POLITECNICO DI TORINO

Corso di Laurea Magistrale in Ingegneria Matematica

Tesi di Laurea

GDPR: Privacy e Utilità del dato. Applicazioni sulla differential privacy

Relatore:
Silvia Chiusano

Candidato:
Leonardo Turturiello

Anno Accademico 2017-2018

Ai miei genitori

Indice

1 Normativa Gdpr **7**

1.1 Definizioni 7

1.2 Principi 9

2 Soluzione in azienda **11**

2.1 RGI S.P.A 11

2.2 Struttura Database 12

2.3 Soluzione 15

2.3.1 Estrazione e Cancellazione 15

2.3.2 Individuazione dati sensibili 24

2.3.3 Offuscamento 25

3 Pianificazione della Sicurezza **27**

3.1 Sicurezza Database 27

3.2 Anonimizzazione 28

3.3 K-Anonimity 31

3.4 Linkage attack 33

3.5 Statistical Database 35

3.6 Questionari Polarizzati 36

3.6.1 Stima per questionari randomizzati 37

3.6.2 Precisione dei questionari randomizzati 39

4 Differential Privacy **41**

4.1 Nozioni Preliminari 41

4.2 Definizione di Differential Privacy 44
4.3 Sensitività 45
4.4 I Principali Meccanismi della Differential Privacy 48
4.4.1 Il Meccanismo di Laplace 48
4.4.2 Il Meccanismo Gaussiano 49
4.4.3 Il Meccanismo Esponenziale 49
4.5 Misura di utilità della differential privacy 52
4.6 Decision Tree: Privacy e Utilità 59
4.6.1 Overview Alberi Decisionali 59
4.6.2 Criterio di Split 61
4.6.3 Valutazione della classificazione 67
4.7 Applicazione della Differential Privacy in un problema di classificazione con Decision Tree 70
4.8 Conclusioni 76

5 Conclusioni e sviluppi futuri **78**

Bibliografia e Sitografia **79**

Introduzione

Il 28 maggio 2018 rappresenta una data fondamentale per la tutela della privacy dei cittadini europei: tutti gli Stati membri dell' Unione Europea si attengono al nuovo regolamento GDPR (General Data Protection Regulation) sulla privacy e sulla protezione e circolazione dei dati personali.
Si tratta quindi di disciplinare il trattamento, totalmente o parzialmente, automatizzato di tutte le informazioni che consentono di risalire ad una persona fisica direttamente (nome, cognome, e codice fiscale) o indirettamente (residenza). Ogni cittadino europeo può godere di nuovi diritti, come, ad esempio, il diritto all'oblio e alla portabilità, che verranno descritti successivamente. Tale regolamento ha un impatto significativo su tutte le organizzazioni che, in modo diretto o indiretto, trattano dati personali. Ogni azienda, infatti, processa, ogni giorno, un'enorme quantità di informazioni, e non sempre vi è una giusta attenzione al rispetto e alla tutela dei dati sensibili.
Nell'epoca dei Big Data, la disponibilità ad accedere ad una tale quantità di informazioni è divenuta oggetto di riflessione: " Oggi l'interesse principale della maggior parte dei colossi del mondo della tecnologia è l'accumulo di qualsiasi genere di dato per poterlo utilizzare in futuro, anche per scopi oggi a noi sconosciuti " - afferma il Garante della Privacy.
Per descrivere il fenomeno dei Big Data spesso ci si riconduce al paradigma delle tre V: Volume, Velocità, Varietà.
La prima V (Volume) fa riferimento alla mole di dati quotidianamente prodotta: stimata a 2,5 miliardi di byte in forma di file, immagini, video, ecc. Oltre alla quantità dei dati, ciò che è cresciuta esponenzialmente negli ultimi

anni è la disponibilità e la facilità di accesso agli stessi.
Con la seconda V (Velocità) si intende il modo con cui vengono prodotti i dati: dinamico e non più statico. Si parla ormai di flusso di informazioni che deve essere elaborato ad una certa velocità.
La terza V (Varietà) si riferisce alla grande varietà di fonti (pagine Facebook, Twitter, ecc.) da cui provengono i dati[1].
All'interno del contesto dei Big Data, alla privacy va attribuito un ruolo da protagonista, al fine di tutelare non solo la protezione ma anche la libera circolazione dei dati personali. La richiesta dei dati personali è in costante crescita. Infatti, privacy e Big Data non devono più viaggiare su due binari paralleli, ma essere l'uno il presupposto dell'altro.
Il primo capitolo della tesi è un'introduzione alla normativa GDPR e ai punti principali che riguardano sia i nuovi diritti dei cittadini europei sia le misure che le aziende devono adottare nel rispetto della normativa. A tal proposito è posta particolare attenzione al punto della normativa *Privacy by Design*, secondo cui qualunque base dati, progetto, sistema va realizzato considerando, fin dalla progettazione, la protezione dei dati sensibili.
La seconda parte descrive la soluzione implementata all'interno di RGI, azienda che mi ha seguito nel percorso di stesura dell'elaborato e durante il periodo di tirocinio. Sono stato inserito nel team di analisi, nel quale è stata ideata la soluzione funzionale al raggiungimento dei requisiti della normativa.
L'ultima parte è un approfondimento sulle tecniche utilizzate per garantire la sicurezza dei dati personali per far fronte a possibili tentativi di violazione della privacy. Al tempo stesso verrà posta particolare attenzione al rapporto tra livello di privacy raggiunto e utilità, ovvero all'integrità dell'informazione protetta. In questo contesto si colloca la teoria della *Differential Privacy* formulata da Dwork nel 2006, che permette di aggiungere un livello di incertezza all'applicazione di meccanismi di estrazione dei dati.

[1] Cfr. Giuseppe D'Acquisto, Maurizio Naldi *Big Data and Privacy by Design*, I diritti nella "rete" della rete, Anonimizzazione Pseudonimizzazione Sicurezza, Prefazione di Franco Pizzetti, G.Giappichelli Editore (2017).

Capitolo 1

Normativa Gdpr

Il regolamento generale sulla protezione dei dati GDPR (*General Data Protection Regulation*) mira a rafforzare e a rendere omogeneo il trattamento dei dati personali dei cittadini e dei residenti dell'Unione Europea, sia all'interno che all'esterno dei confini della stessa.

1.1 Definizioni

Dato personale qualsiasi informazione riguardante una persona fisica identificata o identificabile («interessato»); si considera identificabile la persona fisica che può essere identificata, direttamente o indirettamente, con particolare riferimento a un identificativo come il nome, un numero di identificazione, dei dati relativi all'ubicazione, un identificativo online o a uno o più elementi caratteristici della sua identità fisica, fisiologica, genetica, psichica, economica, culturale o sociale;

Trattamento: qualsiasi operazione o insieme di operazioni, compiute con o senza l'ausilio di processi automatizzati e applicate a dati personali o a insiemi di dati personali, come la raccolta, la registrazione, l'organizzazione, la strutturazione, la conservazione, l'adattamento o la modifica, l'estrazione, la consultazione, l'uso, la comunicazione mediante trasmissione, diffusione o

qualsiasi altra forma di messa a disposizione, il raffronto o l'interconnessione, la limitazione, la cancellazione o la distruzione;

Profilazione: qualsiasi forma di trattamento automatizzato di dati personali consistente nell'utilizzo degli stessi per valutare specifici aspetti relativi a una persona fisica, in particolare per analizzare o prevedere aspetti riguardanti il rendimento professionale, la situazione economica, la salute, le preferenze personali, gli interessi, l'affidabilità, il comportamento, l'ubicazione o gli spostamenti di detta persona fisica;

Pseudonimizzazione: il trattamento dei dati personali in modo tale che gli stessi non possano più essere attribuiti a un interessato specifico senza l'utilizzo di informazioni aggiuntive, a condizione che tali informazioni siano conservate separatamente e soggette a misure tecniche e organizzative intese a garantire che i dati personali non siano attribuiti a una persona fisica identificata o identificabile;

Titolare del trattamento: la persona fisica o giuridica, l'autorità pubblica, il servizio o altro organismo che, singolarmente o insieme ad altri, determina le finalità e i mezzi del trattamento di dati personali; quando le finalità e i mezzi di tale trattamento sono determinati dal diritto dell'Unione o degli Stati membri, il titolare del trattamento o i criteri specifici applicabili alla sua designazione possono essere stabiliti, rispettivamente, dai diritti sopra citati;

Responsabile del trattamento: la persona fisica o giuridica, l'autorità pubblica, il servizio o altro organismo che tratta dati personali per conto del titolare del trattamento;

Violazione dei dati personali: la violazione di sicurezza che comporta accidentalmente o in modo illecito la distruzione, la perdita, la modifica, la

divulgazione non autorizzata o l'accesso ai dati personali trasmessi, conservati o comunque trattati[2];

1.2 Principi

Consenso - Art. 7 un valido consenso deve essere esplicitamente dato per la raccolta dei dati e per i propositi per i quali sono usati. Per trattare i dati sensibili, il Regolamento prevede che il consenso debba essere anche "esplicito". Il consenso potrà essere revocato in ogni momento. Consenso ai genitori o a chi esercita la potestà genitoriale per trattare i dati personali dei minori di 16 anni.

Sicurezza dei dati - Art. 25: la sicurezza dei dati raccolti è garantita dal titolare del trattamento e dal responsabile del trattamento chiamati a mettere in atto misure tecniche e organizzative idonee per garantire un livello di sicurezza adeguato al rischio, come la crittografia e l'offuscamento.

Portabilità - Art. 20: una persona deve essere in grado di trasferire i propri dati personali da un sistema elettronico a un altro, senza che il controllore dei dati possa impedirlo.

Diritto alla cancellazione, limitazione e rettifica (all'oblio-Art. 17): gli interessati potranno ottenere la cancellazione dei propri dati personali anche *on line* da parte del titolare del trattamento qualora ricorrano alcune condizioni previste dal Regolamento: se i dati sono trattati solo sulla base del consenso ; se i dati non sono più necessari per gli scopi rispetto ai quali

[2]Cfr. Gazzetta ufficiale dell'Unione europea *REGOLAMENTO (UE) 2016/679 DEL PARLAMENTO EUROPEO E DEL CONSIGLIO del 27 aprile 2016*, L 119/1

sono stati raccolti; se i dati sono trattati illecitamente; oppure se l'interessato si oppone legittimamente al loro trattamento.

Violazione dei dati - Art. 33: il titolare del trattamento dovrà comunicare eventuali violazioni dei dati personali (*data breach*) comunicarle entro 72 ore da quando ne è venuto a conoscenza.[3]

[3] Cfr. Wikipedia, https://it.wikipedia.org/wiki/Regolamento_generale_sulla_protezione_dei_dati

Capitolo 2

Soluzione in azienda

2.1 RGI S.P.A

RGI è una Software Vendor indipendente, leader in Europa nella trasformazione digitale delle compagnie assicurative.[4] RGI offre alle compagnie assicurative una varietà di software che comprende la gestione delle polizze, a partire dal processo di emissione fino alle operazioni di post vendita, le denunce dei sinistri, le analisi e la gestione del mercato. La soluzione principale per le compagnie è Pass-Insurance, che permette di gestire molteplici funzionalità in un'unica soluzione. Con un fatturato di circa sessanta milioni e con più di ottocento dipendenti, RGI lavora per sviluppare ed implementare sistemi e pacchetti per le Compagnie, Bancassurance, Agenti, Broker e Promotori finanziari.
L'azienda si sviluppa su tredici sedi in sei paesi differenti: Francia, Irlanda, Germania, Lussemburgo, Tunisia e Italia. In particolare, le sedi italiane sono presenti ad Ivrea (sede amministrativa), Torino, Milano, Roma, Verona e Bologna.
Il prodotto *core* è PASS-Insurance, che è diventato negli anni lo standard di mercato e si è affermato come Sistema direzionale e agenziale nei più importanti Gruppi Assicurativi italiani.

[4] http://www.rgigroup.it/overview

PASS-Insurance è un sistema informativo completo, che racchiude l'intero ciclo del Portafoglio tecnico. Attraverso PASS-Insurance la Compagnia realizza la centralizzazione dei vari Sottosistemi: Prodotti, Anagrafe, Portafoglio, Contabilità, Analitycs, Sinistri, etc.
Il sistema di gestione del portafoglio è costituito principalmente dai seguenti moduli:

- *Pass-P&C*: Emissione e gestione delle polizze assicurative del ramo Danni.
- *Pass-Life*: Emissione e gestione delle polizze assicurative del ramo Danni.
- *Pass-Claims*: Registrazione dei sinistri
- *Pass-Product*: Modulo di *back end* dedicato alla progettazione e alla configurazione di tutti i prodotti richiesti da ogni cliente. Rappresenta il motore di calcolo tariffario dei premi di polizza.

2.2 Struttura Database

La struttura dei database all'interno di RGI è di tipo relazionale. Per ogni compagnia assicurativa cliente vengono create circa duemila tabelle contenenti dati relativi ad anagrache, contatti, polizze , sinistri, ecc
Nella Figura 2.1 sono riportate alcune delle tabelle presenti nei database, legate attraverso le opportune relazioni.

In particolare sono riportati i legami dell'entità Beneficiario (tabella PVVBENEFICIARIO)[5] con le tabelle PPVVPolizzaBenefic, PCPOLIZZA, PCRuolo, costituita da più attributi di tipo *date*, *varchar*, *varchar*2, *timestamp*, *boolean*. Ad esempio la struttura della tabella PVVBENEFICIARIO è la mostrata nella figura 2.2:

[5]Il prefisso Pv fa riferimento alle tabelle dell'ambiente Vita

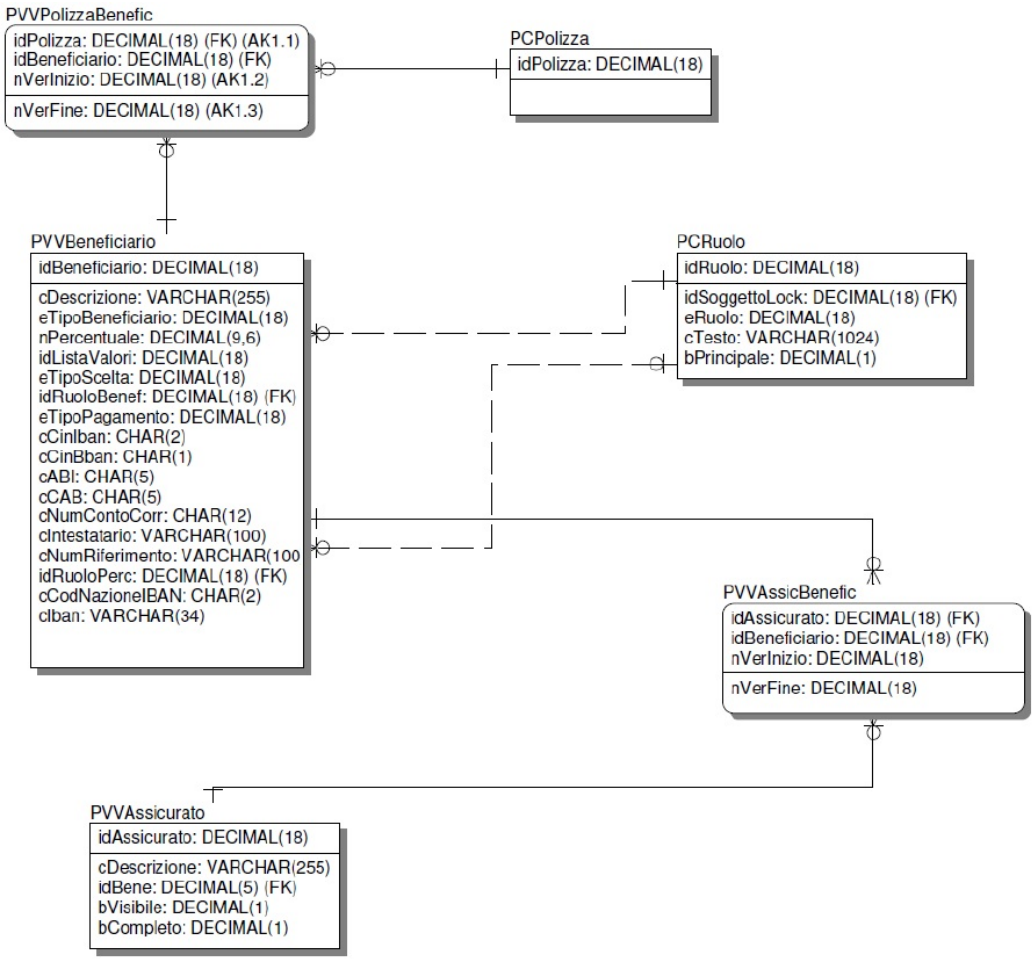


Figura 2.1: Struttura Database

COLUMN_NAME	TYPE_NAME
IDBENEFICIARIO	NUMBER
CDESCRIZIONE	VARCHAR2
ETIPOBENEFICIARIO	NUMBER
NPERCENTUALE	NUMBER
IDLISTAVALORI	NUMBER
ETIPOSCELTA	NUMBER
IDRUOLOBENEF	NUMBER
ETIPOPAGAMENTO	NUMBER
CCINBBAN	VARCHAR2
CCINIBAN	VARCHAR2
CABI	VARCHAR2
CCAB	VARCHAR2
CINTESTATARIO	VARCHAR2
CNUMRIFERIMENTO	VARCHAR2
CNUMCONTOCORR	VARCHAR2
IDRUOLOPERC	NUMBER
CCODNAZIONEIBAN	VARCHAR2
CIBAN	VARCHAR2
ERELAZIONECONTR	NUMBER
IDQUESTIONARIOBEN	NUMBER
IDPAYMENT	NUMBER

Figura 2.2: PVVBENEFICIARIO

Come si può notare, il database contiene un'enorme quantità di informazioni, molte delle quali riguardanti dati sensibili, di pagamento, di domicilio di un determinato soggetto. La normativa prevede che tali informazioni debbano essere protette al fine di tutelare la privacy o addirittura eliminiate fisicamente dal database.

2.3 Soluzione

Per adempiere alla Normativa, RGI ha dovuto effettuare degli sviluppi software al fine di garantire la cancellazione/offuscamento e crittografia dei dati "sensibili" salvati sul database ed adeguare la relativa visualizzazione di tali dati sul front-end di Pass. Nel periodo di stage e tesi in azienda ho seguito lo studio della Normativa GDPR insieme al team di lavoro di RGI. Inizialmente abbiamo analizzato la normativa e individuato i requisiti minimi al raggiungimento della compliance. Successivamente tali requisiti e la relativa soluzione applicativa sono stati condivisi con alcune compagnie assicurative clienti di RGI. Solo dopo tale condivisione, il team di lavoro di RGI ha iniziato a sviluppare la soluzione, i cui principali punti di impatto sono descritti nei paragrafi seguenti:

2.3.1 Estrazione e Cancellazione

Una volta trascorsi i termini di legge che prescrivono l'obbligo di conservazione dei dati, non deve essere più possibile risalire ad una persona fisica a partire dalle informazioni presenti. Nel caso di un sistema gestionale, questo implica l'eliminazione di polizze, contratti, titoli, sinistri e soggetti anagrafici, al termine del periodo durante il quale devono essere conservati obbligatoriamente in base ai termini di legge.
Le esigenze di business di una compagnia assicurativa prevede quindi la cancellazione di dati personali censiti e di non permettere più l'identificazione dei soggetti stessi.

Figura 2.3:

Descrizione Generale del Processo

Il processo prevede l'interazione di due figure principali:

- **Il titolare del trattamento** del dato, cioè colui che determina le finalità del trattamento dei dati personali

- **Il responsabile del trattamento**, cioè clui che tratta i dati personali in nome e per conto del titolare

I dati personali raccolti e trattati sono quelli del **soggetto interessato**, cioè della persona fisica identificata o identificabile.
Per poter cancellare definitivamente un soggetto dalla base dati in cui sono censite le anagrafiche, occorre effettuare degli step preliminari:
Durante la prima fase il responsabile del trattamento elabora i sinistri, successivamente viene elaborato il portafoglio delle polizze vita e danni ed infine l'anagrafe. Il processo si compone di due macro step: Il task di ricerca e il task di eliminazione, che si ripetono per i tre sistemi sopra indicati.

Il primo step è quello di estrazione dei sinistri cancellabili, in quanto tale sistema è indipendente e slegato da altre basidati.

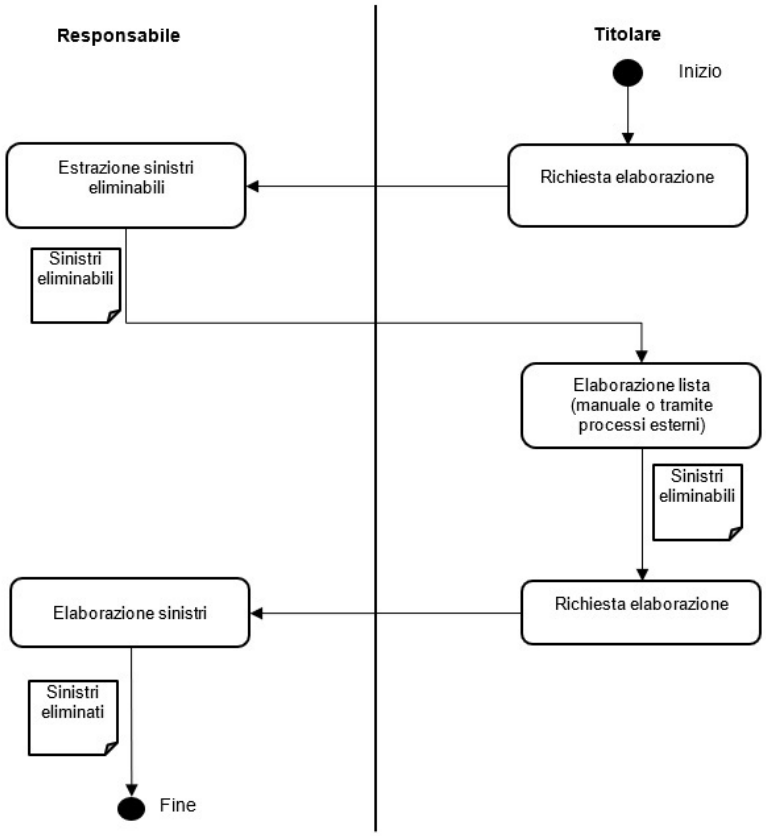


Figura 2.4: Fase 1 - Elaborazione Sinistri

- Task di ricerca: esegue un'estrazione dei sinistri per i quali sono scaduti i termini di conservazione, e quindi elaborati in modo che non sia possibile, a partire dal sinistro, risalire ad un soggetto fisico specifico. il task produce in output un tracciato csv contenente l'elenco dei sinistri che possono essere elaborati. Tale file può essere elaborato dal titolare, sia manualmente che attraverso procedure automatiche. Tale procedura permette all'utente di avere uno step intermedio prima di lanciare la funzione di cancellazione, e quindi di intervenire puntualmente su ogni singolo sinistro.

- Task di eliminazione : esegue l'elaborazione vera e propria. L'input per questo task è costituito dal tracciato di output del task di ricerca(

indicato nello schema come "Sinistri eliminabili"). Al termine dell'elaborazione il task produce un tracciato del tutto analogo a quello di input, aggiornato con l'indicazione dei sinistri elaborati o meno (indicato nello schema come "Sinistri eliminati").

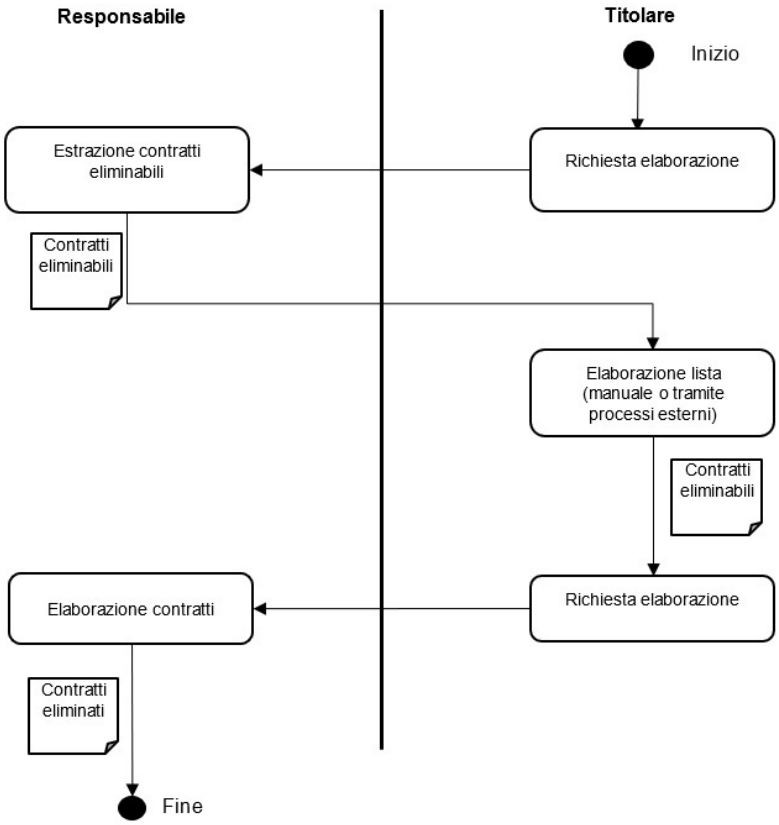


Figura 2.5: Fase 2 - Elaborazione Contratti

- Task di ricerca : esegue un'estrazione dei contratti per i quali sono scaduti i termini di legge, e quindi elaborati in modo che non sia possibile risalire ad un soggetto fisico a partire dal contatto. L'estrazione si basa suula data scadenza dei contratti di assicurazione.
  In seguito alla prima estrazione, il task esegue dei controlli sui sottosistema dei Sinistri per verificare che i contratti estratti non siano legati a sinistri i cui dati devono essere conservati.
  Terminata la verifica, il task produce un tracciato csv (indicato nello schema come "Contratti eliminabili") contenente l'elenco dei contratti eliminabili e quindi processabili. Tale file può essere rielaborato dal titolare manualmente.

- Task di eliminazione : esegue l'eliminazione dei contratti e degli oggetti ad essi legati, dopo aver effettuato controlli di fattibilità. L'input di questo task è costituito dal tracciato di output del task di ricerca. Al termine dell'elaborazione il task produce un tracciato analogo a quello di input, aggiornato con l'indicazione dei contratti elaborati o meno (indicato nello schema come "Contratti eliminati").

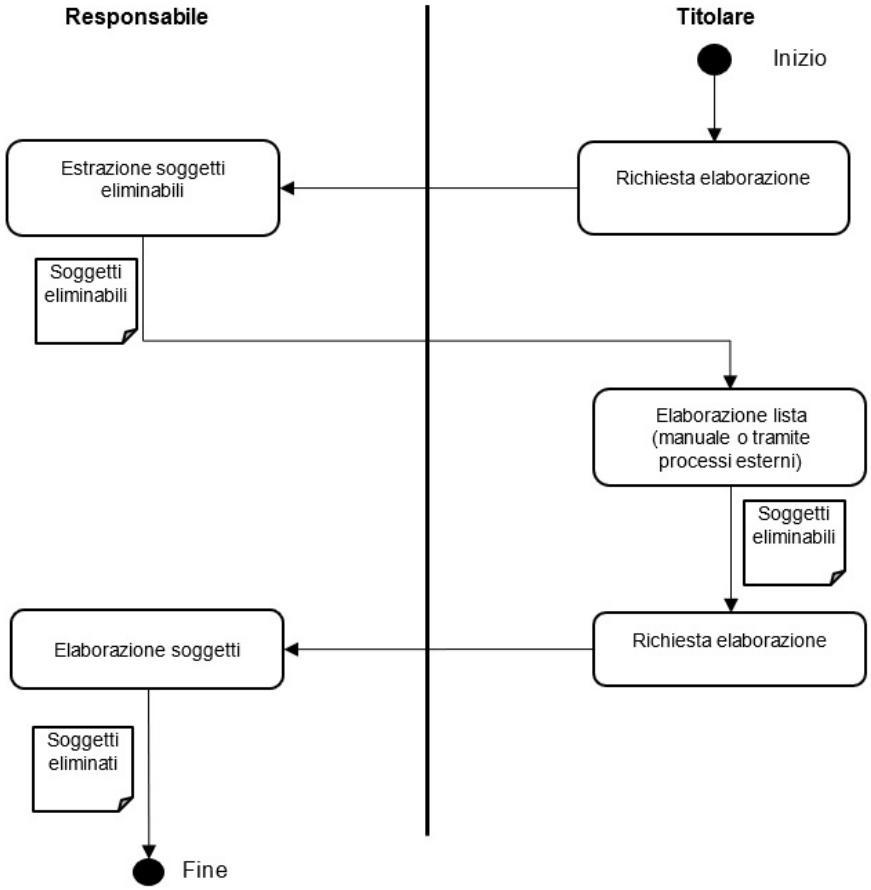


Figura 2.6: Fase 3 - Elaborazione Soggetti

- Task di ricerca : esegue un'estrazione dei soggetti anagrafici per i quali sono scaduti i termini di conservazione per legge e i cui dati devono essere eliminati. L'estrazione si basa sulla data in cui è stato fornito un consenso commerciale al trattamento dei dati personali.
  In seguito, il task esegue dei controlli sui sottosistemi Sinistri e Portafoglio per verificare che i clienti estratti non siano coinvolti in alcun sinistro e non siano presenti su contratti i cui dati devono essere conservati.
  Terminata la fase di verifica, il task produce un tracciato csv contenente l'elenco dei soggetti estratti (indicato nello schema come "Soggetti

eliminabili"), sia nel caso abbiano superato i controlli, sia nel caso abbiano invece dei legami con uno o più sottosistemi.Nel primo caso i soggetti vengono contrassegnati come da eliminare, nel secondo caso vengono resettati gli eventuali dati commerciali da essi forniti.

- Task di eliminazione esegue l'elaborazione dei soggetti e degli oggetti ad essi legati(eliminazione o reset dati commerciali) dopo aver effettuato i controlli di fattibilità. L'input per questo task è costituito dal tracciato di output del task di ricerca. Al termine dell'elaborazione il task riporta il risultato analogo al tracciato di input (indicato nello schema come "Soggetti elaborati"), aggiornato con l'indicazione dei soggetti elaborati o meno.

Tracciati di output

Al termine di ogni processo di estrazione viene quindi prodotto un tracciato di output che sarà l'input del corrispondente task di cancellazione. Si tratta di un file csv con il carattere "," come separatore d campo. Sono previsti un record d'intestazione e un record per ciascun sinistro/contratto o soggetto a seconda che si faccia riferimento alla fase 1, alla fase 2 o alla fase 3.
La struttura standard di un tracciato di output è la seguente:

Nome Campo	Tipo	Descrizione
Compagnia	S	Codice ANIA della compagnia
Nodo	S	Codice identificativo dell'agenzia
Numero polizza	S	Numero univoco della polizza
Contraente	S	Nome e cognome del contraente
Data decorrenza	S	Data decorrenza del contratto nel formato AAMMGG
Data scadenza	S	Data scadenzaa del contratto nel formato AAMMGG
Eliminabile	B	Valori possibili: 1 = Eliminabile 0 = Non eliminabile

2.3.2 Individuazione dati sensibili

Come illustrato precedentemente, all'interno delle tabelle del dataset sono contenute tutte le informazioni relative ad una polizza assicurativa. Tuttavia non tutti gli attributi sono da considerare sensibili e soltanto alcuni di essi devono essere offuscati. Un concetto importante è quello di anonimizzazione:
anonimizzare un dato personale significa trattare un dato in modo da introdurre un'incertezza nell'attribuzione di quel dato ad un soggetto. In particolare , se si eliminano le informazioni che rendono identificabile un determinato soggetto, allora tale informazione risulta anonimizzata. Utilizzando questa tecnica, vengono mantenuto all'interno del database tutte le informazioni che possono essere utilizzate a fini statistici. E' stata effettuata una prima analisi sui tipi di attributi all'interno di una tabella, al fine di considerare solo i tipi che possono contenere dati sensibili, Per questo motivo sono stati presi in considerazione soltanto i tipi carattere e numerico. Inoltre sono stati esclusi dalla ricerca tutti i campi contenenti codici ID dei soggetti, in quanto non contenenti informazioni che possano ricondurre all'dentificazione. Tramite la

seguente query in linguaggio SQL sono stati estratti i record significatici per l'individuazione degli attributi sensibili:

```
SELECT TABLE_NAME, COLUMN_NAME, DATA_TYPE, DATA_LENGTH
FROM ALL_TAB_COLUMNS
WHERE COLUMN_NAME
NOT LIKE 'ID%' AND
DATA_TYPE IN ('NUMBER', 'VARVHAR2','CHAR')
```

Una volta estratto il risultato della query si è passati all'analisi nel dettaglio, selezionando i campi con i seguenti criteri:

- Attributi DATE significativi, come DataNascita, DataDecesso, DataDocumento, ecc.
- Attributi CHAR riguardanti nome, cognome e altri identificativi diretti
- Attributi CHAR riguardanti la localizzazione del soggetto fino al dettaglio di CAP
- Attributi char contenenti not, messaggi, codici di errore

Mettere esempio tabella in cui scelgo i campi da offuscare

2.3.3 Offuscamento

Una volta terminata l'individuazione dei campi sensibili occorre creare delle soluzioni per proteggere tali dati. In tal senso esistono due possibilità: l'offuscamento e la crittografia.
Tali tecniche consistono nel modificare un'informazione rendendola non più fruibile mediante l'utilizzo di una o più chiavi. In particolare l'offuscamento consiste nell'applicazione di chiavi unilaterali, ovvero una volta utilizzata non è più possibile tornare indietro. Offuscare un dato implica quindi la perdita

completa di quell'informazione, ma a differenza della completa cancellazione, i dati rimangono comunque persistenti nella base dati. In maniera analoga ai task di cancellazione descritti precedentemente la logica di mascheramento dei dati sviluppato all'interno di RGI è la seguente:

- Ai valori contenuti nei campi di timpo stringa viene sostituito il valore ***
- Ai valori contenuti nei campi di tipo numerico vine sostituito il valore 99
- Ai valori contenuti nei campi di tipo data viene sostituito il valore 01/01/3000

Il task di offuscamento può essere eseguito quindi in alternativa al task di cancellazione successivamente al task di ricerca. Quindi anche in questo caso, prima di offuscare un soggetto, è necessario controllare che non sia legato ai sottosistemi dei sinistri e dei contratti.

Capitolo 3

Pianificazione della Sicurezza

In questo capitolo si pone attenzione al tema della Privacy by Design e by Default, ovvero all'articolo 25 della Normativa, secondo cui il titolare del trattamento mette in atto misure tecniche adeguate volte ad attuare i principi della protezione dati. In particolare, verranno presentate alcune delle tecniche utilizzabili per far fronte ai possibili attacchi informatici.

3.1 Sicurezza Database

Oggi, il pensiero ricorrente riguardo al tema della *privacy* è quello di trovare il giusto compromesso tra protezione del dato e utilità dell'informazione in esso contenuta. Anonimizzare più del necessario potrebbe comportare una perdita dell'informazione statistica, quindi la difficoltà consiste nel trovare l'opportuno *subset* di dati da rendere sicuro.
E' opportuno fare una breve descrizione della tipologia di dati che ricorrono in questo contesto:

Identificatori I: dati che riconducono direttamente all'identità del soggetto, identificatori univoci come il codice fiscale.

Quasi Identificatori QI: dati che possono parzialmente permettere di iden-

tificare il soggetto e che, uniti ad altre informazioni, possono ricondurre all'identità. Ad esempio: il codice postale, la data di nascita, il sesso.

Attributi Sensibili S: dati che contengono l'informazione sensibile del soggetto, cioè quella che vogliamo preservare, come ad esempio una specifica patologia.[6]

3.2 Anonimizzazione

L'anonimizzazione è uno dei metodi più utilizzati in questo contesto, per via della non complessa implementazione.
La prima tecnica presentata è l'aggiunta di rumore: ovvero l'aggiunta di una variabile aleatoria che perturba la reale informazione e la rende meno accurata. Supponiamo che il dataset da anonimizzare contenga l'età in centimetri di un certo numero di soggetti:

[6] Cfr. Rosa Meo, *Privacy preserving data mining: una promessa e una sfida*, Forum Machine Learning, Applicazioni ed effetti sulla società, Campus Luigi Einaudi, Torino, 17-18 Novembre 2017

Cognome	**Altezza**	**Rumore**
Rossi	175 cm	+7
Verdi	169 cm	+2
Bianchi	180 cm	−5
Gialli	172 cm	−1

Tabella 3.1: Esempio di aggiunta di rumore

Le altezze nella tabella 3.1[7] vengono così modificate generando numeri pseudocasuali, facendo in modo che la distribuzione complessiva non venga alterata. Il valore medio prima e dopo l'aggiunta di rumore risulta approssimativamente lo stesso, mantenendo così l'accuratezza dell'informazione.
Tale tecnica risulta comunque vulnerabile:
supponiamo che il rumore venga aggiunto in maniera automatica ad ogni interrogazione al database. Simulando questo processo con la generazione di variabili uniformi, si ottiene, dopo un certo numero di iterazioni, che il valor medio dell'altezza approssima bene quello reale:
Ad esempio: supponiamo si voglia stimare l'altezza di un individuo di valore $x = 180$ cm. Ad ogni interrogazione, al database viene aggiunto un rumore, rappresentato da una variabile uniforme:

$$U \sim Unif(-10, 10)$$

Pertanto ad ogni interrogazione la nuova altezza perturbata è:

$$x'_i = x + u_i$$

dove u_i è la realizzazione della variabile uniforme U all'interrogazione i-esima,

[7] Cfr. Giuseppe D'Acquisto, Maurizio Naldi *Big Data and Privacy by Design*, I diritti nella "rete" della rete, Anonimizzazione Pseudonimizzazione Sicurezza, Prefazione di Franco Pizzetti, G.Giappichelli Editore (2017).

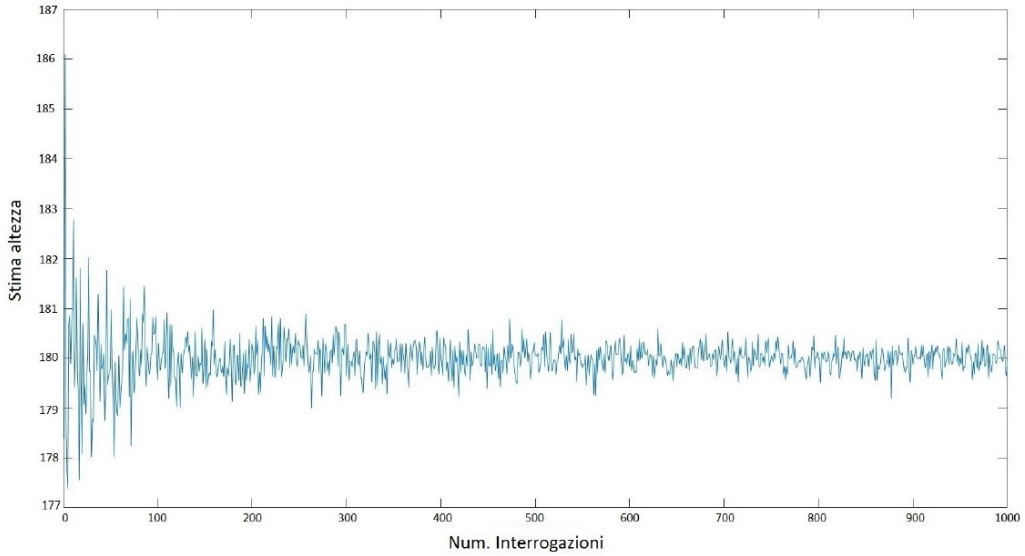


Figura 3.1: Stima dell'altezza x

con valore compreso tra -10 cm e 10 cm.
Dopo n iterazioni è possibile calcolare una stima dell'altezza:

$$x' = \sum_{k=1}^{n} \frac{x_i}{n}$$

Dopo un certo numero n di aggiunte di rumore, la stima dell'altezza risulta essere molto precisa, ed è quindi possibile risalire al dato iniziale.

Nella simulazione in Figura 3.1 sono state calcolate diverse stime dell'altezza iniziale di 180 cm, utilizzando un numero variabile di interrogazioni. In questo esempio, a partire da circa 100 iterazioni, la stima dell'altezza risulta essere molto precisa.
Ciò dimostra che l'aggiunta di rumore, spesso, non riesce a garantire un livello di sicurezza opportuno.

3.3 K-Anonimity

K-Anonimity è una tecnica che rientra nella famiglia dell' anonimizzazione per generalizzazione: l'idea è quella di rendere meno precisi e dettagliati i record all'interno di una tabella.

Definizione K-Anonimity Si dice che un set di dati ha la proprietà di k-anonimity se, per ogni individuo, l'informazione ricercata non può essere distinta dagli altri $k-1$ individui nello stesso set di dati.

In altre parole, l'obiettivo è generalizzare le singolarità, ovvero quei record che identificano direttamente il soggetto.

Supponiamo di disporre del seguente set di dati non anonimizzato:

Nome	Sesso	Età	Domicilio	Religione	Malattia
John	Maschio	29	Londra	Cristiana	No malattia
Marc	Maschio	24	Parigi	Musulmana	Diabete
Joe	Maschio	28	Parigi	Cristiana	Infezione virale
George	Maschio	27	Roma	Cristiana	Diabete
Jack	Maschio	24	Londra	Cristiana	Infezione virale
William	Maschio	23	Roma	Indu	Ipertensione
Oscar	Femmina	19	Roma	Musulmana	Ipertensione
Emily	Maschio	29	Londra	Cristiana	Diabete
Amelia	Femmina	17	Roma	Indu	No malattia
Harry	Femmina	19	Parigi	Musulmana	Diabete
Thomas	Femmina	20	Parigi	Indu	Infezione virale

Figura 3.2: Dataset non anonimizato

Supponiamo di voler proteggere l'informazione sensibile dello stato di salute dei soggetti presenti nella tabella, ma allo stesso tempo di non perdere del tutto la significatività dei record. In genere esistono due modi per raggiungere la *k-anonimity*:

Eliminazione: Consiste nel sostituire i valori di un attributo con il simbolo '*'. In questo caso vengono coperti gli attributi Nome e Religione.

Generalizzazione: Cosniste nel sostituire i valori di un attributo con un intervallo più o meno ampio. Ad esempio consideriamo le fasce d'età.

In questo caso è stata ottenuta la proprietà $2-anonimity$ rispetto agli attributi Sesso, Età e Domicilio. Come si vede dalla figura 3.3[8], esistono almeno 2 record aventi gli stessi valori di sesso, età e domicilio; pertanto non sarà più possibile identificare univocamenteil tipo di malattia associata al singolo soggetto.

[8] Cfr. https://en.wikipedia.org/wiki/K-anonymity

Nome	Sesso	Età	Domicilio	Religione	Malattia
*	Maschio	20 <Età≤ 30	Londra	*	No malattia
*	Maschio	20 <Età≤ 30	Parigi	*	Diabete
*	Maschio	20 <Età≤ 30	Parigi	*	Infezione virale
*	Maschio	20 <Età≤ 30	Roma	*	Diabete
*	Maschio	20 <Età≤ 30	Londra	*	Infezione virale
*	Maschio	20 <Età≤ 30	Roma	*	Ipertensione
*	Femmina	Età ≤ 20	Roma	*	Ipertensione
*	Maschio	20 <Età≤ 30	Londra	*	Diabete
*	Femmina	Età≤ 20	Roma	*	No malattia
*	Femmina	Età≤ 20	Parigi	*	Diabete
*	Femmina	Età≤ 20	Parigi	*	Infezione virale

Figura 3.3: Dataset anonimizzato

3.4 Linkage attack

Un esempio evidente del fallimento dell'anonimizzazione è il caso dell'identificazione dei dati sanitari del governatore del Massachusets William Weld alla fine degli anni novanta.
La seguente applicazione è tratta da un articolo scritto da Latanya Sweeney nel 2002[9] Il GIC, Commissione Insurance Group, è stata un'agenzia governativa che raccoglieva dati sanitari di migliaia di dipendenti statali. Il database del GIC conteneva queste informazioni e erano stati rimossi alcuni attributi come il nome e l'indirizzo, cercando di rendere il più possibile anonimi i soggetti interessati.
La struttura del dataset era la seguente:

- Zip Code (QI)
- Data Nascita (QI)

[9]Latanya Sweeney *K-Anonimity: a model for protecting privacy*, International Journal on Uncertainty, Fuzzies and Knowledge-based System, 10(5),(2002);557-570

- Sesso (QI)
- Nazionalità (QI)
- Diagnosi (S)
- Cura (S)

Nel 1997 Latanya Sweeney, allora studente del MIT, comprò alla cifra di 20 dollari le registrazioni al voto di due distretti di Cambridge, in cui risedeva il governatore William Weld. Le registrazioni contenevano varie informazioni tra cui identificatori e quasi-identificatori dei votanti:

- Nome (I)
- Indirizzi (QI)
- Data registrazione (QI)
- Zip Code (QI)
- Data Nascita (QI)
- Sesso (QI)
- Ultimo voto (QI)

Sweeny dimostrò che incrociando i due dataset in maniera opportuna, era possibile recuperare i dati sanitari del governatore Weld e quindi ricondursi al dato sensibile della diagnosi.

Attraverso una semplice "Join" tra i due dataset, con chiave ZIP CODE, DATA NASCITA e SESSO, Sweeny scoprì la presenza di sei record con la stessa data di nascita del governatore; soltanto tre di loro erano uomini e uno con il codice zip di Weld. In questo modo riuscì a risalire ai dati clinici di Weld individuando la presenza della sua malattia.

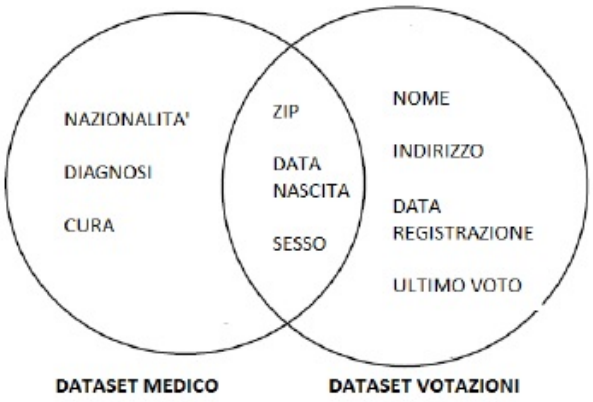


Figura 3.4: Join tra due dataset

3.5 Statistical Database

I database statistici sono database a cui è possibile accedere solo attraverso interrogazioni aggregate (SUM, COUNT, AVG). E' possibile estrarre informazioni riguardanti non uno specifico soggetto ma un insieme di persone. Tuttavia in alcuni casi è possibile risalire al singolo individuo.
Supponiamo di avere a disposizione un database medico, il cui accesso è consentito solo per scopi di ricerca o statistici. Ad esempio, supponiamo contenga il nome della persona e una variabile boolana che indica l'essere affetto o non affetto da una determinata malattia:
Ipotizziamo che siano permesse solamente interrogazioni del tipo: *"Quante persone tra le prime n sono affette da malattia?"*. Siamo interessati in particolare al cognome X_5.
Sapendo che il record corrispondente di X_5 si trova in posizione 5, è sufficiente fare la differenza tra il risultato di due query aggregate:
"Quante persone tra le prime 5 sono affette dalla malattia?"

Cognome	**Affetto da malattia**
X_1	SI
X_2	NO
X_3	NO
X_4	NO
X_5	SI
X_6	NO

Tabella 3.2: Esempio di tabella con attributo sensibile

"Quante persone tra le prime 4 sono affette dalla malattia?"
Se la differenza è 1 allora la persona X_5 è affetta dalla malattia. In questo caso il risultato della prima query è 2, mentre quello della seconda query è 1. Come si vede, quindi, anche laddove non sia possibile creare query specifiche su un singolo record, si riesce a risalire comunque all'informazione sensibile di interesse.

3.6 Questionari Polarizzati

Molto spesso accade che si vogliano reperire informazioni statistiche su un gruppo di persone attraverso un questionario individuale. Se, però, le domande riguardano temi particolarmente sensibili o privati, difficilmente le persone risponderanno onestamente. Per questo motivo, il risultato del questionario non sarebbe reale, ovvero affetto da un errore sistematico (indagine polarizzata).
Per ovviare a questo problema, esiste un meccanismo di randomizzazione delle risposte che permette alla persona che risponde al questionario di fornire una risposta più o meno casuale, e allo stesso tempo permette a chi conduce l'indagine di avere una stima accurata della risposta complessiva.
L'idea di base è la seguente:

Supponiamo essere interessati a stimare il numero di persone che posseggono una certa proprietà A all'interno di un gruppo. Ad ogni partecipante viene fornita la seguente procedura per rispondere al questionario:

- Lancia una moneta
- Se il risultato è **croce**, allora rispondi sinceramente
- Se il risultato è **testa**, allora lancia di nuovo la moneta: - Rispondi "SI" se **testa** oppure "NO" se **croce**

La privacy è garantita dal fatto che chi conduce l'indagine non può sapere se la singola risposta è reale o casuale. Tuttavia è possibile stimare con una certa accuratezza il numero di persone che possiedono la proprietà A. Sia a la frazione reale di persone aventi la proprietà A, allora il numero atteso di "SI" è S:

$$S = 1/4(1 - a) + 3/4(a) = 1/4 + a/2$$

Quindi possiamo stimare la frazione reale di persone aventi la proprietà A come:

$$a = 2S - 1/2$$

3.6.1 Stima per questionari randomizzati

Tale meccanismo fu formalizzato per la prima volta da Stanely Warner nel 1965.
Supponiamo di essere sempre interessati a stimare il numero di persone appartenenti ad un gruppo piuttosto che ad un altro. I due gruppi (A e B)

Ruota	**Persona**	
	Gruppo A	Gruppo B
Gruppo A	SI	NO
Gruppo B	NO	SI

Tabella 3.3: Risposte randomizzate fornite da un questionario

potrebbero rappresentare la presenza o meno di una malattia, oppure l'appartenenza ad una fascia di reddito (informazioni sensibili)
Introduciamo un dispositivo che indica casualmente uno dei due gruppi, ad esempio una ruota che gira e che si posiziona virtualmente sul gruppo A con probabilità p e sul gruppo B con probabilità $1 - p$.
Ogni persona fa girare privatamente la ruota e risponde "SI" se la ruota indica il reale gruppo di appartenenza, e "NO" nel caso contrario (Tabella 3.3)[10]
Indichiamo ora con v la probabilità effettiva che una persona appartenga al

gruppo A. Dalla tabella si evince che una persona risponde SI con probabilità:

$$P[SI] = pv + (1 - p)(1 - v)$$

ovvero la probabilità che la ruota indichi A per la probabilità effettiva di appartenere ad A più la probabilità che la ruota indichi B per la probabilità effettiva di appartenere a B. In modo analogo la probabilità che la risposta sia NO è:

$$P[NO] = (1 - p)v + p(1 - v)$$

Sia n_{SI} la frazione di risposte SI ottenute sul totale degli n questionari. Possiamo quindi scrivere la funzione di verosimiglianza L corrispondente a n_{SI}

[10]Cfr. Giuseppe D'Acquisto, Maurizio Naldi *Big Data and Privacy by Design*, I diritti nella "rete" della rete, Anonimizzazione Pseudonimizzazione Sicurezza, Prefazione di Franco Pizzetti, G.Giappichelli Editore (2017).

risposte SI e $n - n_{SI}$ risposte NO:

$$L = P[SI]^{n_{SI}} \cdot P[NO]^{(n-n_{SI})}$$
$$= [pv + (1-p)(1-v)]^{n_{SI}} \cdot [(1-p)v + p(1-v)]^{(n-n_{SI})}$$

. Essendo interessati a stimare la probabilità v, ovvero la reale probabilità di appartenenza ad A, è possibile cercare il valore che massimizza la funzione di verosimiglianza. Utilizziamo per semplificazione il logaritmo della verosimiglianza (essendo il *log* una funzione monotona, il risultato sarà lo stesso).

$$logL = n_{SI} \cdot log[pv + (1-p)(1-v)] + (n - n_{SI}) \cdot log[(1-p)v + p(1-v)]$$

Cerchiamo il valore di v che massimizza L:

$$\frac{\partial logL}{\partial v} = 0$$
$$\frac{(n - n_{SI})(2p-1)}{(1-p)v + p(1-v)} = \frac{n_{SI}(2p-1)}{pv + (1-p)(1-v)}$$

da cui si ottiene la stima di massima verosimiglianza

$$\hat{v} = \frac{1}{2p-1}\left(p - 1 + \frac{n_{SI}}{n}\right)$$

3.6.2 Precisione dei questionari randomizzati

Nei questionari a risposta randomizzata è presente una componente di casualità dovuta al modo in cui vengono ottenute le risposte (esempio testa o croce), ed un'altra componente di casualità dovuta al campione di popolazione scelta. Siamo interessati a valutare l'errore presente all'interno dello stimatore calcolato precedentemente, in quanto variabile aleatoria.
Il numero atteso di SI n_{SI} non è altro che una variabile binomiale con media

n per la probabilità di ottenere un SI:

$$E[n_{SI}] = nP[SI]$$

e varianza:

$$Var[n_{SI}] = nP[SI](1 - P[SI])$$

dove abbiamo calcolato precedentemente

$$P[SI] = pv + (1 - p)(1 - v)$$

Quindi:

$$\begin{aligned} &E[\hat{v}] = \frac{1}{2p-1}\left(p - 1 + \frac{E[n_{SI}]}{n}\right) = \\ &\frac{1}{2p-1}\left(p - 1 + \frac{nP[n_{SI}]}{n}\right) = \\ &\frac{1}{2p-1}\left(p - 1 + \frac{n[pv + (1-p)(1-v)]}{n}\right) = v \end{aligned}$$

(3.1)

$$\begin{aligned} &Var[\hat{v}] = \frac{Var[n_{SI}]}{(2p - 1^2)n^2} = \\ &\frac{nP[SI]P[NO]}{(2p - 1^2)n^2} = \\ &\frac{[pv + (1-p)(1-v)] \cdot (1-p)v + (1-v)]}{(2p - 1^2)n^2} = \\ &\frac{1}{n}\left(\frac{1}{16(p - 1/2)^2} - (v - 1/2)^2\right) \end{aligned}$$

Il valore atteso della stima di v risulta esattamente uguale a v, quindi la stima non è affetta da errore *bias*. La varianza diminuisce al crescere della numerosità n del campione, migliorando la precisione.

Capitolo 4

Differential Privacy

La *Differential Privacy* (Privacy differenziale) è una tecnica che rappresenta una solida garanzia di privacy individuale.
I primi passi verso il modello di privacy differenziale furono compiuti nel 2003, ma soltanto nel 2006 Dwork iniziò a formalizzare una vera e propria teoria. L'assunto di base è che anche se l'attaccante conosce tutti i record in un dataset, eccetto uno (quello di interesse), ancora non sarà in grado di inferire l'informazione contenuta nel record sconosciuto. Sotto quest'ipotesi, la privacy differenziale garantisce l'esistenza di una bassa probabilità di risalire al dato sconosciuto di interesse.

4.1 Nozioni Preliminari

La figura seguente è un esempio di modello di attacco. Supponiamo si voglia preservare la privacy per n record all'interno di un dataset D. Inoltre, l'attaccante possiede tutte le informazioni su $n-1$ record, eccetto l'n-esimo di suo interesse. La conoscenza sugli $n-1$ record viene definita *Background Information.*
Come si è visto precedentemente, facendo la differenza di query aggregate sugli n record del dataset D, è possibile identificare facilmente l'informazio-

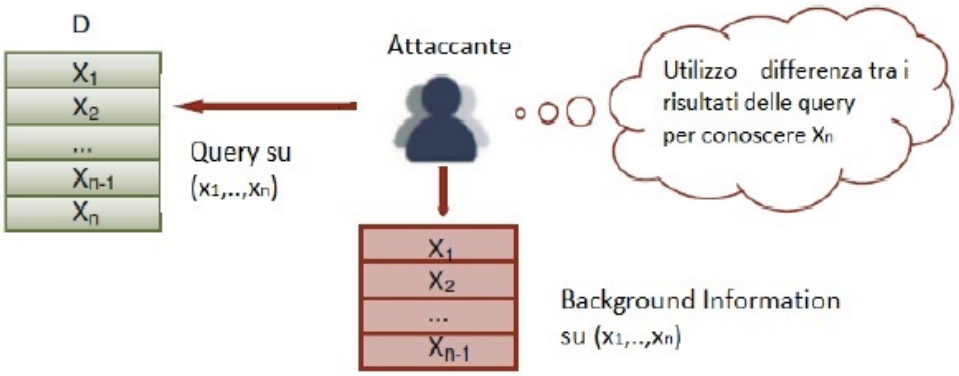


Figura 4.1: Principio della Differential Privacy

ne contenuta nel record n.

La privacy differenziale ha come obiettivo quello di proteggere il dataset da questo tipo di attacco.
Definiamo un nuovo dataset D' (*Neighbour Dataset*) che differisce dal dataset iniziale D solo per un record, x_n. Sia, inoltre, f una query a disposizione dell'attaccante che effettua su entrambi i dataset. Essendo i due dataset molto simili, la probabilità che le due query diano lo stesso risultato S è molto alta. Quando l'attaccante non riesce ad individuare una differenza tra i risultati delle due query, allora il record x_n si può considerare protetto. Se tale proprietà è valida per tutti i record di D, allora il dataset riesce a preservare interamente la privacy[11].
Prima di formalizzare la definizione di *Differential Privacy* introduciamo le seguenti notazioni:

[11]Tianqing Zhu, Gang Li, Wanlei Zhou, Philip S. Yu *Differential Privacy and Apllications*, Advances in Information Security, vol.69, Springer (2017).

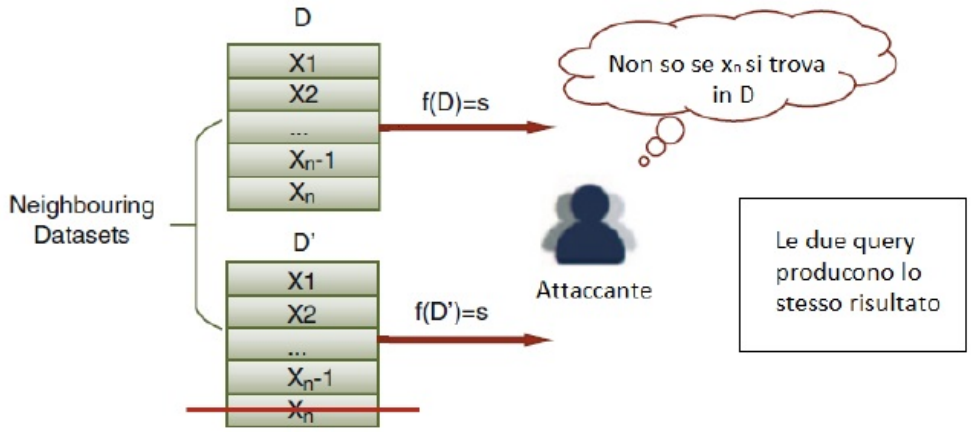


Figura 4.2: Neighbouring Dataset

χ :spazio finito di dati di dimensione $|\chi|$
Sia r un record con d attributi, e D un dataset con n record campionati dallo spazio χ. Due dataset D e D' si dicono 'vicini' (*neighbouring datasets)* se differiscono solo per un record.
Una query f è una funzione $f : D \to Y$ a valori in uno spazio generico ad esemio R. Inoltre denotiamo con F un gruppo di query e con m il numero di query all'interno di F.
L'obiettivo della *differential privacy* è quello di mascherare la differenza della query sui due dataset vicini. Definiamo la sensitività Δf cone la massima differenza tra i risultati della query f, che indica la quantità di perturbazione necessaria. La risposta perturbata è denotata con il simbolo $'\wedge'$, ad esempio $\hat{f}(D)$ denota la risposta randomizzata della query f sul dataset D.
La tabella seguente riassume le notazioni che verranno utilizzate.

Notazione	**Significato**
χ	Spazio o universo
D	Dataset, training set
D'	Neighbour dataset
D	Distribuzione dtataset
n	Numero di record nel dataset
r	Record
d	Numero di attributi nel dataset
f	Query
F	Query set
m	numero di query in F
M	Meccanismo
$\hat{f}$	Output perturbato
Δf	Sensitività
ϵ	Privacy budget

4.2 Definizione di Differential Privacy

Defiizione-Differential Privacy Si dice che un Meccanismo/Algoritmo randomizzato M soddisfa una (ϵ, δ)-*differential privacy* se per ogni set di output S e per ogni *neighbour dataset* D, D', M soddisfa:

$$Pr[M(D) \in S] \leq e^{\epsilon} \cdot Pr[M(D') \in S] + \delta$$

Dove M rappresenta il meccanismo/algoritmo randomizzato, ovvero un algoritmo che alle query su un dataset fornisce risposte a cui è stato introdotto rumore. $M(D)$ e $M(D')$ sono le risposte fornite dal meccanismo applicato a D e D'. La figura 4.3[12] mostra il meccanismo applicato ai due dataset.

[12]Cfr. Tianqing Zhu, Gang Li, Wanlei Zhou, Philip S. Yu *Differential Privacy and Apllications*, Advances in Information Security, vol.69, Springer (2017).

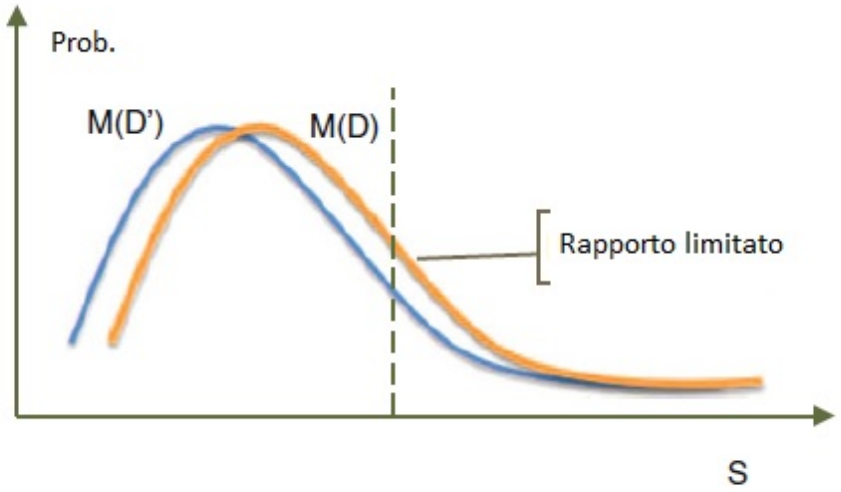


Figura 4.3: Differenza tra meccanismi

Per un particolare insieme S delle possibili risposte, il rapporto tra le due probabilità è limitato da e^{ϵ}.

Se $\delta = 0$, il meccanismo M fornisce una ϵ-*differential-privacy* (pura).

Definizione-Privacy Budget Il parametro ϵ è detto *privacy-budget*, che controlla il livello di protezione del meccanismo M. Minore è il valore di ϵ, maggiore sarà la il livello di protezione garantita dal meccanismo M. Di solito vengono utilizzati valori di $\epsilon = 0.1$ o $\epsilon = \ln 2$

4.3 Sensitività

La *sensitività* determina la quantità di rumore richiesta dal meccanismo per preservare la privacy. In particolare rappresenta la differenza tra le risposte ottenute interrogando i *neighbour dataset*. Un valore piccolo di sensitività

indica una piccola distorsione. Esistono due tipi di sensitività: la *sensitività globale* e la *sensitività locale.*

Definizione-Sensitività Globale Per una query $f : D \longrightarrow R$, la *sensitività Globale* fi f è definita come:

$$\Delta f_{GS} = \max_{D,D'} ||f(D) - f(D')||_1$$

La sensitività globale lavora bene con query come COUNT e SUM. Ad esempio per la query COUNT, di solito $\Delta f_{GS} = 1$. Per query come AVG o MEDIAN, la sensitività globale fornisce valori elevati rispetto a quelli reali. Pertanto introduciamo la sensitività locale.

Definizione-Sensitività Locale La sensitività locale tiene conto della differenza tra i record all'interno dei dataset. E' definita come segue:

Per una query $f : D \longrightarrow R$, la *sensitività locale* è:

$$\Delta f_{LS} = \max_{D'} ||f(D) - f(D')||_1$$

Per query come MEDIAN, la sensitività locale è molto inferiore alla sensitività globale, ma varia localmente quindi non può essere applicata direttamente al meccanismo. E' opportuno quindi definire un limite ad essa.

Definizione-Limite Variabile Sia $\beta > 0$, una funzione $B : D \longrightarrow R$ è *limite variabile superiore* della sensitività locale di f se soddisfa:

$$\forall D \in X : B(D) \geq f_{LS}(D)$$

$$\forall D, D' \in X : B(D) \leq e^{\beta} S(D)$$

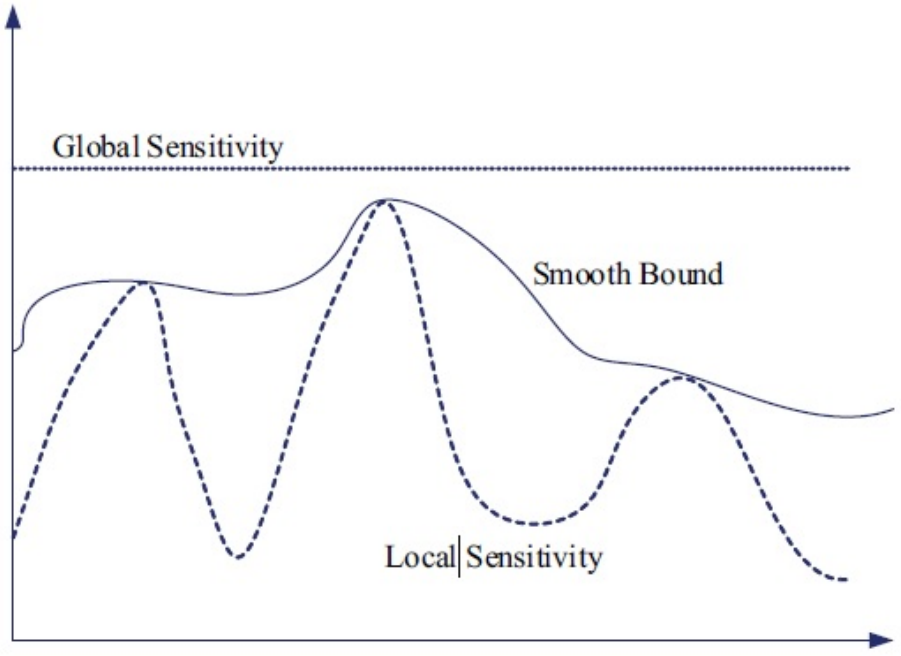


Figura 4.4: Confronto tra Sensitività

La figura 4.4[13] mostra la relazione tra sensitività locale, limite variabile e sensitività globale. Per alcune query la sensitività locale è inferiore alla sensitività globale, mentre per query come COUNT o RANGE, sono identiche.(in genere si utilizza la globale).

[13]Cfr. Tianqing Zhu, Gang Li, Wanlei Zhou, Philip S. Yu *Differential Privacy and Apllications*, Advances in Information Security, vol.69, Springer (2017).

4.4 I Principali Meccanismi della Differential Privacy

I tre principali meccanismi utilizzati per garantire la *differential privacy* sono: il meccanismo di Laplace, il meccanismo Gsussiano e il meccanismo Esponenziale

4.4.1 Il Meccanismo di Laplace

Il meccanismo di Laplace consiste nell'aggiungere rumore al risultato della query prima che arrivi all'utente. Il rumore è generato da una distribuzione di Laplace, centrata in 0 e con parametro di scala b. Il rumore è indicato con $Lap(b)$, in cui maggiore è b, maggiore è la quantità di rumore. Ricordiamo la corrispondente funzione di densità di Laplace:

$$Lap(x) = \frac{1}{2b} e^{-\frac{|x|}{b}}$$

Il meccanismo è definito nel modo seguente:

Definizione-Meccanismo di Laplace Sia $f : D \longrightarrow R$ una funzione definita su un dataset D. Si dice che il meccanismo M fornisce una ϵ *differential privacy* se soddisfa:

$$M(D) = f(D) + Lap\left(\frac{\Delta f}{\epsilon}\right)$$

Il meccanismo mostra che la quantità di rumore è legata alla sensitivtà della query f. Maggiore è la sensitività, maggiore è la quantità di rumore da aggiungere.

4.4.2 Il Meccanismo Gaussiano

Prima definire il meccanismo, introduciamo i concetto di sensitività ℓ_2

Definizione-Sensitività ℓ_2 Per una query $f : D \longrightarrow R$, la sensitività ℓ_2 di f è definita:

$$\Delta_2 f = \max_{D,D'} ||f(D) - f(D')||_2$$

Il meccanismo Gaussiano con parametro σ aggiunge rumore da una distribuzione Normale di media 0 e varianza σ.

Definizione-Meccanismo Gaussiano Sia $f : D \longrightarrow R$ una funzione definita su un dataset D, e sia N una variabile gaussiana di media 0 e varianza σ^2, il meccanismo M fornisce una (ϵ, δ)-*differential privacy* se soddisfa:

$$M(D) = f(D) + N$$

Il meccanismo Gaussiano utilizza lo stesso principio di aggiunta di rumore del meccanismo di Laplace.

4.4.3 Il Meccanismo Esponenziale

Il meccanismo esponenziale è di solito utilizzato per le query non numeriche. Si utilizza una *score-function* $q(D, \phi)$ che misura la qualità dell'output ϕ.
In particolare un Meccanismo M si dice ϵ-differential privacy se:

$$M(D) = \text{ritorna } \phi \text{ con probabilità proporzionale a } :$$
$$exp(\frac{\epsilon q(D, \phi}{2\Delta q})$$

Nome	**Malattia**
Ally	HIV
John	Diabete
Harry	Ipertensione
Bob	Diabete
David	HIV
...	...
Marc	Ipertensione

Tabella 4.1: Esempio di Dataset con attributo sensibile Malattia

Esempio

In questo esempio vengono illustrati i concetti principali di sensitività, privacy budget e meccanismo. La tabella 4.1 mostra un datatset D contenente dati medici di un distretto.

Siamo interessati a capire come il meccanismo M possa garantire la privacy per ogni individuo in D.

Supponiamo che l'attaccante sia in possesso della query f_1 : *Quante persone sono affette da HIV?*

Essendo una query numerica, possiamo utilizzare il meccanismo di Laplace. Per prima cosa calcoliamo la sensitività di f_1. Modificando un record alla volta dal dataset D, il risultato della query può variare al più di 1. Quindi la sensitività di f_1 è $\Delta f_1 = 1$. Inoltre impostiamo il valore della *privacy budget* $\epsilon = 1.0$. Come da definizione, viene generato un rumore di Laplace $Lap(1)$ che si somma alla reale risposta $f_1(D)$: $M(D) = f_1(D) + Lap(1)$. Se ad esempio la risposta reale alla query è 10, allora la risposta rumorosa potrebbe essere 11.

Supponiamo ora che l'attaccante abbia a disposizione la query f_2:

Qual'è la malattia più comune presente nel database?. Essendo una query non numerica, possiamo applicare il meccanismo esponenziale. La tabella

Malattia	**Numero di persone**	$\epsilon = 0$	$\epsilon = 0.1$	$\epsilon = 1$
Diabete	24	0.25	0.32	0.12
Ipertensione	8	0.25	0.15	$4x10^{-5}$
Infezione virale	28	0.25	0.40	0.88
HIV	5	0.25	0.13	$8.9x10{-}6$

Tabella 4.2: Meccanismo Esponenziale

4.2[14] mostra nelle ultime tre colonne le probabilità di scegliere l'output applicando il meccanismo, utilizzando una privacy budget $\epsilon = 1.0$, $\epsilon = 0.1$ e $\epsilon = 0$.

Definiamo la *score function* q di f_2 come il numero di persone per ogni malattia. Anche in questo caso, modificando un record della tabella, q varia al massimo di 1, quindi la relativa sensitività $\Delta q = 1$.
Per $\epsilon = 0$ il meccanismo fornisce la stessa risposta per tutti e 4 gli output, avremo quindi una probabilità uniforme pwe ciascuna malatia. Forniamo ora un esempio di calcolo quando $\epsilon = 1$, seguendo la definizione di meccanismo esponenziale:

$$P(Diabete) \simeq exp(\tfrac{1 \cdot 24}{2\Delta q}) = 0.12$$

$$P(Ipertensione) \simeq exp(\tfrac{1 \cdot 8}{2\Delta q}) = 4x10^{-5}$$

$$P(Infezionevirale) \simeq exp(\tfrac{1 \cdot 28}{2\Delta q}) = 0.88$$

$$P(HIV) \simeq exp(\tfrac{1 \cdot 5}{2\Delta q}) = 8.9x10^{-6}$$

Con $\epsilon = 1$ la differenza tra le probabilità di avere l'HIV rispetto alle altre malattie è significativo; ciò implica un buon livello di utilità del meccanismo,

[14] Cfr. Tianqing Zhu, Gang Li, Wanlei Zhou, Philip S. Yu *Differential Privacy and Apllications*, Advances in Information Security, vol.69, Springer (2017).

ma un livello di privacy più basso. Infatti per l'attaccante non sarà difficile intuire che l'HIV è l'infezione con una probabilità molto inferiore rispetto alle altre.
$\epsilon = 0.1$ rappresenta un buon compromesso tra privacy e utilità: la differenza tra la probabilità di HIV e la probabilità di infezione virale, che è la più alta, non è così significativa. Ciò permette di ottenere un livello di privacy e di utilità accettabile. Il caso $\epsilon = 0$, ovviamente, determina il massimo livello di privacy possibile, a discapito dell'utilità.

4.5 Misura di utilità della differential privacy

Fissato il livello di pricacy budget a ϵ, è opportuno utilizzare delle misure di utilità per valutare la performance del meccanismo della differential privacy:

- **Misura della quantità di rumore** : il modo più semplice è valutare quanto rumore è stato aggiunto all'output della query. Un rumore più piccolo implica una misura di utilità migliore
- **Misura dell'errore** : l'utilità viene calcolata come differenza tra l'output perturbato e l'output reale $|f(D) - \hat{f}(D)|$. Minore è la differenza, minore sarà la misura di utilità.

Esempio

Riprendiamo la tabella in cui sono riportati alcuni soggetti con la patologia associata:

Nome	**Malattia**
Ally	HIV
John	Diabete
Harry	Ipertensione
Bob	Diabete
David	HIV
...	...
Marc	HIV
Joe	Ipertensione

Ipotizziamo siano presenti un numero di soggetti pari a $n = 30$ e che l'attaccante sia interessato all'informazione sensibile del soggetto Marc contenuta nel record 29 della tabella. Siano inoltre f_1 e f_2 due query a sua disposizione:

f_1 : *Quante persone tra le prime 19 sono affette da HIV*
f_2 :*Quante persone tra le prime 18 sono affette da HIV*

Come illustrato nella sezione 3.6, è possibile risalire all'informazione sensibile attraverso la differenze delle due query. Ipotizziamo che la risposta alla prima query sia ad esempio 10. Se la query f_2 da come riusltato 9 allora è facile dedurre dalla differenza che Marc è affetto da HIV.
Utliliziamo il meccanismo di Laplace con l'obiettivo di preservare la privacy di Marc, ma allo stesso tempo ottenere un buon livello di utilità.
A tal proposito verranno illustrati i risultati di una simulazione con diversi valori della *privacy budget* ϵ al fine di ottenere il giusto compromesso tra privacy ed utilità.
L'idea di base è di aggiungere rumore di Laplace al risultato delle query f_1 ed f_2 in modo da ottenere una differenza perturbata:

$$\hat{f}_1 = 10 + Lap_1(\Delta f/\epsilon)$$

$$\hat{f}_2 = 9 + Lap_2(\Delta f/\epsilon)$$

Dove $Lap_1(\Delta f/\epsilon)$ e $Lap_2(\Delta f/\epsilon)$ sono la variabili aleatorie generate da una distribuzione di Laplace centrata in 0 e con parametro di scala $\Delta f/\epsilon$; Δf è la sensitività delle due query, in questo caso $\Delta f = 1$, essendo delle query COUNT. Al diminuire di ϵ, la quantità di rumore aumenta, ottenendo un livello di prvacy più elevato.
Ricordiamo che la privacy di Marc verrebbe violata solo ne caso in cui la differenza tra i due output è 1.

- $\epsilon = 2$

$$\hat{f}_1 = 10 + Lap_1(1/2) = 10 + 0.13 \approx 10$$
$$\hat{f}_2 = 9 + Lap_2(1/2) = 9 - 0.35 \approx 9$$

In questo caso avremo una differenza tra le due query = 1, violando la privacy di Marc, poichè il rumore non ha perturbato abbastanza la situazione di partenza (utilità elevata)

- $\epsilon = 1$

$$\hat{f}_1 = 10 + Lap_1(1/1) = 10 - 0.8 \approx 10$$
$$\hat{f}_2 = 9 + Lap_2(1/1) = 9 - 0.1 \approx 9$$

Anche in questo caso il rumore prodotto non è stato ancora sufficiente.

- $\epsilon = 0.5$

$$\hat{f}_1 = 10 + Lap_1(1/0.5) = 10 + 0.4 \approx 10$$
$$\hat{f}_2 = 9 + Lap_2(1/0.5) = 9 + 0.7 \approx 10$$

Finalmente con $\epsilon = 0.5$ si ottiene una differenza tra output= 0, conservando l'integrità del record relativo a Marc.
Gli esempi precedenti rappresentano dei risultati statisticamente non robusti, in quanto il rumore prodotto è una realizzazione di una variabile casuale che

può far variare facilmente il risultato atteso.
Pertanto è opportuno chiedersi: Ripetendo il meccanismo un numero elevato di iterazione, quante volte viene violata la privacy del soggetto interessato al variare di ϵ? Come varia il valore dell'utilià all'aumentare del livello di privacy? Qual'è il valore di ϵ che garantisca il giusto compromesso tra privacy e utilità?

Algoritmo

1. Fissato privacy budget ϵ
2. Applicare meccanismo di Laplace a f_1 e f_2
3. Calcolare $d = \hat{f}_1 - \hat{f}_2$
4. Se $d = 1$ allora la privacy è stata violata
5. Iterare dal punto 1 a 4 n volte
6. Calcolare quante volte sul totale di n iterazioni è stata violata la privacy
7. Funzione di (NON) utilità = $(\sum_{k=1}^{n} |10 - \hat{f}_1| + \sum_{k=1}^{n} |9 - \hat{f}_2|)/n$

L'algoritmo è stato applicato 2000 volte al variare di ϵ tra 0.05 e 2.

Come si vede dalla figura 4.5, all'aumentare del parametro di privacy budget ϵ, il livello di protezione dell'informazione di Marc diminuisce linearmente. Il criterio utilizzato per stimare la violazione della privacy di Marc è il calcolo del numero di volte sul totale in cui la differenza tra la prima e la seconda query è 1.

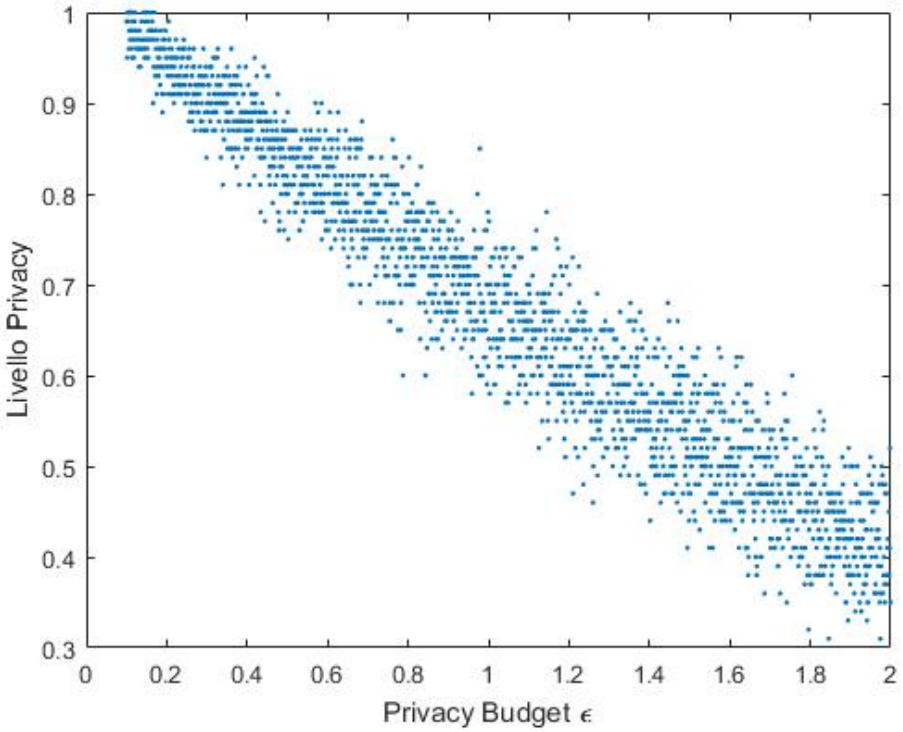

Livello Privacy
Privacy Budget ϵ
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2

Figura 4.5:

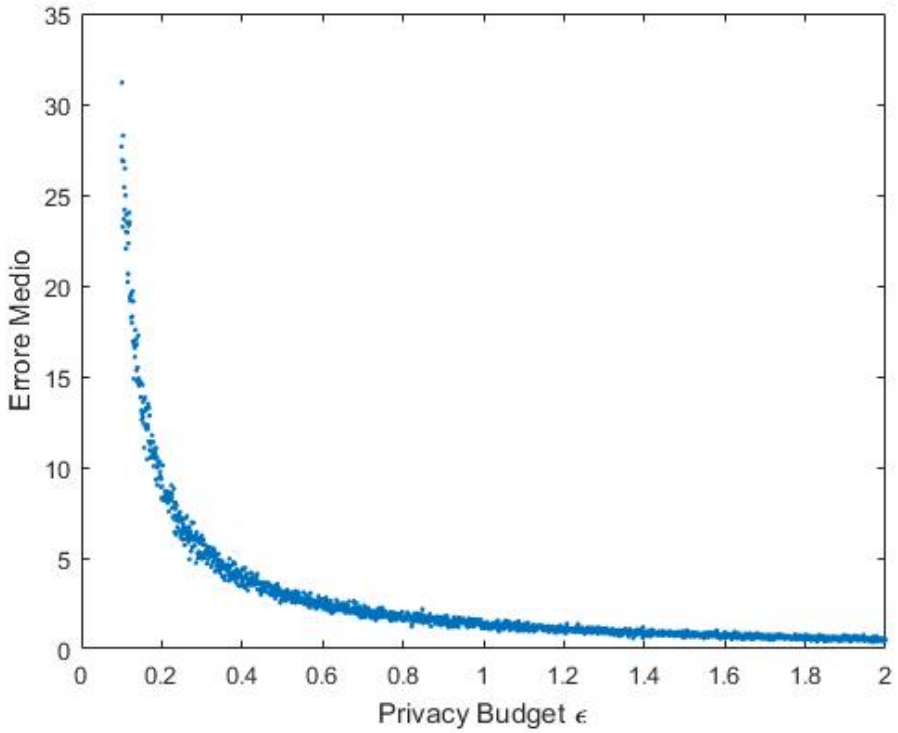

35
30
25
20
15
10
5
0
Errore Medio
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Privacy Budget ϵ

Figura 4.6:

Ovviamente, il massimo livello di privacy si otterrebbe con ϵ prossimo allo 0, che produrrebbe una quantità di rumore molto elevata.
Tuttavia, per scegliere un adeguato valore di ϵ, è opportuno valutare anche il comportamento della funzione di utilità. Nella figura 4.6 è riportato l'errore calcolato nel punto 7 dell'algoritmo, ovvero la differenza tra gli output reali delle query e quelli perturbati. L'errore medio cala drasticamente dopo un certo valore di ϵ. Il valore di $\epsilon = 0.3$ potrebbe essere un ottimo compromesso, in quanto determina un decremento significativo dell'errore, un incremento dell'utilità dell'informazione, e un livello di privacy elevato, pari – circa - al 90%. La scelta di ϵ non è sempre immediata, ma necessita di un attento confronto tra funzione di utilità e livello di privacy raggiunto. Scegliere un ϵ molto piccolo determina sì un livello di privacy elevato, ma una quantità di rumore tale da perdere completamente l'informazione di partenza.

4.6 Decision Tree: Privacy e Utilità

4.6.1 Overview Alberi Decisionali

Un albero decisionale (Decision Tree) è un grafo in cui ogni nodo interno rappresenta una variabile. Esso è utilizzato come modello predittivo: le foglie dell'albero rappresentano i valori predetti. E' possibile costruire alberi decisionali a partire da un insieme di dati e utilizzare gli attributi del *dataset* come variabili di decisione all'interno dei nodi. A partire dal nodo radice (*root*) si arriva a classificare il nodo foglia (*leaf*), attraverso un determinato cammino(*path*) in cui i record di un *dataset* vengono raggruppati a seconda della decisione presa in un nodo.
Viene, quindi, utilizzato per classificare attraverso un processo di apprendimento:
il set di dati a disposizione viene generalmente diviso in due sottoinsiemi: il *training set*, sul quale viene realizzato l'albero, e quindi il modello predittivo, e il *test set* che viene utilizzato per testare la precisione del modello.

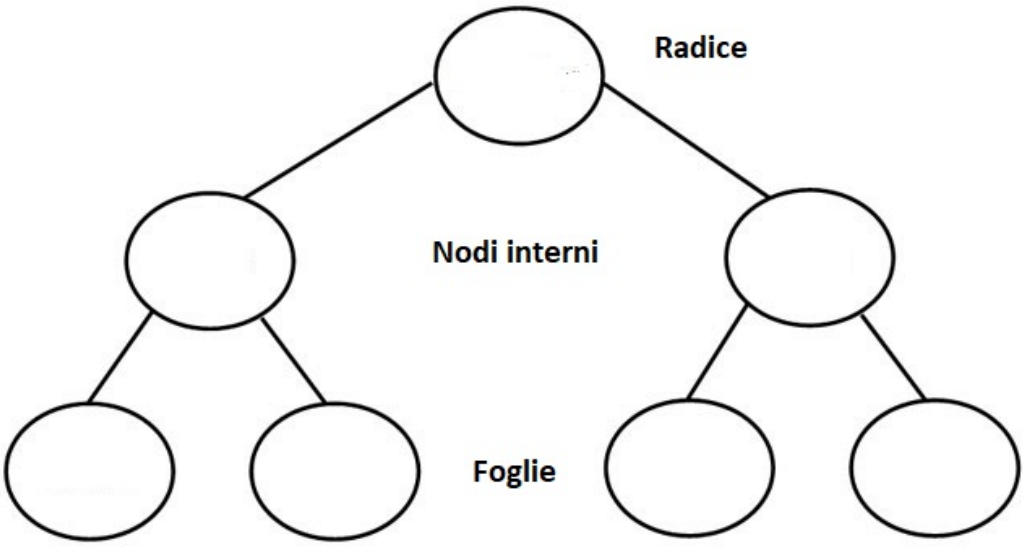


Figura 4.7: Struttura di un albero decisionale

Nel nostro contesto, i nodi sono etichettati con il nome degli attributi, gli archi rappresentano i possibili valori dell'attributo "vincente", mentre ogni foglia dell'albero rappresenta un valore della variabile *target*, corrispondente all'informazione sensibile di interesse.
In genere, il set di dati è rappresentato nel modo seguente:

$$(x, Y) = (x_1, x_2, x_3, ..., x_k, Y)$$

dove Y è la variabile dipendente, il target che proviamo a classificare, e il vettore x è costituito dagli attributi *splittati* durante la creazione dell'albero. L'idea è quindi quella di addestrare l'albero su un primo set di dati già classificati con un attributo target, per poi testare il modello e stimare la bontà delle nuove classificazioni.

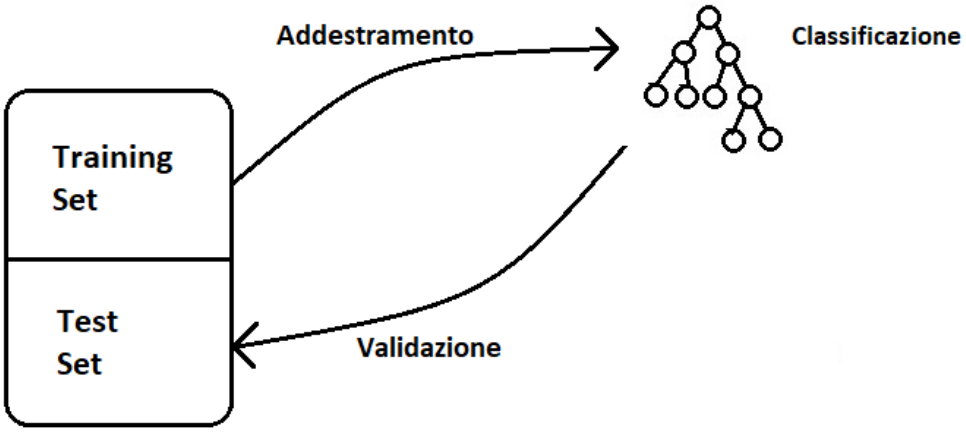


Figura 4.8: Training set e Test set

Una volta creato il modello sulla base del *training set*, è necessario testarlo attraverso una misura d'errore che stimi la bontà della classificazione. Spesso, creando un modello troppo fedele al *training set* si rischia di incorrere al fenomeno dell'*overfitting*: ovvero, l'albero "fitta" così bene i dati del training set, tanto da risultare, poi, poco flessibile e poco preciso nella classificazione di nuovi dati. Per evitare l'*overfitting*, è possibile implementare una fase di *pruning* (potatura) in cui l'albero individua gli attributi che non hanno contribuito ad una consistente suddivisione dei record ed elimina i rispettivi nodi, rimandando le istanze al livello superiore.

4.6.2 Criterio di Split

Gli algoritmi di costruzione degli alberi di decisione si basano sul modo in cui vengono scelti gli attributi da splittare. I rami vengono creati iterativamente

partizionando tutti i record e scegliendo ad ogni step l'attributo "migliore". Due delle tecniche di *split* più utilizzate sono: l'indice GINI e l'*Information Gain*.

Indice GINI

Sia t un nodo dell'albero, l'indice di Gini è definito come:

$$GINI(t) = 1 - \sum_{i=1}^{n} [p(i|t)]^2$$

dove $p(i|t)$ è la frequenza relativa della classe i al nodo t. L'indice di Gini raggiunge il suo minimo zero quando il nodo appartiene ad un unica casse. Misura l'impurità/disordine del dataset in corrispondenza del nodo t.
Una volta calcolato l'indice di Gini in corrispondenza del nodo t, si valuta la qualità della suddivisione come:

$$GINI_{split} = \sum_{i=1}^{k} Gini(i) \cdot \frac{n_i}{n}$$

k=numero di figli del nodo t
n_i=numero di record della partizione (figlio) i
n=numero di record del dataset al nodo t
Nella ricerca del nodo da splittare, si sceglie il nodo t in corrispondenza dell'attributo che fornisce il più piccolo valore di $GINI_{split}$ per partizionare il dataset.

Information Gain

Sia t un nodo dell'albero, l'*Information Gain* è defnita come:

$$Entropia(t) = -\sum_{i=1}^{n} p(i)|t \cdot log(p(i|t))$$

dove $p(i|t)$ è la frequenza relativa della classe i al nodo t.

Misura l'omogeneità/ordine di un nodo. Raggiunge il suo minimo a zero quando tutti i record appartengono ad una sola classe (massima informazione).
Una volta calcolato l'indice di Gini in corrispondenza del nodo t, si valuta la qualità della suddivisione come:

$$GAIN_{split} = Entropia(t) - [\sum_{i=1}^{k} Entropia(i) \cdot \frac{n_i}{n}]$$

k=numero di figli del nodo t
n_i=numero di record della partizione (figlio) i
n=numero di record del dataset al nodo t
L'indice $GAIN_{split}$ è denominato *Information Gain*. Si sceglie il nodo di split che raggiunge la maggiore *Information Gain*.

L'indice di Gini e la variazione di entropia sono parametri che guidano la costruzione dell'albero. Poichè, in genere, in un buon modello i nodi foglia dovrebbero essere il più possibile *puri* (ovvero contenere solo record che appartengono ad una solo classe), si cerca minimizzare il livello di impurità man mano che si scende dalla radice verso le foglie.
A seconda del criterio di split scelto, esistono diversi algoritmi per costruire un albero decisionale: l'indice di Gini viene utilizzato nell'algoritmo **Cart**, mentre l'Entropia è utlizzata in algoritmi come **ID3** e **C4.5**. In genere il rischio di usare l'**Information Gain** è quello di creare piccole e tante partizioni e di tendere ad un albero *overfitted* rispetto al training set.

Esempio di costruzione di albero con il criterio di Gini

Vogliamo ora costruire un albero decisionale considerando il seguente dataset costituito da 16 record, avente come variabile target la presenza o meno di una certa malattia e come predittori le variabili Età, Sesso e la presenza in passato della stessa malattia.

Record	**Età**	**Sesso**	**Malattia Prima**	**Malattia**
x_1	41	F	NO	SI
x_2	20	M	SI	SI
x_3	28	F	NO	NO
x_4	62	M	SI	SI
x_5	70	F	NO	NO
x_6	79	F	NO	SI
x_7	50	M	NO	NO
x_8	35	M	SI	SI
x_8	22	F	NO	SI
x_{10}	77	F	SI	NO
x_{11}	64	F	SI	NO
x_{12}	38	M	NO	NO
x_{13}	55	F	SI	NO
x_{14}	77	M	NO	SI
x_{15}	78	F	SI	NO
x_{16}	23	F	SI	NO

Al fine di utilizzare il criterio di split di Gini è necessario discretizzare la variabile quantitativa Età. Possiamo ad esempio creare tre classi A, B, C, D con i seguenti valori:

A=[20,39] B=[40,59] C=[60,80]

Il primo step è quello di scegliere il nodo radice ovvero il primo nodo da splittare. Seguendo la definizione dell'indice di Gini cerchiamo l'attributo migliore in partenza:

$$Gini(Sesso = F) = 1 - [\left(\frac{3}{10}\right)^2 + \left(\frac{7}{10}\right)^2] = 0.42$$

$$Gini(Sesso = M) = 1 - [\left(\frac{4}{6}\right)^2 + \left(\frac{2}{6}\right)^2] = 0.4444$$

$$GINI_{Split}(Sesso) = Gini(Sesso = F) \cdot \frac{10}{16} + Gini(Sesso = M) \cdot \frac{6}{16} = 0.4292$$

$$Gini(Prima = SI) = 1 - [\left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2] = 0.4898$$

$$Gini(PrimaNO) = 1 - [\left(\frac{5}{9}\right)^2 + \left(\frac{4}{9}\right)^2] = 0.4938$$

$$GINI_{Split}(Prima) = Gini(Prima = SI)\cdot\frac{7}{16} + Gini(PRIMA = NO)\cdot\frac{9}{16} = 0.4921$$

$$Gini(Eta \in A) = 1 - [\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2] = 0.5$$

$$Gini(Eta \in B) = 1 - [\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2] = 0.444$$

$$Gini(Eta \in C) = 1 - [\left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2] = 0.4898$$

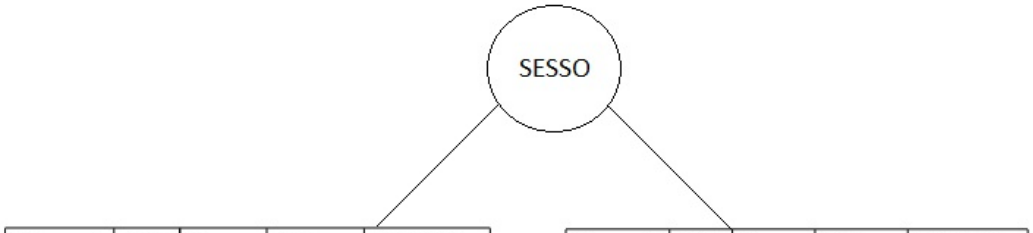
$$GINI_{Split}(Eta) = Gini(Eta \in A)\cdot\frac{6}{16} + Gini(Eta \in B)\cdot\frac{3}{16} + Gini(Eta \in C)\cdot\frac{7}{16} = 0.48918$$

Il più piccolo valore dell'indice Gini è quello relativo alla variabile Sesso, pertanto è il nodo scelto come radice dell'albero. Si ottiene quindi una prima partizione del dataset di partenza in due dataset contenenti rispettivamente Maschi e Femmine.
A questo punto basta ripetere il ragionamento per cercare un'altro nodo di split. Dalla parte del subset costituto dai maschi, si ottiene un valore di Gini index identico per le variabili Prima e Eta:

$$GINI_{Split}(Prima) = 0.2222$$

$$GINI_{Split}(Eta) = 0.2222$$

In questo caso l'algoritmo sceglierà casualmente uno dei due attributi (per semplicità scegliamo Prima).
Una volta splittata la variabile PRIMA è chiaro che è possibile classificare



Record	Età	Sesso	Prima	Malattia
x_2	20	M	SI	SI
x_4	62	M	SI	SI
x_7	50	M	NO	NO
x_8	35	M	SI	SI
x_{12}	38	M	NO	NO
x_{14}	77	M	NO	SI

Record	Età	Sesso	Prima	Malattia
x_1	41	F	NO	SI
x_3	28	F	NO	NO
x_5	70	F	NO	NO
x_6	79	F	NO	SI
x_8	22	F	NO	SI
x_{10}	77	F	SI	NO
x_{11}	64	F	SI	NO
x_{13}	55	F	SI	NO
x_{15}	78	F	SI	NO
x_{16}	23	F	SI	NO

Figura 4.9: Primo *split*

questa parte di albero, in quanto i maschi con l'attributo $Prima = SI$ hanno tutti il valore target $Malattia = SI$. Siamo giunti quindi alle prime due foglie dell'albero: Iterando il calcolo dell'indice nel subset femminile, si arriva alla configurazione dell'albero in figura 4.10.

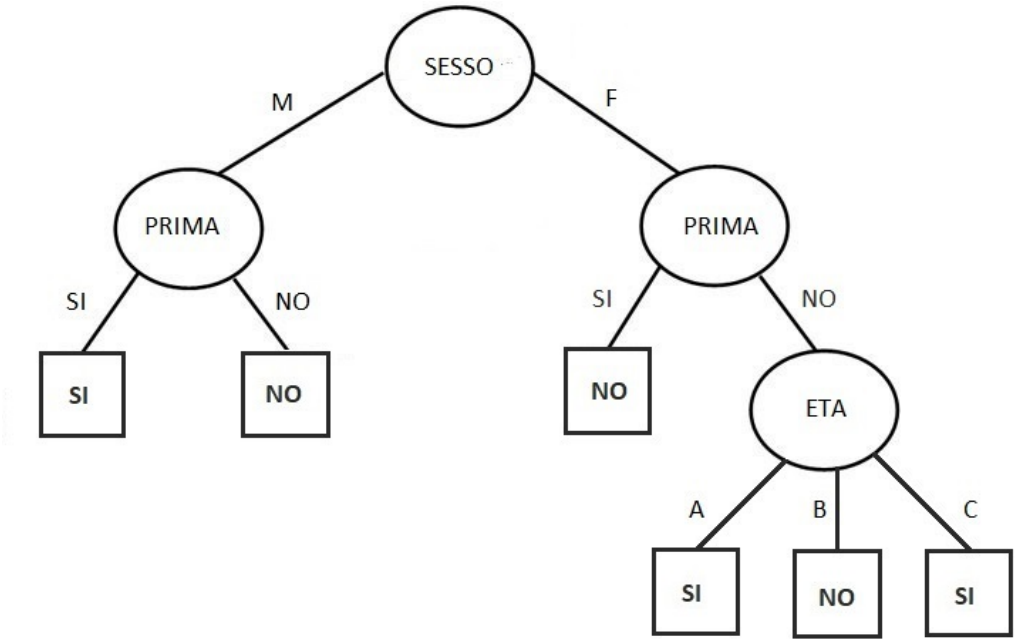


Figura 4.10:

4.6.3 Valutazione della classificazione

Una volta costruito l'albero e classificati i nodi foglia, si procede a valutare la qualità della classificazione. L'idea è quella di calcolare la distanza tra i valori reali dell'attributo target (valori osservati) e i valori classificati (valori predetti). Quindi una prima misura d'errore è quella calcolata nella classificazione sul training set (training error). Sia n_c il numero di classificazioni corrette e n il numero totale di predizioni. Due misure per valutare la classificazione sono:

$$\text{Accuratezza} = \frac{n_i}{n}$$

$$\text{Tasso d'errore} = 1 - \frac{n_i}{n}$$

Se si ottiene una buona classificazione, si procede a testare e validare il modello sul test set. In questo caso, l'algoritmo di decison tree utilizzerà il modello creato sul training set e cercherà di classificare i record presenti nel test set.
Allo stesso modo possiamo valutare la classificazione sul test set , con le misure definite precedentemente.
Tuttavia contare il numero di classificazioni corrette talvolta non è il modo migliore per valutare il modello, in quanto non abbastanza dettagliato. Ad esempio un'accuratezza dello 0.90 non è un buon indicatore se il 90 per cento dei record appartiene alla classe predetta correttamente. In questo caso la bontà della predizione deriva dal fatto che si sta classificando correttamente la classe di gran lunga più comune.

Confusion Matrix

La matrice di confusione o di errore, è una tabella che permette di valutare la performance di un algoritmo di classificazione. Le righe della matrice rappresentano i valori predetti per ogni classe, mentre le colonne i valori reali: l'elemento in posizione (i, j) è il numero di record in cui l'albero ha classificato la classe "vera" i come j. Permette di comprendere non solo la quantità di errore della classificazione, ma anche il tipo di errore commesso.
Nel caso di un classificatore binario, esistono due tipi di errore: può classificare in maniera errata un individuo con i falsi positivi (predizione errata della classe 1) ed i falsi negativi (predizione errata della classe 0).

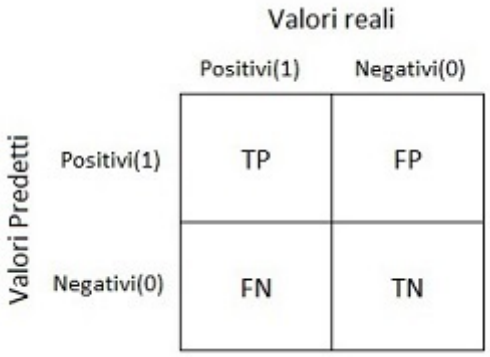


Figura 4.11: Confusion Matrix

Gli indici per la valutazione della matrice di confusione sono:

$$\textbf{Accuratezza} = \frac{TP + TN}{TP + FP + FN + TN}$$

L'accuratezza è il miglior indice nel caso in cui il dataset ha una struttura simmetrica in cui i valori dei falsi positivi e dei falsi negativi sono molto simili.

$$\textbf{Precisione} = \frac{TP}{TP + FP}$$

La precisione è la frazione delle classificazioni corrette positive sul totale delle predizioni positive Indica quanto corretta è stata la predizione (più alto possibile).

$$\textbf{Sensitività} = \frac{TP}{TP + FN}$$

La sensitività è la frazione delle classificazioni corrette positive sul tutte le osservazioni della classe dei positivi (1).

4.7 Applicazione della Differential Privacy in un problema di classificazione con Decision Tree

In questa sezione verrà utilizzato il concetto di *Differential Privacy* in un algoritmo di classificazione, applicato ad un *dataset*[15] in cui ogni *record* contiene informazioni sulla situazione economico-sociale di 1000 cittadini statunitensi.
In particolare, ipotizziamo che un attaccante abbia la possibilità di accedere al *dataset* e di costruire un modello basato sugli alberi decisionali. Lo scopo dell'attaccante è quello di estrarre informazioni riguardanti il livello di reddito dei cittadini statunitensi, con l'obiettivo di creare un modello utile per predire tale informazione sensibile.
Il *dataset* è costituito da 1000 osservazioni e 15 variabili, di cui la variabile *target* che costituisce il reddito annuo del cittadino. L'obiettivo della classificazione è quello di capire, in base alle informazioni contenute all'interno delle altre variabili, se un individuo ha un reddito maggiore - o minore di - 50.000 dollari; tale informazione è considerata sensibile e, quindi, il titolare del *dataset* è tenuto a proteggerla. La struttura del *dataset* è la seguente:

- Age - età del soggetto
- Type-employer - tipo di impiego
- Fnlwgt (non verrà utilizzata)
- Education - livello d'istruzione
- Education-lvel - livello d'istruzione in forma numerica
- Marital - stato sociale

[15] https://archive.ics.uci.edu/ml/index.php

- Occupation - occupazione
- Reletionship - situazione familiare (marito, moglie, non sposato, non in famiglia)
- Race - bianco, nero, asiatico, america-latina, altro
- Capital-gain guadagni
- Capital-loss perdite
- H-per-week - ore di lavoro settimanali
- Country - paese d'origine
- Income - reddito annuo in dollari

Prima di implementare il modello, il primo step è quello di cercare di ridurre la dimensionalità del *dataset*, individuando gli attributi non significativi. Ad esempio, la variabile *education* è strettamente correlata alla variabile $education - level$; pertanto verrà utilizzata soltanto quest'ultima. Quindi, viene semplicemente rimossa dal dataset, così come la variabile *fnwgt*, che risulta inutile al fine della classificazione.
Inoltre, gli attributi *marital* e *relationship* tendono a sovrapporsi e risultano strettamente correlati, in quanto la condizione $relationship = marito/moglie$ implica necessariamente l'attributo $marital = sposato$. Otteniamo così un dataset costituito da 11 variabili più la variabile target.
Per simulare la procedura utilizzata dall'attaccante, dividiamo il *dataset* in due parti: il *training set*, su cui verrà creato il modello, e il *test set*, su cui si valuterà la classificazione. L'obiettivo del titolare del trattamento sarà, invece, quello di preservare il più possibile la *privacy* del *test set*, ovvero fare in modo che il modello costruito dall'attaccante abbia un errore di *missclassificazione* più alto possibile. In questo modo, proveremo a preservare l'informazione sensibile del reddito del maggior numero possibile di individui.
Lo schema è – dunque - il seguente:

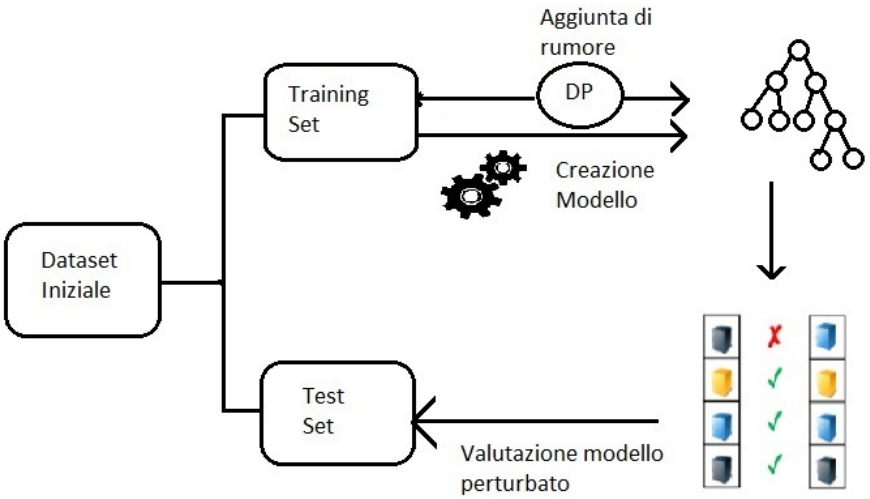


Figura 4.12: Procedura

L'attaccante crea un modello basato sui dati di training per predire l'informazione del reddito su record futuri. Per impedire una corretta classificazione il titolare del dataset perturberà i dati di partenza, cercando comunque di non perdere l'informazione statistica al loro interno. L' obiettivo è quello di cercare la giusta quantità esatta di rumore che preservi l'informazione, ma al contempo perturbi il modello costruito dall'attaccante, quel tanto che basta da renderne difficile la classificazione.
Essendo interessato a scoprire se il reddito di un soggetto supera i $50k$ dollari, la prima operazione che l'attaccante effettua è quella di trasformare la variabile reddito in una variabile booleana, per utilizzarla come variabile di target nella costruzione dell'albero (1 se il reddito annuo supera i $50k$ e 0 altrimenti). Pertanto utilizzerà una certa funzione $f : D \longrightarrow R$, che per ogni soggetto estrae il reddito annuo, per poi trasformarlo in una variabile booleana. Il ruolo della Differential Privacy, in questo caso, è quello di aggiungere rumore prima che il modello venga costruito: seguendo la definizione di meccanismo Laplaciano, verrà aggiunta una quantità di rumore alla

variabile reddito pari a:

$$X = lap(\frac{\Delta f}{\epsilon})$$

dove Δf è la sensitività di f. Possiamo attribuire a Δf il valore di 50.000 in quanto , secondo la definizione, è la quantità massima da aggiungere al reddito per determinare l'appartenenza o meno ad una classe. ϵ è il parametro di privacy budget che influisce direttamente sulla quantità di rumore prodotta. Il nostro obiettivo è quello di individuare il giusto valore di ϵ che conduca l'attaccante ad ottenere un errore di classificazione significativo, e allo stesso tempo non perturbi troppo il dataset di partenza.
Definiamo quindi il Training set 'rumoroso' come:

$$Traing\ Set\ Perturbato = Training\ Set + X$$

A questo punto verrà costruito l'albero decisionale utilizzando i seguenti parametri:

- Algoritmo = Cart
- Criterio di Split = Indice di Gini
- Variabile da classificare = Soglia reddito annuo($50k$).
- Criterio di potatura=Si

Lo step successivo è quello di misurare l'errore di classificazione, confrontando i valori predetti con quelli osservati nel Test Set.
Inoltre per stimare quanto la perturbazione abbia distorto il modello creato dall'attaccante, quest'ultimo verrà confrontato con un albero decisionale costruito sul Training set iniziale (privo di rumore).
L'algoritmo è stato simulato 2000 volte al variare di ϵ da 0.5 a 2.
La procedura utilizzata è la seguente:

Algoritmo

1. Creazione Training Set e Test Set
2. Applicazione del Meccanismo di Laplace su Training Set
3. Costruzione modello 'perturbato'
4. Calcolo della misura d'errore sul Test Set (livello di Privacy ottenuto)
5. Calcolo della misura di utilità del dataset perturbato
6. Confronto con modello non perturbato
7. Procedura (1.-6.) iterata al variare di ϵ
8. Stima del miglior valore di ϵ

Ovviamente per $\epsilon \longrightarrow \infty$ il modello perturbato tende a quello non perturbato, in quanto la quantità di rumore aggiunta dalla Distribuzione di Laplace tende a 0.
Ad ogni iterazione, viene calcolato l'accuratezza della classificazione che è una stima del numero di volte in cui l'informazione sensibile relativa al reddito viene violata dall'attaccante. Per valutare l'utilità del meccanismo, è possibile stimare la quantità media di rumore aggiunta al dataset iniziale al variare di ϵ. Una misura di utilità alternativa è rappresentata dal numero di volte in cui la perturbazione applicata alla variabile reddito, determina scostamento dalla classe di reddito di partenza:
ad esempio se un valore di reddito annuo di partenza è pari a 40.000 e la quantità di rumore applicata è 20.000, si otterrà il passaggio di classe da "$\leq 50k$ a "$\geq 50k$". In questo modo l'informazione relativa a quel particolare individuo viene perturbata quel tanto che basta da renderla statisticamente non valida.

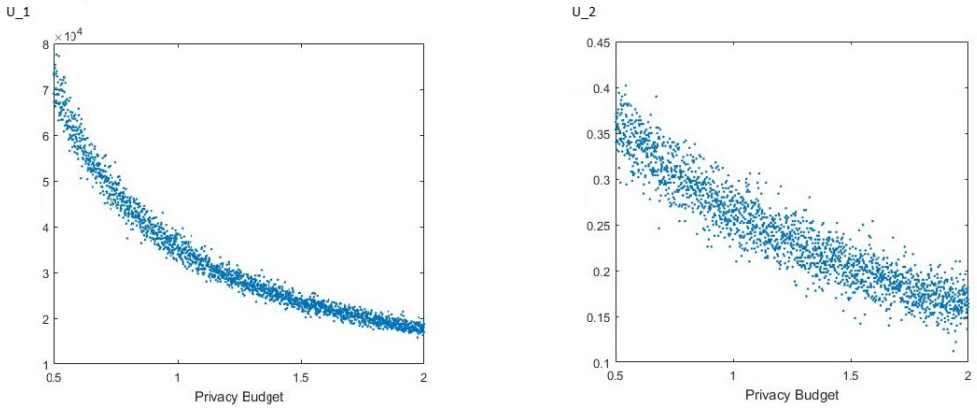


Figura 4.13: Funzioni di utilità U_1 e U_2

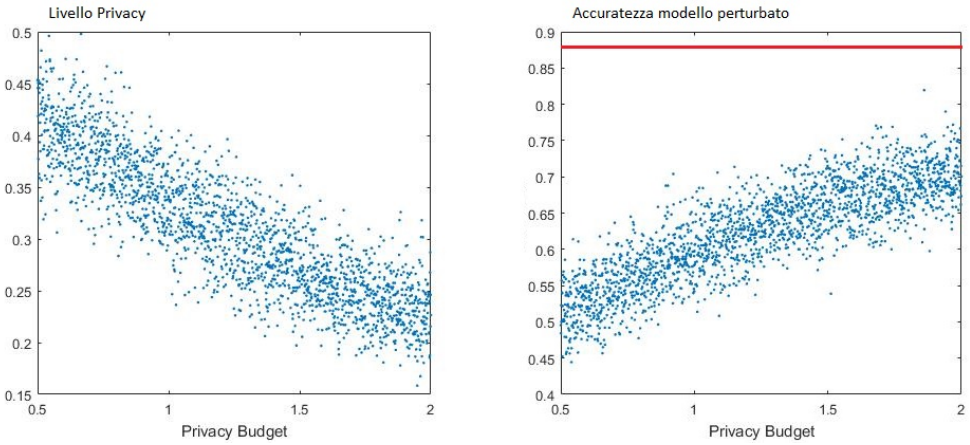


Figura 4.14: Livello di Privacy e Accuratezza del modello

4.8 Conclusioni

Nell'analisi dei risultati ottenuti, consideriamo le seguenti funzioni che indicano il livello di utilità e privacy, al variare di ϵ:

$U_1(\epsilon)$=*Rumore medio aggiunto alla variabile reddito annuo nel Training Set.*
$U_2(\epsilon)$=*Percentuale di volte in cui il cui reddito annuo supera – o scende al di sotto de - i* $50k$ *a causa del rumore aggiunto.*

Le due misure U_1 e U_2 sono un indicatore di utilità "inversa", ovvero maggiore è il valore di U_1 e U_2, minore sarà l'utilità della variabile reddito annuo nel dataset di partenza.
Come misura del livello di privacy:

Error Rate(ϵ)=1− Accuratezza(ϵ).

Maggiore è l'errore del modello perturbato, maggiore sarà il livello di privacy ottenuto, grazie all'errata classificazione.

Come si evince nella figura 4.13, all'aumentare di ϵ, l'errore di classificazione del modello perturbato diminuisce significativamente. Con valore di *epsilon* > 2, il modello perturbato si avvicina di molto al modello costruito in assenza di rumore (Test error 0.17, linea in rosso in figura 4.14). Ciò implica che, per circa l'83 percento degli individui presenti nel test Set, la privacy relativa alla variabile reddito $> 50k$, sarebbe violata. Al contrario, un valore di ϵ vicino a 0.5 produce un errore di circa 0.5, che rappresenta un ottimo risultato in termini di privacy (circa la metà delle classificazioni totali sarebbero errate-Figura 4.13). Tuttavia la quantità di rumore medio aggiunta alla variabile reddito annuo sarebbe molto elevata. (circa $70k$). Al fine di conservare l'integrità dell'informazione, occorre scegliere il privacy budget ϵ in maniera tale da avere un giusto compromesso tra privacy e utilità.
Considerando il grafico di U_2 in figura 4.13, è possibile calcolare l' *elbow/knee*

point, ovvero il punto in cui la curva presenta la maggiore inclinazione. Il valore esatto, calcolato tramite un'apposita funzione, è: $\epsilon = 0.98$ che corrisponde ad un utilità di $25k$ (reddito annuo aggiunto). Da quel valore di ϵ in poi il decremento del rumore aggiunto non è così significativo da giustificare la scelta di un ϵ maggiore. Il livello di privacy corrispondente sarebbe pari a 0.35. Tuttavia se si vuole ottenere un livello di privacy più elevato, occorre aumentare la quantità di rumore, diminuendo quindi la *privacy budget*.
Il valore di $\epsilon \approx 0.9$ sembra essere una buona scelta in tal senso: l'errore relativo alla classificazione sugli individui del Test Set è comunque elevato (circa 0.42), con un incremento di 0.25 rispetto al modello non perturbato. Parallelamente, valutando la funzione d'utilità 2 nel valore $\epsilon = 0.9$, si ottiene una percentuale del 20 percento sul totale di individui nel Training Set.
La scelta del valore appropriato del parametro di *privacy budget* ϵ è ancora oggi oggetto di studio e analisi. Non esiste infatti una regola valida in generale, ma occorre valutare ogni applicazione della *Differential Privacy*, in modo che vengano sempre rispettate e tutelate le informazioni sensibili. La tecnica utilizzata in quest'applicazione vuole essere un'alternativa alla soluzione dell' offuscamento del dato sensibile: invece di perdere completamente l'informazione contenuta in variabile, l'aggiunta di rumore dalla distribuzione di Laplace permette di trovare un giusto compromesso tra privacy ed utilità, rendendo il dato ancora utilizzabile a fini statistici.

Capitolo 5

Conclusioni e sviluppi futuri

In questa tesi sono stati presentati gli impatti della normativa GDPR sul trattamento dei dati personali. L'entrata in vigore, nel Maggio 2018, di questa normativa rappresenta un punto di svolta nella gestione dei dati, soprattutto per le organizzazioni che ne posseggono grandi moli. Tale cambiamento risulta essere un'opportunità per per innovare le procedure della libera circolazione delle informazioni sensibili, nei contesti dei *Big Data* e del Machine Learning. Durante la mia esperienza di tesi, accompagnata da uno *stage* in azienda, ho avuto modo di conoscere e approfondire tutte le fasi procedurali di adempimento alla normativa. Inoltre, ho analizzato alcune delle tecniche utilizzate in ambito di anonimizzazione e pseudonimizzazione, e ho tentato di applicarle alla realtà odierna. Mi sono dedicato, infine, allo studio della teoria della *Differential Privacy*. Essa può essere considerata, con le dovute precauzioni, un importante compromesso tra la protezione obbligatoria dei dati sensibili e l'altrettanto fondamentale utilizzo di informazioni a fini statistici. L'ingresso recente della normativa lascia spazio allo sviluppo di nuove tecniche che permettano di sfruttare al massimo le informazioni contenute all'interno di determinati dati senza infrangere la legge, e nel rispetto della privacy degli individui.

Bibliografia

[1] Tianqing Zhu, Gang Li, Wanlei Zhou, Philip S. Yu *Differential Privacy and Apllications*, Advances in Information Security, vol.69, Springer (2017).

[2] Cynthia Dwork, Aron Roth, *The Algorithmic Foundations of Differential Privacy*, Foundations and Trends in Theoretical Computer Science, vol.9,Now 2014.

[3] Giuseppe D'Acquisto, Maurizio Naldi *Big Data and Privacy by Design*, I diritti nella "rete" della rete, Anonimizzazione Pseudonimizzazione Sicurezza, Prefazione di Franco Pizzetti, G.Giappichelli Editore (2017).

[4] Latanya Sweeney *K-Anonimity: a model for protecting privacy*, International Journal on Uncertainty, Fuzzies and Knowledge-based System, 10(5),(2002);557-570

[5] Gazzetta ufficiale dell'Unione europea *REGOLAMENTO (UE) 2016/679 DEL PARLAMENTO EUROPEO E DEL CONSIGLIO del 27 aprile 2016*, L 119/1

[6] Rosa Meo, *Privacy preserving data mining: una promessa e una sfida*, Forum Machine Learning, Applicazioni ed effetti sulla società, Campus Luigi Einaudi, Torino, 17-18 Novembre 2017

[7] Wikipedia, https://it.wikipedia.org/wiki/Regolamento_generale_sulla _protezione _dei_dati
(ultimo accesso 30/06/2018)

[8] http://www.rgigroup.it/overview
(ultimo accesso 30/05/2018)

[9] https://en.wikipedia.org/wiki/K-anonymity
(ultimo accesso 20/07/2018)

[10] https://archive.ics.uci.edu/ml/index.php
(ultimo accesso 20/09/2018)

[11] https://www.mathworks.com/matlabcentral/mlc-downloads
/downloads/submissions/47407/versions/2/previews/DMPSK
_over_Rayleigh/functions/laprnd.m/index.html?access_key
(ultimo accesso 20/09/2018)