Curs 2018-2019

Pràctica 2. Privacitat i anonimització de dades

1 Informació important

Setmanes dedicades a la pràctica 2:
4 i 11 d'abril
2, 9 i 16 de maig

Dates de lliurament:
19 de maig (fins les 23:59)

Objectes d'avaluació per aquesta pràctica:

1. **Informe** en PDF que reculli les respostes a totes les preguntes o qüestions formulades a l'enunciat. L'informe ha de seguir les següents consideracions:
   (a) El nom del fitxer ha de ser *GIS_{A,B,C,D,E}{1..12}_P2.pdf*.
   (b) En la capçalera cal indicar, clarament, el codi grup (i.e. GIS_{A,B,C,D,E}{1..12}), els noms dels integrants del grup i el NIU corresponent.
   (c) L'**extensió máxima** de l'informe és de **10 pàgines**. Sobrepassar aquesta extensió por comportar penalitzacions en la nota de la pràctica.
2. **Codi font** complet de la segona part de la pràctica (exercici 10).

Els dos arxius s'hauran de lliurar per l'Aula Moodle a través d'una tramesa que s'obrirà a tal efecte. Només cal realitzar un lliurament per equip de treball, especificant el nom dels integrants de l'equip en els comentaris de la tramesa.

Consideracions addicionals:

- No només s'avaluarà el contingut tècnic de l'informe sinó que també s'avaluarà la redacció i estil del document.
- Qualsevol intent de copia serà penalitzat amb una qualificació de **0 punts** per tots els grups implicats.
- En cap cas s'acceptaran entregues fora dels terminis establerts.

2 Introducció

En aquesta pràctica veurem temes relacionats amb la privacitat de les dades. En concret, ens centrarem en la preservació de la privacitat en processos de publicació de dades.

2.1 Problemàtica i context

La mineria de dades (*data mining*, en anglès) és el procés d'extreure informació útil, interessant, i desconeguda fins al moment de conjunts de dades. L'èxit de la mineria de dades es basa en la disponibilitat de dades de qualitat sobre els quals executar aquests processos. En aquest sentit, la recopilació de la informació digital per part de governs, corporacions i individus ha de facilitar l'intercanvi i la disponibilitat de dades a gran escala per a la seva posterior anàlisi. Hi ha una demanda d'intercanvi de dades entre diversos actors impulsada, d'una banda, pels beneficis mutus i, d'altra banda, per les regulacions que requereixen que certes dades siguin publicades. Generalment, la publicació de dades en obert (*open data*, en anglès) que incloguin dades personals pot induir a problemes de privacitat, si no es tracten de forma adequada.

Una tasca de gran importància és el desenvolupament de mètodes i eines que permetin la publicació de dades, de manera que les dades publicades mantinguin la seva utilitat a la vegada que preserven la privacitat dels usuaris que hi apareixen. Aquest procés s'anomena preservació de la privacitat en la publicació de dades (*privacy-preserving data publishing* o PPDP, en anglès), que pot ser vist com una resposta de caràcter tècnic per a complementar les polítiques de privacitat que cada país o regió implementa.

L'escenari típic de la recopilació i publicació de dades es descriu a la figura 1. En la fase de recollida de dades, el titular de les dades recopila informació dels diferents usuaris. A continuació, el propietari de les dades recopilades, dades originals en la figura, ha de protegir i assegurar la privacitat dels usuaris que apareixen abans de fer públic el conjunt de dades recopilades. Aquest procés, anomenat anonimització o preservació de la privacitat, serà l'encarregat d'assegurar que no és possible identificar un usuari dins el conjunt de dades protegides o anònimes. En la fase de publicació de les dades, el propietari de les dades recopilades publica les dades protegides per a la posterior explotació. Aquesta publicació de dades protegides es pot fer de forma pública i accessible per a qualsevol persona o entitat, o bé de forma privada a un conjunt d'empreses o centres autoritzats.

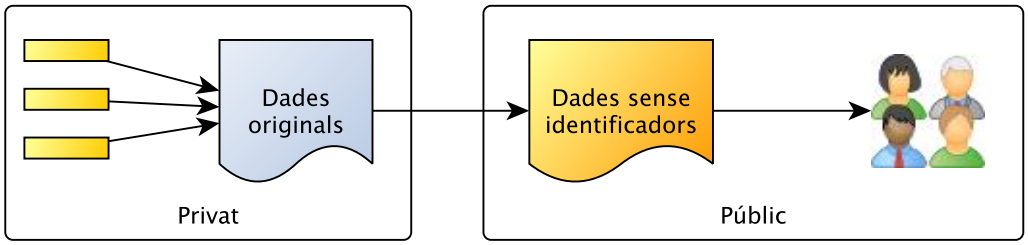

Figure 1: Escenari bàsic de publicació de dades

Per exemple, un hospital recull dades dels pacients i comparteix els registres dels pacients amb un centre mèdic extern. En aquest exemple, l'hospital és el titular de les dades, els pacients són propietaris de les seves pròpies dades i el centre mèdic extern és el receptor de les dades. Les tasques de mineria de dades que el centre mèdic extern pot realitzar sobre les dades protegides poden ser de qualsevol tipus, des d'un simple recompte del nombre d'homes amb diabetis fins a un sofisticat anàlisi de grups de pacients segons les seves característiques fisiològiques i demogràfiques.

Des del punt de vista de la privacitat o anonimització, els atributs d'un conjunt de dades es divideixen en quatre classes, segons el tipus d'informació que contenen:

- Els **identificadors** són un conjunt d'atributs que permeten identificar de forma explícita a un individu. El nom, DNI o número de la seguretat social són exemples d'atributs identificadors.
- Els **quasi-identificadors** són un conjunt d'atributs que potencialment podrien identificar a un individu.
- Els **atributs sensibles** presenten informació específica i sensible d'un individu en concret, com ara les malalties que pateix, el seu salari o les seves preferències sexuals o religioses.
- Finalment, els **atributs no sensibles** són tots els atributs que no caben en cap de les categories anteriors.

Òbviament, els atributs identificadors han de ser eliminats abans de publicar les dades. En cas contrari, la identificació dels usuaris és directa i trivial en les dades publicades.

Encara eliminant tots els identificadors, un estudi de Sweeney l'any 2002 [2] va aconseguir trencar la privacitat d'un governador d'Estats Units. En aquest treball, el nom i altres dades públiques del cens electoral van ser combinats amb una base de dades mèdiques utilitzant el codi postal, la data de naixement i el gènere.

En l'exemple anterior, la identitat d'un registre es veu compromesa a través de la combinació de diferents quasi-identificadors. Per dur a terme aquest tipus d'atacs a la privacitat dels usuaris és necessari que l'atacant tingui cert coneixement extern de l'usuari objectiu. Aquest coneixement pot ser obtingut a través de moltes i molt diverses fonts. Per exemple, un atacant pot adonar-se que el seu cap ha estat hospitalitzat durant uns dies concrets, per tant pot saber que apareixerà en els registres que es facin públics dels pacients de l'hospital. D'altra banda, no li serà molt difícil descobrir el codi postal, data de naixement i gènere de l'individu en qüestió, amb la qual cosa pot efectuar aquest atac per descobrir informació confidencial del seu cap.

Per evitar aquest tipus d'atacs, el propietari de les dades d'aplicar una sèrie d'operacions sobre els quasi-identificadors per impedir que puguin usar per identificar a qualsevol usuari dins de les dades protegides.

La anonimització o PPDP persegueix ocultar la identitat i la informació sensible dels usuaris que apareixen en conjunts de dades, assumint al mateix temps que cal preservar al màxim la utilitat de les dades en el conjunt de dades protegides. És a dir, l'anàlisi executat sobre les dades protegides ha de revelar informació útil i veritable, de manera similar als resultats que s'obtindrien en les mateixes anàlisis utilitzant les dades originals (no protegides).

Hi ha multitud d'estratègies d'anonimització, però en general aquest tipus de tècniques busquen les formes d'ocultar els detalls que puguin fer a un individu únic dins del conjunt de dades. L'objectiu és que un únic individu sigui indistingible respecte a un conjunt d'individus prou gran per protegir la seva identitat, de manera que l'atacant només pot deduir certa informació amb una certa probabilitat. La figura 2 mostra que, a diferència de l'escenari bàsic de publicació de dades vist anteriorment, en aquest cas s'afegeix la tasca d'anonimització o PPDP prèviament a la publicació de les dades.

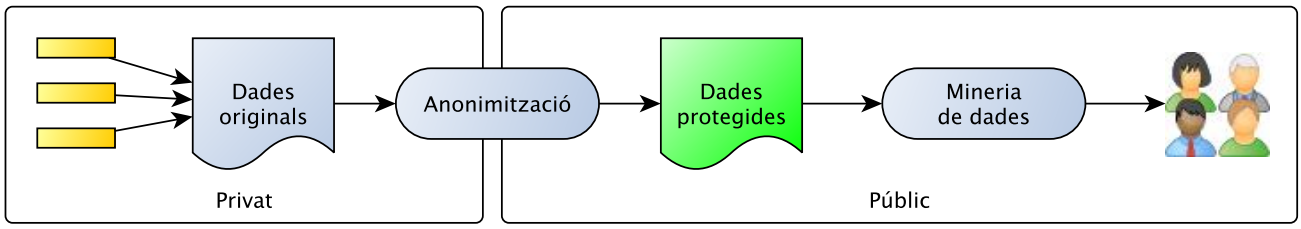

Figure 2: Escenari bàsic considerant la preservació de la privacitat en la publicació de dades

3 Mètodes d'anonimització

Essencialment, hi ha dos enfocaments principals per limitar el risc de divulgació en processos de publicació de dades:

- **Protecció no interactiva**, mitjançant la qual es genera i s'allibera una versió protegida del conjunt de dades original recopilat dels subjectes de dades.
- **Protecció interactiva**, mitjançant la qual es realitza una consulta de dades amb finalitats analítiques anàlisi en el conjunt de dades original i, a continuació, es retorna una versió protegida dels resultats a l'usuari que ha realitzat la consulta.

Quan el tipus d'anàlisi de dades és desconegut en el moment de la protecció i publicació de dades, la protecció no interactiva és l'única solució viable. No obstant això, per a una anàlisi de dades conegut prèviament, es pot optar per la protecció interactiva, ja que permet ajustar el nivell de protecció a l'anàlisi que s'està fent, el que teòricament permet maximitzar la utilitat i precisió dels resultats.

3.1 Anonimització aleatòria

El model d'aleatorització o pertorbació de dades és un mètode de protecció no interactiva. Consisteix, simplement, a introduir soroll en les dades originals, de manera que un atacant no pugui saber, del cert, si la informació que està extraient és certa o ha estat alterada durant aquest procés d'anonimització aleatòria.

No obstant això, si s'introdueix massa soroll en les dades originals, la privacitat quedarà preservada, però la utilitat de les dades pot arribar a ser nul·la. En efecte, quan el grau de pertorbació o introducció de soroll en les dades originals és molt elevat, estem generant dades aleatòries, de manera que el risc de trencar la privacidad serà nul·la, però també serà nul·la la informació que es pot extreure de les dades.

Per tant, un procés d'aleatorització ha d'introduir una quantitat de soroll que:

- sigui suficient perquè un atacant no pugui estar segur de la veracitat d'una dada en concret,
- i al mateix temps, les dades i la informació general del conjunt de dades s'ha de preservar perquè les anàlisis realitzades sobre el conjunt de dades anònims sigui proper (tan pròxim com sigui possible) al resultat obtingut utilitzant el conjunt de dades original.

3.2 k-Anonimitat

El model de k-anonimitat, introduït per Samarati [3] i Sweeney [2], és un dels models de protecció no interactiva més àmpliament investigat i emprat en la publicació de dades.

La k-anonimitat és una propietat de les dades que garanteix que un individu no pugui ser distingit d'altres $k-1$ individus també representats en aquestes dades. Per aconseguir aquest objectiu, podem aplicar diferents tècniques diferents, com ara reemplaçar valors concrets per altres valors d'una categoria més general o eliminar certs valors.

4 Anonimització de dades tabulars

Tradicionalment, les dades s'han presentat en forma de taules, on cada fila correspon a un registre o element (un usuari en el nostre cas) i cada columna correspon a un atribut, propietat o característica. Per tant, cada registre té un valor concret per a cada un dels atributs de la taula. La taula 1 mostra el format típic d'un conjunt de dades en forma de taula. En aquest cas, podem veure un total de n files o registres i m atributs.

	Atribut 1	**Atribut** 2	**...**	**Atribut** m
Registre 1	$valor_{1,1}$	$valor_{1,2}$	...	$valor_{1,m}$
...	...	...	...	...
Registre n	$valor_{n,1}$	$valor_{n,2}$	...	$valor_{n,m}$

Table 1: Exemple de taula amb n registres i m atributs

4.1 Mètodes d'emmascarament

Els mètodes d'emmascarament permeten modificar les dades originals amb l'objectiu d'impedir o dificultar la identificació d'un usuari en les dades protegides. Aquests mètodes es poden classificar en tres categories bàsiques en funció de com es manipulen les dades originals per definir el conjunt de dades protegides.

- **Mètodes pertorbatius**. El conjunt de dades original és pertorbat d'alguna manera, i el nou conjunt de dades pot contenir informació errònia. Per exemple, es pot introduir soroll en alguns atributs, és a dir, alterar el seu valor de forma més o menys aleatòria. D'aquesta manera, algunes combinacions de valors desapareixen en el conjunt de dades protegides. Alhora, les combinacions en les dades protegides ja no corresponen als del conjunt de dades original. Aquesta ofuscació dificulta la identificació d'usuaris en el conjunt de dades protegides per part dels atacants.
- **Mètodes no pertorbatius**. La protecció s'aconsegueix a través de la substitució del valor original per un altre valor que no és incorrecte però és menys específic, és a dir, més general. Per exemple, reemplacem un nombre per un interval. En general, els mètodes no pertorbatius redueixen el nivell de detall del conjunt de dades. Aquesta reducció del nivell de detall provoca que diferents registres tinguin les mateixes combinacions de valors, la qual cosa dificulta la identificació d'usuaris per part d'un atacant.
- **Generadors de dades sintètiques**. En aquest cas, en lloc de distorsionar les dades originals, es creen noves dades artificials per substituir els valors originals. Formalment, els generadors de dades sintètics construeixen un model de dades noves a partir del conjunt de dades original i, posteriorment, generen de forma aleatòria un nou (i protegit) conjunt de dades que si bé segueix les pautes de les dades originals, no conté informació privada de cap usuari. Per exemple, podem substituir l'edat d'un conjunt d'individus per valors aleatoris generats a partir del valor mitjà i la variància observada en les dades originals.

D'altra banda, els mètodes d'emmascarament han de considerar la peculiaritats dels diferents tipus d'atributs per reduir les probabilitats d'identificació mentre es manté la utilitat de les dades protegides. Els atributs es poden classificar en dues categories bàsiques:

- Els **atributs numèrics** permeten realitzar operacions aritmètiques entre ells, com per exemple, la resta o addició. Els ingressos i l'edat són exemples típics de tals atributs. Pel que fa al risc d'identificació, els valors dels atributs numèrics són propensos a ser únics en una base de dades i, per tant, poden provocar la identificació d'usuaris si no es prenen mesures oportunes per a la seva anonimització.
- Els **atributs categòrics** poden prendre valors en un conjunt finit i les operacions numèriques estàndard no tenen sentit en aquest tipus d'atributs. Podem distingir tres grups principals dins dels atributs categòrics:
  - **Nominals**. El valor d'aquests atributs es representa mitjançant etiquetes que proporcionen informació. Per exemple, el color dels cabells o l'estat civil són atributs categòrics nominals.
  - **Ordinals**. En aquest cas els atributs presenten un ordre o escala rellevant entre ells. Per exemple, el nivell d'estudis (primària, secundària, batxillerat, grau, etc.) és un atribut categòric ordinal. En aquests atributs les operacions de mínim i màxim tenen sentit.
  - **Estructurats**. Aquests atributs mantenen una relació de classe i subclasse entre ells. Per exemple, les professions poden seguir una jerarquia donada, on per exemple dins de la classe "metge" podem trobar molts especialitzacions, com ara "ginecòleg", "pediatre", etc. En alguns casos la jerarquia pot ser explícita i en altres es pot inferir a partir dels valors donats i de les seves relacions.

A continuació veurem cadascun dels tres tipus de mètodes d'emmascarament i presentarem alguns mètodes concrets i exemples que faciliten la comprensió del seu funcionament.

4.1.1 mètodes pertorbatius

Els mètodes pertorbatius alteren les dades i introdueixen soroll per dificultar el procés d'identificació d'un usuari per part d'un atacant. Hi ha multitud de mètodes en la literatura, encara que una revisió completa de tots ells escapa als objectius d'aquesta pràctica i només veurem aquí alguns dels mètodes més comuns i utilitzats en l'actualitat.

El primer mètode que veurem és conegut com a **soroll additiu** (*additive noise*) i consisteix a afegir distorsió o soroll en les dades originals, ja sigui seguint o no la mateixa distribució de les dades originals. Un exemple simple d'aquest tipus de mètode és introduir el soroll seguint una distribució normal $N(0, p\sigma)$, on σ representa la desviació estàndard de les dades originals i p és el paràmetre que controla la quantitat de soroll introduït en les dades [1]. Aquest mètode va ser desenvolupat originalment per a tractar amb atributs numèrics, encara que posteriorment s'han afegit extensions per poder treballar amb atributs categòrics.

El segon mètode que veurem, i que és àmpliament conegut i utilitzat per empreses i administracions, és conegut com **micro-agregació** (*microaggregation*). Aquest mètode es basa en crear grups de dades segons la seva similitud, i posteriorment reemplaçar el valor original de cada dada pel valor mitjà de tots els valors del grup al qual pertany. Per tant, per a cada valor específic d'un o més atributs existiran sempre un conjunt de registres; mai un registre únic que permeti que sigui identificat un usuari. Aquest mètode es pot aplicar sobre un únic atribut, i llavors es coneix com a micro-agregació univariant, o sobre dos o més atributs, i en aquest cas es coneix com a micro-agregació multivariant.

El mètode permet decidir quants registres s'ajunten en un mateix grup. És important notar que com més registres s'ajunten en cada grup, més gran és el nivell de privacitat que s'aconsegueix, però també és més gran el nivell de soroll. Igual que el mètode anterior, aquest també va ser desenvolupat per atributs numèrics, encara que posteriorment també s'han desenvolupat extensions que permeten tractar amb atributs categòrics.

Per finalitzar amb els mètodes pertorbatius, veurem un tercer mètode que també és molt conegut i utilitzat en entorns empresarials i governamentals. El mètode, conegut com a **intercanvi de rang** (*rank swapping*, en anglès), es basa en intercanviar aleatòriament els valors d'un mateix atribut entre diferents registres. Per aconseguir que les dades no siguin excessivament pertorbades, aquest mètode ordena tots els valors de l'atribut presents a la taula, i tot seguit l'intercanvi es realitza entre valors que es troben dins d'un rang acotat i definit com a paràmetre del mètode. D'aquesta manera s'intenta minimitzar el soroll i mantenir la utilitat de les dades protegides. Aquest mètode pot ser aplicat a atributs numèrics i categòrics ordinals.

4.1.2 Mètodes no pertorbatius

A diferència dels mètodes vistos a la secció anterior, els mètodes que veurem en aquesta secció no introdueixen soroll o distorsió en les dades originals. La informació protegida que es publica continua sent totalment veritable, encara que es generalitzen o suprimeixen algunes parts de la informació que podrien ajudar a un atacant a identificar de forma única a un usuari dins de les dades protegides.

Distingirem dos mètodes bàsics, d'una banda la generalització d'atributs, i per un altre la supressió d'atributs.

El mètode de **generalització** s'aplica normalment a atributs categòrics, encara que es pot aplicar a atributs numèrics sense cap problema. En aquest cas les dades es protegeixen a través de reemplaçar un conjunt d'atributs per un valor més general que els inclogui a tots. Per tant, aquest mètode no introdueix soroll o falseja les dades, simplement els fa més generals o menys específiques, de manera que les individualitats de

[1] La distribució normal, també coneguda com distribució gaussiana, és una distribució de probabilitat de variable contínua. La gràfica de la seva funció de densitat té una forma acampanada i és simètrica respecte d'un determinat punt.

cada usuari del conjunt de dades original queden difuminades entre els altres usuaris en el conjunt de dades protegit.

En el cas dels atributs numèrics, l'agregació es pot implementar mitjançant la construcció de rangs de valors. Per exemple, si tenim registres amb els valors 3,11,7,19 podem crear dos grups i associar un rang concret a cada grup. En aquest cas, podríem crear els rangs [0,10) i [10,20), i en aquest cas el primer i el tercer registre tindrien el mateix valor [0,10) en les dades protegides.

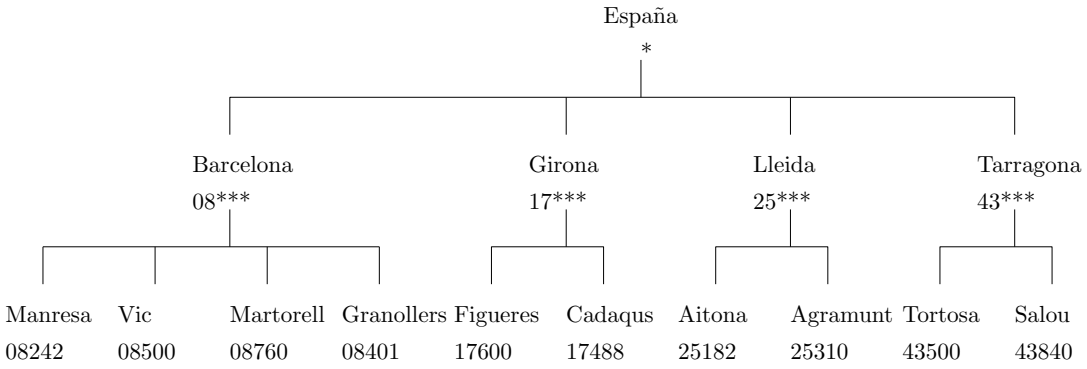

Figure 3: Jerarquia de l'atribut "codi postal"

D'altra banda, en el cas dels atributs categòrics, cal disposar de la jerarquia dels atributs, ja sigui de forma implícita o explícita. La figura 3 mostra la jerarquia de l'atribut "codi postal". Podem veure que els diferents municipis s'agrupen en províncies, i aquestes, al seu torn, s'agrupen formant l'arrel de tota l'estructura, que en aquest cas seria el nivell del país. Veiem que podríem haver inclòs un nivell extra que fes referència a les comunitats autònomes, però s'han obviat per mantenir la simplicitat de l'exemple. D'aquesta manera, si volem generalitzar la informació sense pertorbar, podem generalitzar el codi postal a nivell de província, de manera que les dades són absolutament correctes, però més generals que les dades originals.

El mètode de **supressió** d'atributs consisteix, simplement, en eliminar els atributs. Quan no és possible utilitzar la generalització o un altre mètode i l'atribut pot presentar una bretxa de privacitat, s'utilitza aquest recurs per eliminar el valor de l'atribut i indicar que el valor ha estat suprimit.

5 k-Anonimitat en taules

El model de k-anonimitat [2] no és un mètode d'emmascarament o protecció, sinó un model o condició que ha de ser satisfet pel conjunt de dades protegit. Tot i això, generalment aconseguim complir la k-anonimitat a través dels mètodes de protecció o emmascarament que hem vist en les seccions anteriors. És a dir, sobre el conjunt de dades originals se'ls aplica un o més dels mètodes d'emmascarament vistos anteriorment, amb la finalitat d'aconseguir un conjunt de dades protegit que compleixi amb les restriccions o condicions necessàries per al model de la k-anonimitat.

Un conjunt de dades compleix el model de la **k-anonimitat** si, i només si, per a qualsevol combinació d'atributs quasi-identificadors existeixen k o més registres que comparteixen els mateixos valors.

En altres paraules, cada registre en un conjunt de dades k-anònim és indistingible de, com a mínim, altres $k-1$ registres pel que fa al conjunt de quasi-identificadors. Per tant, la probabilitat d'identificació d'un usuari en un conjunt de dades k-anònim pel que fa als quasi-identificadors és de com a màxim $\frac{1}{k}$.

Per a més detalls sobre anonimitat en taules i en altres formats de dades (p.ex. grafs o xarxes socials, localitzacions, etc.) es pot consultar [1].

6 Programari

En aquesta pràctica farem servir el programari R Package[2] per a minpular diferents conjunts de dades, evaluant-ne el nivell de privacitat i efectuant les alteracions necessàries per garantir el nivell de protecció desitjat.

Tot i que és possible realitzar la pràctica amb la instal·lació bàsica de R, recomanem l'ús de l'entorn de desenvolupament integrat (IDE) R Studio[3]. En concret, caldria descarregar i instal·lar la versió RStudio Desktop (Open Source License) que és gratuït i por ser descarregat del següent enllaç:

https://www.rstudio.com/products/rstudio/download/

Addicionalment, ens caldrà fer ús del paquet `sdcMicro` que implementa els mètodes de pertorbació que farem servir. Podem trobar el manual de referència, així com altra informació d'aquest paquet al següent enllaç:

https://cran.r-project.org/web/packages/sdcMicro/index.html

Per la instal·lació d'aquest paquet, només cal iniciar R Studio (o en el seu defecte, la consola de R) i introduir la següent comanda:

```
1 install.packages("sdcMicro")
```

[2] https://www.r-project.org/
[3] https://www.rstudio.com/

7 Part 1. Mètodes de pertorbació

El primer pas consisteix, evidentment, en la càrrega del conjunt de dades. El següent fragment de codi carrega la llibreria `sdcMicro` (línia 2) i, seguidament, carrega el fitxer (línia 5). El fitxer de dades (en la ruta "data/salaris.csv") conté les dades que utilitzarem en aquesta primera part de la pràctica.

```
1  # Carrega de les llibreries
2  library(sdcMicro)
3
4  # Carregar les dades
5  dades <- read.csv(file="data/salaris.csv", header=TRUE, sep=",", colClasses=c("CP"="character
       "))
6
7  # Mostra de les dades
8  colnames(dades)
9  head(dades)
10 summary(dades)
```

La comanda `colnames` ens mostra els atributs del conjunt de dades que hem carregat:

```
1 [1] "DNI"    "SS"     "CP"     "Edat"   "Salari"
```

Mentre que la comanda `head` mostra les 5 primeres files de dades (resgistres) del conjunt de dades:

```
1         DNI           SS    CP Edat Salari
2 1 34241937Q 129886629522 17800   32     98
3 2 59945988K 197366959277 25560   25     92
4 3 24742991J 553617564644 17001   53     40
5 4 38336226I 831748773654 08018   47     21
6 5 61635114K 431747216456 43500   59     79
7 6 93993451P 994453133343 25560   31     69
```

Finalment, la comanda `summary` mostra un resum dels tipus d'atributs que conté el conjunt de dades, proporcionant informació sobre el rang de valors de cadascun d'ells. En aquest cas concret, els valors obtinguts són:

```
1        DNI           SS                  CP                Edat           Salari
2  11751941Q: 1   Min.   :1.154e+11   Length:100         Min.   :18.00   Min.   : 18.00
3  12464486B: 1   1st Qu.:3.355e+11   Class :character   1st Qu.:30.00   1st Qu.: 42.75
4  16256795X: 1   Median :5.385e+11   Mode  :character   Median :44.00   Median : 67.00
5  17176344Z: 1   Mean   :5.564e+11                      Mean   :41.32   Mean   : 62.40
6  17669148T: 1   3rd Qu.:7.824e+11                      3rd Qu.:52.25   3rd Qu.: 79.00
7  17736629N: 1   Max.   :9.950e+11                      Max.   :65.00   Max.   :100.00
8  (Other)  :94
```

A partir d'aquesta informació podem veure el contingut del conjunt de dades que utilitzarem en aquesta primera part de la pràctica. En concret, veiem que el conjunt de dades conté els següents atributs:

1. **DNI**: camp amb informació del DNI de les persones identificades en el conjunt de dades, formades per una cadena alfanumèrica de 8 dígits + 1 caràcter de control. Per exemple: 34241937Q.
2. Número de la seguretat social (**SS**): número de identificació de la seguretat social de persona, formada per una cadena numèrica de 12 dígits. Per exemple: 129886629522.

3. Codi postal (**CP**): codi postal de la població on resideix cadascun dels individus del conjunt de dades. Es representa per una cadena numèrica de 5 dígits, on els 2 primers indiquen la provincia. Per exemple: 08500.
4. **Edat**: edat de la persona, en el rang 18-65 anys. Per exemple: 30.
5. **Salari**: salari (quantitat en milers € bruts / any). Per exemple: 32 indica 32.000 € bruts/any.

El conjunt de dades conté un total de 100 individus, i.e. 100 files.

Per poder entendre una mica els valors i comportament dels atributs cal, d'alguna manera, poder-los visualitzar per fer-nos una idea dels valors que hi ha, la dispersió, etc. En el següent fragment realitzem algunes accions en aquest sentit:

```
# Visualitzar els 10 primers valors
dades[1:10, 'Edat']
# Calcul mean i sd
print(paste("Atribut 'Edat': mean value =", mean(dades[, 'Edat']), "and SD =", sd(dades[, '
    Edat']), sep=" "))
# Visualitzacio dels valors (ordenats)
plot(sort(dades[, 'Edat']), type="p", col="red", xlab="Registres", ylab="Valor", main="Edat")
```

La primera comanda simplement imprimeix per pantalla els 10 primers valors (files) de l'atribut 'Edat'. En la segona comanda, utilitzem les funcions `mean` i `sd` per a calcular el valor mig i la desviació estàndard d'aquest atribut numèric. I finalment, per poder visualitzar tots els valors d'aquest atribut, una de les opcions més senzilles i efectives és visualitzar en una gràfica els valors ordenats. La figura 4 mostra el resultat. És fàcil veure que la dispersió en el rang d'edats és lineal, i per tant, tenim una distribució equiprobable d'edats en el conjunt de dades.

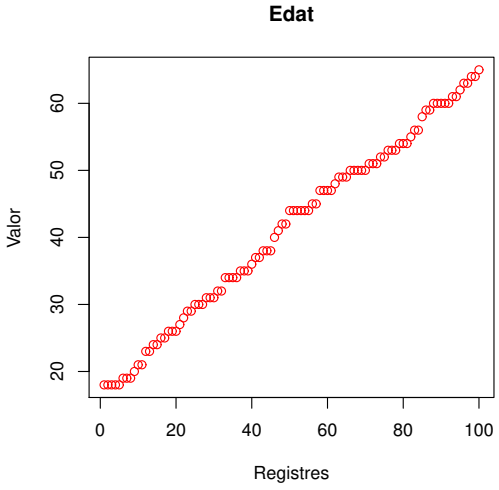

Figure 4: Visualització de l'atribut 'Edat' (ordenat)

7.1 Eliminar els identificadors

Tal i com hem vist, el primer pas per anonimitzar el conjunt de dades consisteix en eliminar els identificadors.

Exercici 1.- Eliminar els identificadors del conjunt de dades (0.5p)

Elimina els identificadors del conjunt de dades, creant una nova variable que contingui totes les files del conjunt anterior, però només els atributs que no siguin identificadors.

Activitat

Responeu a les següent qüestions:

a. Quins són els atributs identificadors en el conjunt de dades?
b. Descriu les comandes que has emprat, indicant breument la funció de cadascun dels paràmetres que has utilitzat.

7.2 Re-identificació en el conjunt de dades

Una vegada hem eliminat els identificadors, caldria definir quins són els potencials quasi-identificadors. En aquest sentit, és força útil poder veure si hi ha valors d'atributs o combinacions d'atributs que siguin únics i, per tant, puguin conduir a un atacant a re-identificar una persona dins del conjunt de dades.

Per exemple, en el nostre cas, si un atacant sap que un cert individu va participar en aquest estudi i coneix el seu lloc de residència i/o edat, seria possible que conegués el seu salari si pot identificar un registre únic dins del conjunt de dades emprant aquests atributs.

Exercici 2.- Re-identificació (0.5p)

Comprova si els atributs `CP` i/o `Edat` (és a dir, cadascun d'ells per separat o la combinació d'ambdós) poden conduir a un atacant a identificar de forma única un registre dins el conjunt de dades.

Nota: Pots fer servir la funció `table` de R.

Activitat

Responeu a les següent qüestions:

a. Quins d'aquests atributs o combinacions poden conduir a la identificació única d'un registre del conjunt de dades?
b. Indica quines comandes has fet servir per obtenir aquesta informació, explicitant el significat de les comandes i dels seus paràmetres.

7.3 Mètodes d'emmascarament

Una vegada eliminats els identificadors i identificats els quasi-identificadors, passarem a aplicar els diferents mètodes d'emmascarament. En aquesta secció de la pràctica veurem alguns dels mètodes d'emmascarament més coneguts i emprats en l'actualitat, i compararem els resultats obtinguts sobre els mateixos obtinguts, amb la finalitat de poder veure les diferències entre ells.

Exercici 3.- Aplicació de soroll additiu (*additive noise*) (0.5p)

Aplica el mètode pertorbatiu basat en soroll additiu sobre l'atribut `Edat` del conjunt de dades, explicitant una quantitat de soroll del 20%. Cal guardar el conjunt de dades resultant en una variable amb el nom `dades.an`.

Nota: Pots fer servir la comanda `addNoise` de la llibreria `sdcMicro`.

Activitat

Responeu a les següent qüestions:

a. Indica la comanda que has fet servir, junt amb els paràmetres emprats i el significat de cadascun d'ells.

b. Executa el següent codi adjunt i explica els resultats obtinguts, justificant la resposta.

Codi que cal executar per respondre la part b) d'aquesta pregunta. Cal remarcar que la variable `dades` correspon al conjunt original de dades, mentre que `dades.an` correspon al mateix conjunt, però amb els valors de l'atribut `Edat` després d'aplicar el mètode de soroll additiu:

```
plot(cbind(dades$Edat, dades.an$Edat),
     ylim=c(min(dades$Edat),max(dades$Edat)),
     xlim=c(min(dades$Edat),max(dades$Edat)),
     xlab="Original", ylab="Masked", main="Additive␣Noise␣-␣Edat␣(P=0.20)")
abline(a=0, b=1, col="red")
```

Exercici 4.- Aplicació del mètode d'intercanvi de rang additiu (*rang swap*) (0.5p)

Aplica el mètode pertorbatiu basat en l'intercanvi de rang sobre l'atribut `Edat` del conjunt de dades original, explicitant un percentatge d'intercanvi del 10%. Cal guardar el conjunt de dades resultant en una variable amb el nom `dades.rs`.

Nota: Pots fer servir la comanda `rankSwap` de la llibreria `sdcMicro`.

Activitat

Responeu a les següent qüestions:

a. Indica la comanda que has fet servir, junt amb els paràmetres emprats i el significat de cadascun d'ells.

b. Genera una gràfica similar a la que hem emprat en l'exercici anterior – apartat b), on es pugui veure de forma visual i ràpida la dispersió o alteració de l'atribut `Edat` després d'aplicar el mètode d'intercanvi de rang.

Exercici 5.- Càlcul de la utilitat (0.5p)

En aquest exercici treballarem amb el concepte de utilitat de les dades, o de forma contraria, la pèrdua de informació que es produeix durant el procés d'anonimització.

Activitat

Responeu a les següent qüestions:

a. Utilitza la comanda `dUtility` de la llibreria `sdcMicro` per a estimar la pèrdua de informació que s'ha produït en generar la versió pertorbada de l'atribut `Edat` emprant els mètodes de soroll additiu i intercanvi de rang (generats en els dos exercicis anteriors). Indica la comanda que has fet servir, junt amb els paràmetres emprats i el significat de cadascun d'ells.

b. Comenta els resultats obtinguts i descriu el significat dels valors obtingut i com es calculen.

Exercici 6.- Càlcul del nivell de privacitat (0.5p)

En aquest exercici treballarem amb el concepte de risc o nivell de privacitat de les dades.

Activitat

Responeu a les següent qüestions:

a. Utilitza la comanda `dRisk` de la llibreria `sdcMicro` per a estimar el risc o nivell de privacitat en la versió pertorbada de l'atribut `Edat`, emprant els mètodes de soroll additiu i intercanvi de rang (generats en els dos exercicis anteriors). Indica la comanda que has fet servir, junt amb els paràmetres emprats i el significat de cadascun d'ells.

b. Comenta els resultats obtinguts i descriu el significat dels valors obtingut i com es calculen.

Exercici 7.- Micro-agregació univariant i multivariant (1p)

En aquest exercici veurem com aplicar el mètode pertorbatiu basat en micro-agregació, tant en la seva modalitat univariant com multivariant.

Nota: Pots fer servir la comanda `microaggregation` de la llibreria `sdcMicro`.

Activitat

Responeu a les següent qüestions:

a. Aplica el mètode de micro-agregació univariant amb un nivell d'agregació igual a 3 als atributs `Edat` i `Salari`, de forma independent, sobre conjunt de dades original. Indica les comandes i paràmetres que has fet servir i la comenta la seva funció.

b. Els histogrames són gràfiques que presenten la freqüència d'aparició segons els valors. Per tant, són molt interessants per veure quins són els valors més freqüents i quins són únics.

   Crea quatre gràfiques que presentin la informació de l'atribut `Edat` i `Salari` abans i després del procés de micro-agregació, i que permetin comparar de forma visual i senzilla, com s'han modificat el valors d'aquests dos atributs.

   Indica el codi que has emprat, detallant la funció dels diferents

   Nota: la funció `par(mfrow=c(2,2))` us pot ajudar a posar les 4 gràfiques en una sola imatge, de manera que simplifica la visualització.

c. Aplica el mètode de micro-agregació multivariant MDAV amb un nivell d'agregació igual a 3 als atributs `Edat` i `Salari` sobre conjunt de dades original. Indica les comandes i paràmetres que has fet servir i la comenta la seva funció.

d. Crea les gràfiques que presentin la informació dels atributs `Edat` i `Salari` anonimitzats en la seva versió multivariant, i compara els resultats obtinguts amb els obtinguts en el cas univariant.

e. Realitza una estimació del risc i de la utilitat en els dos casos (univariant i multivariant). Comenta i justifica els resultats obtinguts.

Exercici 8.- Generalització d'atributs (1p)

Un dels mètodes d'emmascarament no pertorbatiu més coneguts i empleats és la generalització d'atributs, que pot ser aplicat tant a atributs numèrics com a atributs categòrics estructurats.

Tal i com hem explicat anteriorment, i exemplificat en la figura 3, el codi postal és un exemple d'atribut categòric estructurat. Els diferents municipis s'agrupen en províncies, i aquestes, al seu torn, s'agrupen en nivells superiors (ja sigui de comunitat autònoma o de país) fins arribar a l'arrel de tota l'estructura, que en aquest cas seria el nivell del país.

Activitat

Responeu a les següent qüestions:

a. Crea una funció pròpia que permeti generalitzar els codis postals de població, convertint-los en codis de província. Detalle el codi de la funció, així com alguns exemples del seu funcionament i valors de retorn.

b. Aplica la funció que has creat a l'atribut `CP` del conjunt de dades, per a tots els individus (registres) existents. Comenta el resultat obtingut.

c. Aplica la funció que has creat a l'atribut `CP` del conjunt de dades, però només a aquells registres que siguin únics, és a dir, que la seva freqüència d'aparició sigui igual a 1 en tot el conjunt de dades. Comenta el resultat obtingut i compara-ho amb els resultats de l'apartat anterior.

Exercici 9.- Alternatives per mesurar la pèrdua de informació (1p)

Com mesurar la pèrdua de informació que es produeix durant el procés d'anonimització és un dels problemes que hi ha quan es vol comparar diferents mètodes i paràmetres per trobar un model òptim a aplicar.

Executeu el codi que es mostra a continuació, tenint en compte que cal verificar els noms de les variables per a què funcioni correctament:

```
names <- c("AN", "RS", "MA_one", "MA_mul")
colors <- c("grey", "red", "blue", "green")
lty <- 1:4
lwd <- 3

dif_an <- abs(dades[, 'Edat']-dades.an[, 'Edat'])
dif_rs <- abs(dades[, 'Edat']-dades.rs[, 'Edat'])
dif_ma_one <- abs(dades[, 'Edat']-dades.ma_one[, 'Edat'])
dif_ma_mul <- abs(dades[, 'Edat']-dades.ma_mul[, 'Edat'])

ymin <- min(dif_an, dif_rs, dif_ma_one, dif_ma_mul)
ymax <- max(dif_an, dif_rs, dif_ma_one, dif_ma_mul)

par(mfrow=c(1,1))
plot(sort(dif_an), type="l", col=colors[1], lty=lty[1], lwd=lwd, ylim=c(ymin,ymax), xlab="
    Registers", ylab="Error", main="Edat")
lines(sort(dif_rs), col=colors[2], lty=lty[2], lwd=lwd)
lines(sort(dif_ma_one), col=colors[3], lty=lty[3], lwd=lwd)
lines(sort(dif_ma_mul), col=colors[4], lty=lty[4], lwd=lwd)
legend(x="topleft", legend=names, col=colors, lty=lty, lwd=lwd)
```

On:

- `dades.an` és el conjunt de dades després d'aplicar el mètode de soroll additiu.
- `dades.rs` és el conjunt de dades després d'aplicar el mètode de intercanvi de rang.
- `dades.ma_one` és el conjunt de dades després d'aplicar el mètode de micro-agregació univariant.
- `dades.ma_mul` és el conjunt de dades després d'aplicar el mètode de micro-agregació multivariant.

Activitat

Responeu a les següent qüestions:

a. Mostra la gràfica obtinguda.

b. Comenta i justifica els resultats obtinguts.

8 Part 2. Anonimització d'un conjunt de dades

En aquesta segona part es demana aplicar els coneixements que has adquirit en els exercicis anteriors d'aquesta pràctica.

En concret, es demana que realitzis el procés d'anonimització complet d'un nou conjunt de dades, i que presenta i justifiquis els resultats obtinguts, argumentant sempre el perquè de les decisions que has anat prenent durant el procés.

El conjunt de dades que hauràs d'anonimitzar en aquest exercici es troba a:

`"data/hipoteca.csv"`

Aquest conjunt presenta dades de clients d'un banc que han sol·licitat una hipoteca. Es mostren dades bàsiques dels clients i la probabilitat (valors en el rang [0,1]) de que se li concedeixi la hipoteca. A continuació veiem els detalls dels atributs:

1. **DNI**: número del document de identitat nacional.
2. **Sexe**: Masculí (M) o femení (F)
3. **Tel**: Telèfon de contacte
4. **CP**: Codi postal de la residència habitual
5. **Edat**: Edat de la persona
6. **NumFills**: Número de fills
7. **NivellEstudis**: Nivell d'estudis (0: primària, 1: secundària, 2: batxillerat, 3: grau/llicenciatura/diplomatura, 4: doctorat)
8. **Ocupacio**: Ocupació actual
9. **Salari**: Salari actual (milers de € bruts / any)
10. **Prob**: Probabilitat que se li concedeixi una hipoteca

En aquest cas, l'atribut objectiu del conjunt de dades és la probabilitat de què es concedeixi una hipoteca a la persona indicada (atribut `Prob`).

Exercici 10.- Anonimització del conjunt de dades "hipoteca" (4p)

A partir del conjunt de dades que permet predir la probabilitat de concedir una hipoteca, realitzeu un procés d'anonimització complet, considerant tots els punts vistos anteriorment i d'altres que permetin assegurar la preservació de la privacitat dels usuaris involucrats en aquest conjunt de dades.

Activitat

Responeu a les següent qüestions:

a. Expliqueu, atribut per atribut, quin tipus d'atribut és (numèric, categòric nominal, etc.) i a quina classe correspon segons la classificació de privacitat (identificador, quasi-identificador, etc.).

b. Expliqueu, atribut per atribut, quin mètode d'emmascarament heu utilitzar en cas de què calgués alterar l'atribut. Justifiqueu les decisions presses en cada cas.

c. Avalueu el nivell de privacitat que heu assolit en el conjunt final, una vegada anonimitzats tots els atributs necessaris.

c. Avalueu el nivell de utilitat de les dades que heu assolit en el conjunt final, una vegada anonimitzats tots els atributs necessaris. Us podeu ajudar de gràfiques i altres recursos que considereu rellevants.

Nota: Tal i com s'indica a l'inici d'aquest enunciat, cal entregar tot el codi (en un fitxer .R) que heu generat per aquesta part, partint de la càrrega del fitxer (ubicat en "data/hipoteca.csv") i fins la generació dels resultats i tots els recursos necessaris (gràfiques, etc.).

References

[1] Casas-Roma, J.; Romero-Tris, C. (2017). “Privacidad y anonimización de datos”. Barcelona, Spain: Editorial UOC. 150 pp. ISBN: 9788491169383.

[2] Sweeney, L. (2002). “k-anonymity: a model for protecting privacy”. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (IJUFKS), Vol. 10(5), pp. 557-570.

[3] Samarati, P. (2001). “Protecting Respondents’ Identities in Microdata Release”. IEEE Transactions on Knowledge and Data Engineering (TKDE), Vol. 13(6), pp. 1010-1027.

9 Annex 1: Principals comandes de R

```
Funcions rellevants
===================

    Commanda        Description
    --------        -----------
    ?<command>      Ajuda sobre una comanda
    abline          Afegir una linea recta en un plot
    cbind           Concatenar columnes
    hist            Generar un histograma
    legend          Afegir la llegenda a una gr fica (plot)
    lines           Afegir seq  ncies de valors a una gr fica (plot)
    max             Retorna el valor m xim dels par metres
    min             Retorna el valor m nim dels par metres
    plot            Genera un gr fic a partir d'una s rie de valors
    print           Imprimeix per pantalla
    read.csv        Llegeix un fitxer separat per comes (CSV) i el retorna com dataframe
    sapply          Aplica una funci  sobre cadascun dels valors d'un vector o array
```