# Progetto Buzzi e lo screening neonatale esteso (1 min)

Progetto nato all’interno del contesto dell’ospedale Buzzi, parte dell’ASST (Azienda Socio Sanitaria Territoriale) Fatebenefratelli-Sacco.

Obiettivo: analizzare uno dei più ampi e significativi database in Europa relativi allo screening neonatale esteso, comprendente i dati clinici e demografici di tutti i bambini nati in Regione Lombardia a partire da Giugno 2012 fino ad arrivare ad Aprile 2022. Progetto per l’analisi dei dati per individuare correlazioni, patterns o signatures all’interno dei dati raccolti, e prevedere la comparsa di malattie metaboliche nella popolazione pediatrica lombarda.

Screening neonatale esteso: test per l’individuazione preventiva di patologie e condizioni cliniche nel neonato. Screening attraverso il prelievo di poche gocce di sangue dal tallone del neonato, effettuato tra le 48 e le 72 ore di vita; campioni analizzati nel Laboratorio specializzato per lo Screening, e risultati inviati all’Ospedale di nascita.

# Le malattie metaboliche (30 sec)

- Ipotiroidismo congenito (IC): condizione permanente, causata da fattori neonatali, con gravi effetti sul sistema nervoso centrale; incidenza di un caso ogni 3000 nati; necessità di precoce terapia ormonale.   
- Fibrosi cistica (mucoviscidosi): malattia ereditata dai genitori, dovuta alla mutazione del gene CTFR; altera secrezioni di organi come polmoni e pancreas, causando danni gravi; necessità di intervenire con farmaci dalla nascita per limitarne i danni.   
- Fenilchetonuria (PKU): raro difetto metabolico, porta a gravi disturbi sul sistema nervoso, fino a disabilità cognitive se non individuata e trattata; circa 50000 casi in tutto il mondo; per contrastarla, necessario rigoroso regime alimentare a basso contenuto proteico.   
- Iperplasia surrenalica congenita: gruppo di rare malattie genetiche, provocate da insufficienza di cortisolo o aldosterone; circa un caso ogni 16000 nati; trattamento farmacologico e chirurgico previsti.   
- Atrofia Muscolare Spinale (SMA): rara malattia neuromuscolare, causata da perdita di motoneuroni; provoca debolezza e atrofia muscolare progressiva; circa un caso ogni 10000 nati.

# Il dataset (1.30min)

4 fasi di raccolta dei dati:   
1. Raccolta ed “accettazione” (inserimento) dei dati nel sistema   
2. Invio dei cartoncini al Buzzi   
3. Analisi dei cartoncini al Buzzi   
4. Analisi dei risultati e individuazione dei positivi

Dataset completo formato da 985792 records, raccolti da circa metà 2012 fino all’Aprile 2022, con un totale di 266 variabili. In questo progetto analizzato un campione del dataset completo (30% circa di osservazioni totali, ovvero 295738 records), con 109 delle 266 features originali così suddivise: 76 variabili quantitative (analiti, peso ed età gestazionale), 28 variabili qualitative, 3 variabili in formato data e 2 variabili necessarie per riconoscimento dei pazienti e dei campioni.

# Obiettivi e fasi del progetto (1 min)

1. Fase di import e pulizia dei dati;
2. Fase di analisi esplorativa;
3. Fase di analisi esplorativa stratificata per la variabile “Reparto”;
4. Fase di applicazione delle tecniche di riduzione della dimensionalità;
5. Fase di applicazione delle tecniche di cluster analysis;
6. Fase di applicazione delle tecniche di cluster analysis ai dati ridotti con applicazione di tecniche di riduzione della dimensionalità;
7. Valutazione risultati ottenuti.

# Pulizia dati ed analisi esplorativa (1min)

Tutti i passaggi sono svolti col linguaggio di programmazione Python, tramite utilizzo di metodi e librerie utili per l’analisi di dati, la creazione di infografiche e il calcolo di indici statistici.

Statistiche descrittive variabili quantitative, con istogrammi collegati; distribuzioni variabili qualitative, con barplot; analisi correlazioni (con heatmap e VIF, senza escludere variabili su richiesta dei clinici esperti di dominio); inferenza; esplorazioni ulteriori (istogrammi distribuiti per variabili qualitative; barplot stratificati per qualitative; scatterplot stratificati…). Fase utile per scoprire le caratteristiche dei dati presenti all’interno del dataset di interesse, ed utili in vista dell’implementazione delle fasi successive.

# Analisi esplorativa stratificata per reparto (30 sec)

Ripetizione di molti dei passaggi della prima fase di analisi esplorativa sul dataset stratificato per la variabile Reparto (indica il reparto in cui è ricoverato il neonato al momento dello screening, ovvero ‘generico’, ‘neo-patologico’ o ‘nido’). Importante questo perché indica generalmente complicazioni o meno in gravidanza o parto, dunque avere informazioni sulle distribuzioni di alcune variabili in base al reparto di appartenenza.

# Riduzione di dimensionalità (1.30 minuto)

Metodi di machine learning nati con l’obiettivo di ridurre il numero di variabili in un dataset di interesse, con minima perdita di informazione (variabilità).

Applicata in diversi ambiti (text mining, finanza, elaborazione di immagini, biostatistica), e spesso usata come operazione preliminare ad altre tecniche quali cluster analysis, modelli di classificazione e modelli di regressione.

Fondamentali i valori dei parametri in ogni metodo applicato (scelta dei valori sempre delicata, si tratta di metodi di apprendimento non supervisionato, senza particolari metriche di valutazione della bontà dei risultati ottenuti).

Tra le principali tecniche applicate in questo progetto: PCA, t-SNE ed UMAP.

Applicazioni della riduzione di dimensionalità su dati puliti e resi adatti ai metodi seguenti (quindi un solo record per neonato, esclusione features con tanti valori mancanti, dati normalizzati x, create le dummies per dati qualitativi; totale di 272114 osservazioni con 69 variabili, 25 qualitative e 44 quantitative).

# UMAP (1.30 min)

Tecnica recente (2018 circa), eccelle in catturare strutture dati non lineari, molto utilizzata in diversi ambiti (tra cui biomedico).

Parametri fondamentali: n\_neighbors (quanto UMAP mantiene della struttura locale), min\_dist (quanto distano i punti in rappresentazioni a basse dimensioni), n\_components (dimensionalità dello spazio risultante dall’algoritmo) e metrics (come viene calcolata la distanza tra punti).

Molto efficiente per grandi insiemi di dati, preserva strutture locali, ma molto sensibile ai cambi di valore dei parametri, i risultati variano parecchio in base al seed (quindi risultati diversi in base alla Random Initialization) e di non semplice interpretazione.

A sinistra neigh 5 e min\_d 0.05; centro neigh 5 e min\_d 0.01 (x min\_d bassa, cluster più fitto con poche singole osservazioni distanti); destra neigh 50 e min\_d 0.1 (x neigh alti, gruppi più compatti)

# PCA (1.30 min)

Tecnica storica di riduzione dimensionalità, usata in svariati ambiti, proietta i dati in nuovo sistema di coordinate (componenti principali) ordinate per quantità di varianza che spiegano

Diverse implementazioni con valori differenti per i parametri whiten, svd\_solver, tol, iterated\_power, n\_oversamples e power\_iteration\_normalizer, senza particolari variazioni nei risultati.

Cattura le variazioni più significative nei dati, e permette di ottenere componenti linearmente dipendenti (utile per interpretazione e modellazioni successive), ma è sensibile agli outliers, ha prestazioni molto negative per distribuzioni non gaussiane, e porta alla perdita di molte informazioni, soprattutto per spazi finali di dimensioni ridotti.

Grosso cluster molto denso e compatto, con alcune osservazioni molto separate.

# t-SNE (1.30 min)

Tecnica di riduzione dimensionalità, usata in svariati ambiti, eccelle nella cattura di relazioni complesse e non lineari.

Parametri fondamentali: perplexity (equilibrio tra conservazione di strutture globali e locali), early\_exageration (quanto i campioni nei clusters vengono riprodotti vicini nello spazio a dimensionalità ridotta) ed n\_components (dimensionalità dello spazio risultante dall’algoritmo).

È efficace nel catturare relazioni non lineari, robusta rispetto a variazioni globali e preserva strutture locali, ma è molto sensibile agli iperparametri, computazionalmente pesante e sensibile all’inizializzazione casuale.

Poche variazioni coi parametri: grossa nuvola di dati con pochissime variazioni che sono distanti dalla nuvola (meno distanti con perplexity alta, figura al centro).

# Altri metodi non utilizzati

Altre tecniche di riduzione della dimensionalità escluse dai risultati finali:   
- Kernel PCA: non implementata poiché prevedeva la creazione di uno spazio a dimensionalità superiore rispetto a quello in input (con creazione di una matrice di 600GB);   
- ICA (Independent Component Analysis), SVD (Singular Value Decomposition) e NMF (Non-Negative Matrix Factorization): impossibili da eseguire richiedendo la decomposizione della matrice di dati in input (dimensioni vicine ai 300GB).

Problema privacy dati.