. Pel que fa a l'atribut sexe (3), tots els clústers són molt similars, amb una proporció molt homogènia. L'atribut alçada (4) l'agrupació és homogènia en els diferents clústers al dataset original i d'utilitat, però al dataset de privacitat hi ha distanciació entre els clústers, degut a la tècnica utilitzada. Amb els atributs pes (5), bmi (6) i ciutat (7), la relació entre el dataset original i utilitat és la mateixa, però el dataset privacitat, en comptes d'una distanciació, s'hi ha produit una generalització dels valors. L'atribut nombre de fills (8) és igual al dataset original i d'utilitat, amb la diferència que a l'original destaca un clúster amb valors més elevat de fills. Al dataset de privacitat, apareix una generalització dels diferents clústers. És degut al fet que en desaparèixer un tipus de valor les dades han resultat ser més homogènies. Al següent atribut, fumador (9), existeix una generalització dels valors del clúster al dataset d'utilitat i una dispersió al dataset de privacitat, a causa de la diferència de soroll aplicat per generar-los. L'atribut religió (10) és un cas particular, perquè l'augment de soroll ha provocat que els valors dels clústers siguin quasi els mateixos, desapareixent els extrems particulars que existeixen al dataset original i d'utilitat. A l'atribut raça (11) s'ha

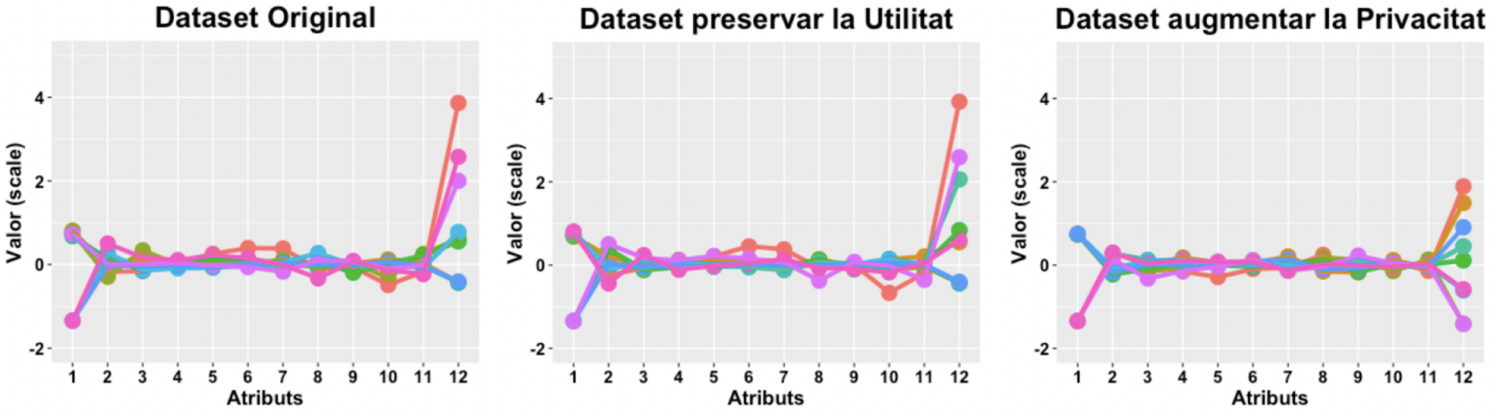

Fig. 6: Gràfiques de la clusterització $k = 9$ als dataset original, utilitat i privacitat

generat una generalització als valors dels clústers que va augmentant amb la quantitat de soroll aplicada. Finalment, a l'atribut Despeses (12), l'augment de soroll ha provocat generalització de les dades dels clústers, com es pot veure a la gràfica 6, on els valors afectats han sigut els extrems, generant nous valors més pròxims entre ells.

Per tant, la quantitat de soroll aplicada als següents atributs ha sigut encertada per la gran variació i distanciació amb les dades originals; USBankNumber (1), Edat (2), bmi (6), Fumador (9), Raça (11) i Despeses (12). Pels altres atributs s'hauria de tornar a realitzar la generació de les dades dels datasets, augmentant la quantitat de soroll aplicada, sobretot pel dataset de privacitat.

6.2.3 Índex de Jaccard

Una vegada creades les tres funcions per poder calcular l'índex de Jaccard amb aquest datasets, s'ha cridat a les funcions per generar la matriu de relació de clústers. S'han obtingut els següents resultats per la clusterització de $k = 2$:

	Clu Uti 1	Clu Uti 2	Clu Pri 1	Clu Pri 2
Clu Ori 1	74,55%	66,19%	71,94%	63,18%
Clu Ori 2	66,08%	74,46%	63.89%	71,67%

TAULA 3: MATRIU DE RELACIONS CLÚSTERS $k = 2$ DATASET 1

S'ha mostrat que existeix una relació entre els clústers 1 dels datasets original, d'utilitat i de privacitat i una relació entre els clústers 2 dels datasets original, d'utilitat i de privacitat. Mentre que si s'ha tornen a cridar les funcions per generar la matriu de relació de clústers, s'han obtingut els resultats per la clusterització de $k = 9$, es pot trobar a la figura 4 de l'apèndix 7.

Després d'analitzar la matriu proporcionada per la clusterització de $k = 9$, sorprèn el fet que hi ha hagut duplicitat amb els clústers del dataset de privacitat. Els clústers que han donat problemes de duplicitat són el quatre i el cinc. El que s'ha indicat en el dataset de privacitat no està format per nou grups, sinó per sis, com ens indica la taula 7. Cosa que no és d'estranyar, pel fet que la funció del càlcul del valor òptim de k, NbClust(), indica que el valor nou no és la millor opció sinó que era cinc, el qual, s'aproxima més al valor obtingut en la matriu de relacions. Però, per poder comparar els datasets amb els mateixos grups de clústers, s'han realitzat nou clústers en comptes de cinc.

7 CONCLUSIONS

Mesures Genèriques: De les dades obtingudes de les mesures genèriques de l'anàlisi comparatiu s'ha extret que: els atributs als quals se'ls ha aplicat la tècnica de rankswap no han tingut cap variació, és a dir, que les dades són les mateixes en els tres datasets, però en diferent ordre. En canvi, els atributs als quals se'ls ha aplicat una tècnica d'addicció de soroll o microagregació, les quals afegeixen soroll i, per tant, modifiquen les dades, han patit canvis en les mesures, on l'atribut edat ha sigut el que menor variació ha tingut, seguit del bmi, pes, alçada i finalment despeses. Per aquest últim, despeses, la variació ha sigut astronòmica, a causa de la quantitat excessiva de soroll aplicada. Repercutint en què la privatització ha sigut molt bona, variant molt les dades, les quals no es poden relacionar amb les originals, però la utilitat de les dades és nul·la. Tenint valors negatius, variància astronòmics, etc. Per tant, la dispersió de les dades ha canviat molt, fent que no hi hagi rcap relació amb les dades originals.

Mesures Específiques: De les dades obtingudes de les mesures específiques de l'anàlisi comparatiu s'ha extret que: Les dades del dataset estan molt agrupades entre elles, fent que la generació de núvols de punts tingui moltes col·lisions i, per tant, la visualització sigui molt difícil o pràcticament nul·la. Aleshores la generació de densiogrames o dendogrames ha sigut la solució, permetent veure la relació entre registres. En voler maximitzar el nombre de clústers de $k = 2$ a $k = 9$ per generar més subgrups, i d'aquesta forma, fer grups més dedicats o específics, el dataset de privacitat ha donat conflictes, perquè el valor de $k = 9$, segons les tècniques per calcular el valor k, no era una bona opció, era millor el valor 5. Això ha repercutit en la taula de relació de clústers, generant duplicitat de valors.

Conclusións finals: S'ha conclòs que els valors de soroll aplicats no han sigut els més idonis en els atributs: sexe, alçada, pes, ciutat, Num_Fills i religio. Que la relació que hi ha entre utilitat i privacitat és com una balança, on s'ha prioritzat una, l'altre decau, fent que trobar aquest equili-

bri no sigui un procés trivial. Per cada dataset, al tindre valors desiguals, fa que les quantitats de soroll a aplicar siguin diferents, tanmateix com les tècniques. Però, sí que és veritat que es pot arribar a generar una guia on tenir una idea de les tècniques a aplicar, però és un procés iteratiu pel fet que amb els resultats actuals se sap quins atributs editar. Per tant, s'haurà de tornar a repetir la generació dels datasets amb nous valors, per després tornar a fer l'anàlisi i veure quines millores s'han obtingut, per tornar a iterar tantes vegades com siguin necessàries fins a obtenir els valors desitjats o arribar a un punt en el qual no hi ha cap millora.

Concloure en què l'estudi de privacitat, no és un procés ràpid, sinó iteratiu, on depenent de les exigències que s'han tingut s'estendrà més o menys.

7.1 Treball futur

Per ampliar més l'estudi, S'haurien de realitzar més anàlisis, com ara, un exploratori per entendre millor el dataset i així acabar de optimitzar la quantitat de soroll idònia. Aplicar diferents tècniques de clusterització com ara: BIRCH, DBSCAN, GAUSIAN Mixture Model, etc. Aplicar més mètodes específics com el Knn. Fer mes iteracions per veure com van millorant els resultats dels anàlisis i els datasets.

8 AGRAÏMENTS

Agraeixo a l'exalumne Daniel Morales per ajudar-me amb la correcció de l'informe, aportant consells de redacció i comprensió lectora.

REFERÈNCIES

[1] K.Paranjape, M Schinkel, RN Panday, J Car, P Nanayakkara, "Introducing Artificial Intelligence Training in Medical Education" Amsterdam University Medical Center Vol 5, Nº 2, Septiembre 2019.

[2] ByteBridge (24 Septiembre 2021) Why the High-Quality Training Data is so Important to AI Machine Learning? [online]. Available: shorturl.at/cqrtR

[3] Unión Europea (30 Septiembre 2021) "Pólitica de Privacidad Protección de Datos Personales"[online]. Available: shorturl.at/mqKQ0

[4] Comisión Europea (30 Septiembre 2021) "Reglamento General de Protección de Datos"[online]. Available: shorturl.at/bqyYZ

[5] E.Gil, Big data, privacidad y protección de datos, XIX edición, España, Accésit, 2015.

[6] Solanas, Agusti & Ballesté, Antoni & DomingoFerrer, Josep & Bujalance, Susana & Mateo-Sanz, Josep. (2006). Métodos de microagregación para k-anonimato: privacidad en bases de datos.

[7] Agencia Española de Protección de Datos (22 Septiembre 2021) L·LA K-ANONIMIDAD COMO MEDIDA DE LA PRIVACIDAD"[online]. Available: shorturl.at/kADH6

[8] Dictamen 05/2014 sobre técnicas de anonimizacion, Vol 29, Derechos Fundamentales y Ciudadanía de la Unión Europea, Bruselas, Abril 2014.

[9] Agencia Española de Protección de Datos (22 Septiembre 2021) "Orientaciones y Garantias en los procedimientos de anonimizacion de datos personales"[online]. Available: shorturl.at/ginHS

[10] C. Dwork, A. Roth, The Algorithmic Foundations of Differential Privacy, Vol 9, Now The sense of Knoletge, 2014.

[11] Comisión Europea (30 Septiembre 2021) "¿Que datos personales se consideran sensibles?"[online]. Available: shorturl.at/nqtBT

[12] A. Narayanan,V. Shmatikov, Robust Deanonymization of Large Datasets(How to Break Anonymity of the Netflix Prize Dataset), The University of Texas at Austi, 2008.

[13] Digital Guide IONOS (March 11, 2019) "El modelo en cascada: desarrollo secuencial de software"[online]. Available: shorturl.at/wyS68

[14] cran.r-project (July 26, 2021) "Package sdcMicro"[online]. Available: shorturl.at/fkJWY

[15] Mateo-Sanz J, Domingo-Ferrer J, Sebé F (2005) Probabilistic information loss measures in confidentiality protection of continuous microdata. Data Mining and Knowledge Discovery 11:181–193

[16] Wikipedia (October 10, 2021) "*k*-means clustering"[online]. Available: shorturl.at/nwyE2

[17] Análisis de Clústers en R y Rstudio (20 Diciembre, 2021) YouTube [online]. Available: shorturl.at/xzIMR

[18] Documentació funció ipak R (20 Diciembre, 2021) Github [online]. Available: shorturl.at/hjFT2

APÈNDIX

Tasques	Data Inici	Data Final
Recerca d'informació sobre Privatització, Anonimització de dades.	20/09/21	15/10/21
Recerca dels Dataset	04/10/21	08/10/21
Generació del Dataset	18/10/21	29/10/21
Creació del Dataset Utilitat i Privacitat	01/11/21	12/11/21
Anàlisi Comparatiu Mesures Genèriques	15/11/21	26/11/21
Anàlisi Comparatiu Mesures Especifiques Dataset Utilitat	29/11/21	10/12/21
Anàlisi Comparatiu Mesures Especifiques Dataset Privacitat	13/12/21	24/12/21
Conclusions Anàlisi Comparatiu	27/12/21	07/01/22

TAULA 4: TASQUES A REALITZAR EN EL PROJECTE

Nom Camps	Classificacio Atribut	Classificacio Privacitat	Descripció
DNI	Numeric	Identificador	Identificador composat per 8 numeros i una lletra.
USBankNumber	Numeric	Quasi-Identificador	Valor numeric de 9 xifres que identifica un banc en els Estats Units.
Edat	Numeric	Quasi-Identificador	Valor numeric de l'edat del individu.
Sexe	Categoric Nominal	Atribut no sensible	Distingeix els usuaris de la BBDD per el seu sexe (home o dona).
Alcada	Numeric	Quasi-Identificador	Valor numeric de l'alçada del individu.
Pes	Numeric	Quasi-Identificador	Valor numeric en escala de Kilograms del pes del individu.
bmi	Numeric	Quasi-Identificador	Valor numeric que representa l'index de masa corporal obtingut de la formula pes(kg) / alcada(m^2)
Estat	Categoric Nominal	Atribut no sensible	Ens enumera l'estat on resideix l'individu
Ciutat	Categoric Nominal	Atribut sensible	Ens enumera la ciutat de residencia del individu entre un dels seguents valors: Los Angeles, San Diego, San Jose, San Francisco i Long Beach
Num Fills	Numeric	Atribut sensible	Valor numeric entre 0-5 amb el numero de fills del individu.
Fumador	Categoric Nominal	Atribut no sensible	Valor Categoric que ens indica si fuma o no.
Religio	Categoric Nominal	Atribut sensible	Valor Categoric que indica la religio del individu.
Raza	Categoric Nominal	Atribut sensible	Valor Categoric que indica la raza del individu.
Despeses	Numeric	Quasi-Identificador	Valor en dolars de les despeses medicas del individu.

TAULA 5: CLASSIFICACIÓ ATRIBUTS DEL DATASET01.

Atribut	Tipus	Tècnica	Valors Soroll
DNI	Identificador	Eliminació	—
USBankNumber	Quasi-Identificador	RankSwap	0.05, 0.10, 0.15, 0.20, 0.30
Edat	Quasi-Identificador	Microagregació	5, 10, 15, 20, 25
		AddNoise	1, 2, 3, 4, 5
Sexe	Atribut No Sensible	RankSwap	0.05, 0.10, 0.15, 0.20, 0.30
Alcada	Quasi-Identificador	Microagregació	5, 10, 15, 20, 25
		AddNoise	10, 20, 30, 40, 50
Pes	Quasi-Identificador	Microagregació	5, 10, 15, 20, 25
		AddNoise	10, 20, 30, 40, 50
bmi	Quasi-Identificador	Microagregació	5, 10, 15, 20, 25
		AddNoise	5, 10, 15, 20, 25
Estat	Atribut No Sensible	—	—
Ciutat	Atribut Sensible	RankSwap	0.05, 0.10, 0.15, 0.20, 0.30
Num fills	Atribut Sensible	RankSwap	0.05, 0.10, 0.15, 0.20, 0.30
Fumador	Atribut No Sensible	RankSwap	0.05, 0.10, 0.15, 0.20, 0.30
Religio	Atribut Sensible	RankSwap	0.05, 0.10, 0.15, 0.20, 0.30
Raza	Atribut Sensible	RankSwap	0.05, 0.10, 0.15, 0.20, 0.30
Despeses	Quasi-Identificador	Microagregació	5, 10, 15, 20, 25
		AddNoise	50, 100, 250, 500, 1000

TAULA 6: TÈCNIQUES APLICADAS ELS ATRIBUTS DEL DATASET01.

	Ori_C1	Ori_C2	Ori_C3	Ori_C4	Ori_C5	Ori_C6	Ori_C7	Ori_C8	Ori_C9	Uti_C1	Uti_C2	Uti_C3	Uti_C4	Uti_C5	Uti_C6	Uti_C7	Uti_C8	Uti_C9	Pri_C1	Pri_C2	Pri_C3	Pri_C4	Pri_C5	Pri_C6	Pri_C7	Pri_C8	Pri_C9
Ori_C1	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,737	0,559	0,577	0,511	0,623	0,578	0,535	0,542	0,588	0,604	0,532	0,608	0,614	0,529	0,614	0,615	0,531	0,527
Ori_C2	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,563	0,569	0,651	0,576	0,629	0,742	0,654	0,532	0,635	0,622	0,585	0,641	0,672	0,623	0,666	0,666	0,578	0,612
Ori_C3	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,587	0,548	0,564	0,543	0,579	0,650	0,566	0,540	0,728	0,576	0,548	0,595	0,596	0,557	0,592	0,593	0,540	0,548
Ori_C4	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,548	0,710	0,662	0,553	0,593	0,577	0,578	0,526	0,536	0,583	0,553	0,603	0,609	0,571	0,607	0,604	0,545	0,558
Ori_C5	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,563	0,651	0,745	0,580	0,629	0,651	0,659	0,531	0,550	0,624	0,585	0,645	0,673	0,628	0,671	0,671	0,582	0,614
Ori_C6	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,506	0,550	0,582	0,725	0,553	0,577	0,664	0,598	0,531	0,553	0,637	0,559	0,569	0,651	0,567	0,566	0,628	0,650
Ori_C7	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,521	0,568	0,658	0,662	0,588	0,652	0,745	0,615	0,550	0,581	0,667	0,602	0,634	0,712	0,631	0,627	0,665	0,698
Ori_C8	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,612	0,588	0,629	0,551	0,709	0,631	0,587	0,520	0,571	0,642	0,554	0,639	0,657	0,569	0,660	0,662	0,552	0,573
Ori_C9	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,541	0,536	0,544	0,618	0,528	0,547	0,628	0,738	0,537	0,528	0,610	0,536	0,536	0,617	0,534	0,535	0,603	0,616
Uti_C1	0,737	0,563	0,587	0,548	0,563	0,506	0,521	0,612	0,541	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,603	0,529	0,604	0,611	0,525	0,612	0,612	0,529	0,522
Uti_C2	0,559	0,569	0,548	0,710	0,651	0,550	0,568	0,588	0,536	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,591	0,557	0,605	0,616	0,574	0,613	0,609	0,557	0,567
Uti_C3	0,577	0,651	0,564	0,662	0,745	0,582	0,658	0,629	0,544	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,637	0,598	0,658	0,687	0,640	0,684	0,682	0,595	0,626
Uti_C4	0,511	0,576	0,543	0,553	0,580	0,725	0,662	0,551	0,618	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,561	0,648	0,565	0,584	0,664	0,579	0,579	0,642	0,663
Uti_C5	0,623	0,629	0,579	0,593	0,629	0,553	0,588	0,709	0,528	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,660	0,571	0,655	0,673	0,582	0,675	0,674	0,563	0,581
Uti_C6	0,578	0,742	0,650	0,577	0,651	0,577	0,652	0,631	0,547	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,637	0,597	0,652	0,686	0,635	0,680	0,678	0,592	0,623
Uti_C7	0,535	0,654	0,566	0,578	0,659	0,664	0,745	0,587	0,628	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,595	0,681	0,616	0,648	0,725	0,643	0,641	0,678	0,710
Uti_C8	0,542	0,532	0,540	0,526	0,531	0,598	0,615	0,520	0,738	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,524	0,609	0,532	0,534	0,615	0,533	0,532	0,601	0,615
Uti_C9	0,588	0,635	0,728	0,536	0,550	0,531	0,550	0,571	0,537	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,575	0,544	0,584	0,591	0,551	0,587	0,588	0,539	0,541
Pri_C1	0,604	0,622	0,576	0,583	0,624	0,553	0,581	0,642	0,528	0,603	0,591	0,637	0,561	0,660	0,637	0,595	0,524	0,575	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Pri_C2	0,532	0,585	0,548	0,553	0,585	0,637	0,667	0,554	0,610	0,529	0,557	0,598	0,648	0,571	0,597	0,681	0,609	0,544	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Pri_C3	0,608	0,641	0,595	0,603	0,645	0,559	0,602	0,639	0,536	0,604	0,605	0,658	0,565	0,655	0,652	0,616	0,532	0,584	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Pri_C4	0,614	0,672	0,596	0,609	0,673	0,569	0,634	0,657	0,536	0,611	0,616	0,687	0,584	0,673	0,686	0,648	0,534	0,591	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Pri_C5	0,529	0,623	0,557	0,571	0,628	0,651	0,712	0,569	0,617	0,525	0,574	0,640	0,664	0,582	0,635	0,725	0,615	0,551	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Pri_C6	0,614	0,666	0,592	0,607	0,671	0,567	0,631	0,660	0,534	0,612	0,613	0,684	0,579	0,675	0,680	0,643	0,533	0,587	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Pri_C7	0,615	0,666	0,593	0,604	0,671	0,566	0,627	0,662	0,535	0,612	0,609	0,682	0,579	0,674	0,678	0,641	0,532	0,588	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Pri_C8	0,531	0,578	0,540	0,545	0,582	0,628	0,665	0,552	0,603	0,529	0,557	0,595	0,642	0,563	0,592	0,678	0,601	0,539	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Pri_C9	0,527	0,612	0,548	0,558	0,614	0,650	0,698	0,573	0,616	0,522	0,567	0,626	0,663	0,581	0,623	0,710	0,615	0,541	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000

Fig. 7: Matriu de relacions clústers $k = 9$ Dataset 1