Analisi Esplorativa

1. Importazione di un dataset di lavoro

Contesto

Secondo l'Organizzazione Mondiale della Sanità (OMS), l'ictus è la seconda causa principale di morte a livello globale, responsabile di circa il 11% del totale dei decessi.

Questo dataset è utilizzato per prevedere se un paziente è probabile che subisca un ictus in base ai parametri di input come sesso, età, varie malattie e abitudine al fumo.

Fonte

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/

2. Comprensione del quadro generale

Obiettivo: capire il problema che vogliamo andare a risolvere, andando a ragionare sull'intero dataset e sul significato delle variabili.

Struttura del dat aset

1. **id:**
    - Tipo: Integer
    - Contesto: Identificazione di un singolo paziente
    - Aspettativa: Bassa
    - Commenti: utile solo in ottica di identificazione di un pazienza
2. **gender:** "Male", "Female" or "Other"
    - Tipo: Object
    - Contesto: Distinzione di genere nei pazienti
    - Aspettativa: Media
    - Commenti: Potrebbe essere un fattore rilevante, ma non ci si aspetta di poter trarre conclusioni molto differenti in base al genere del paziente.
3. **age:** age of the patient
    - Tipo: Float
    - Aspettativa: Media
    - Commenti: Ci si aspetta che la probabilità di ictus incrementi con l'alzarsi della fascia d'età del paziente
4. **hypertension:** 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
    - Tipo: Integer
    - Contesto: L'ipertensione arteriosa è una condizione caratterizzata dall'elevata pressione del sangue nelle arterie, che è determinata dalla quantità di sangue che viene pompata dal cuore e

dalla resistenza delle arterie al flusso del sangue.
    - Aspettativa: Alta
    - Commenti: Ci si aspetta che che l'ipertensione sia uno dei fattori di rischio più importanti

5. **heart_disease:** 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease

    - Tipo: Boolean
    - Contesto: Presenza di malattie cardiache che potrebbero influire
    - Aspettativa: Alta
    - Commenti: Ci si aspetta che la presenza di malattie cardiache possa essere un fattore importante

6. **ever_married:** "No" or "Yes"

    - Tipo: Object
    - Aspettativa: Bassa

7. **work_type:** "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"

    - Tipo: Object
    - Aspettativa: Media
    - Commenti: Il tipo di lavoro del paziente potrebbe influire sui dati, ci si potrebbe aspettare che ad esempio fattori come lo stress che dipendono dal lavoro possano influire.

8. **Residence_type:** "Rural" or "Urban"

    - Tipo: Object
    - Aspettativa: Bassa
    - Commenti: Non ci si aspetta un grande impatto di questa variabile nei dati

9. **avg_glucose_level:** average glucose level in blood

    - Tipo: Float
    - Contesto: Secondo la seguente fonte https://www.grupposandonato.it/news/2020/giugno/glicemia-alta: "La glicemia alta, o iperglicemia, è una condizione da non sottovalutare. Il valore normale della glicemia a digiuno viene mantenuto tra 70 e 100mg/dl. Si parla di iperglicemia quando vi è un innalzamento del glucosio nel sangue che può portare a conseguenze negative sul nostro stato di salute."
    - Aspettativa: Alta
    - Commenti: Ci si aspetta che il livello di glucosio nel sangue possa considerarsi un fattore di grande importanza nel contesto della ricerca.

10. **bmi:** body mass index

- Tipo: Float
- Contesto: Secondo la seguente fonte https://www.salute.gov.it/portale/nutrizione/dettaglioIMCNutrizione.jsp?lingua=italiano&id=5479&area=nutrizione&menu=vuoto: "L'IMC (BMI) è l'indicatore di riferimento per studi epidemiologici e di screening di obesità. E' utile sottolineare che l'IMC in quanto indicatore di studi di popolazione, non è in grado di valutare la reale composizione corporea, così come non permette di conoscere la distribuzione del grasso corporeo nell'individuo." Secondo la fonte indicata, si possono distinguere le seguenti categorie:

    - Grave magrezza: < 16,00
    - Sottopeso: 16,00 - 18,49
    - Normopeso: 18,50 - 24,99
    - Sovrappeso: 25,00-29,99
    - Obeso classe 1: 30,00-34,99
    - Obeso classe 2: 35,00-39,99
    - Obeso classe 3: ≥ 40,00" Inoltre, indici troppo alti o troppo bassi potrebbero indicare la presenza di patologie.
- Aspettativa: Media
- Commenti: Potrebbe essere un indice importante e che potrebbere avere un impatto significativo nei dati.

- importante

```
verificare che l'indice sia calcolato nello stesso modo della fonte modotale da
poter sfruttare correttamente la classificazione stilata
```

11. **smoking_status:** "formerly smoked", "never smoked", "smokes" or "Unknown"*

- Tipo: Object
- Aspettativa: Alta
- Commenti: Ci si aspetta che il fumo possa essere un fattore importante nel contesto della ricerca.

12. **stroke:** 1 if the patient had a stroke or 0 if not *Note: "Unknown" in smoking_status means that the information is unavailable for this patient

- Tipo: Boolean
- Aspettativa: Alta
- Commenti: La ricorrenza di ictus ci si aspetta sia un fattore di rischio importante.