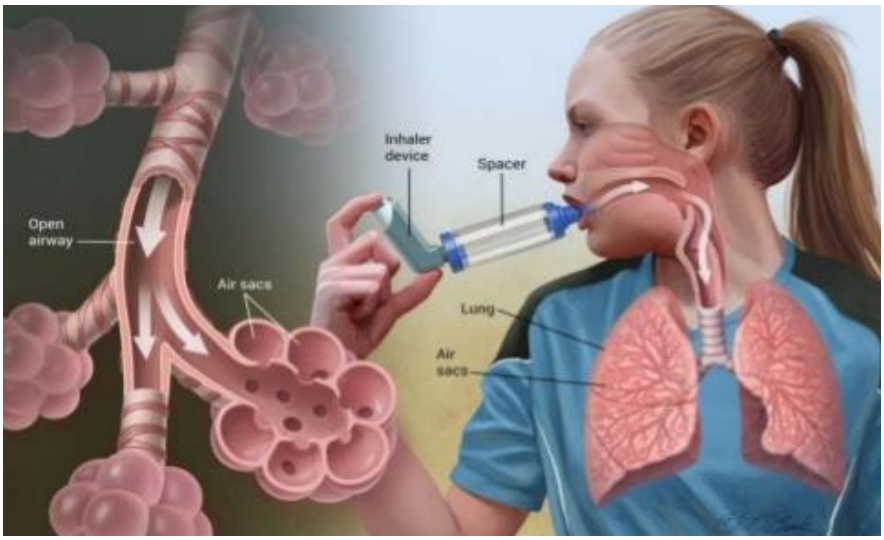
# Università della Calabria

**Dipartimento di Economia, Statistica e Finanza "*Giovanni Anania*"**

Corso di Laurea Magistrale in Statistica e Informatica per le Decisioni e le Analisi di Mercato


# *Reti bayesiane per lo studio dell'asma bronchiale in Italia*





***Docente***

*Prof. Paolo Carmelo Cozzucoli*

***Studenti***

*Tucci Marco Matr. 214216*

*Ruffo Pietro Matr. 214241*



Anno Accademico 2020/2021

# INTRODUZIONE

L'asma bronchiale è una malattia cronica delle vie aeree caratterizzata da ostruzione bronchiale più o meno accessionale, solitamente reversibile spontaneamente o in seguito alla terapia; si associa ad ampia variabilità nel tempo della funzione polmonare, di solito concordante con l'andamento dei sintomi; provoca iperreattività bronchiale e un accelerato declino della funzionalità respiratoria che può evolvere, in alcuni casi, in una ostruzione irreversibile delle vie aeree. L'asma è un problema mondiale e un consistente onere sociale ed economico per i sistemi sanitari. Secondo l'Organizzazione mondiale della sanità, ci sono tra i 100 e i 150 milioni di persone che soffrono di questa condizione in tutto il mondo.
Le morti associate alla malattia, sempre secondo i dati dell'Oms, sono circa 180mila ogni anno.
In Italia, si stima che ogni anno circa nove milioni di persone si ammalano di allergie respiratorie derivanti dalla presenza di pollini nell'aria e quattro milioni di essi ricorrono a cure. Le persone, ma soprattutto i bambini, che vivono in aree urbane la cui aria è inquinata sono più soggetti a sviluppare forme di asma o a vedere peggiorare la malattia.

Lo scopo dell'analisi è quello di costruire una rete bayesiana multinomiale per rappresentare le relazioni di dipendenza tra le variabili che influenzano l'insorgere di questa malattia in Italia. Si propone l'utilizzo di tale strumento metodologico sia perché il problema da trattare è complesso sia perché si vuole ragionare in termini probabilistici sul dominio del problema stesso.

Nell'ambito delle reti bayesiane, usualmente si hanno due possibili approcci: "expert system" e addestramento della rete tramite algoritmi. Nel progetto i due metodi verranno utilizzati in completa sinergia tra loro.

Si procederà dapprima a considerare la componente qualitativa della rete (il DAG – Directed Acyclic Graph – che meglio si adatta ai dati), per cui si costruirà sia un grafo "expert system" sulla base di conoscenze teoriche del problema da risolvere sia diversi grafi tramite l'algoritmo di apprendimento automatico Hill Climbing. Dopodiché si passerà ad individuare la componente quantitativa della rete (stima delle CPTs). Infine, si utilizzerà il modello di rete bayesiana costruito per fare inferenza probabilistica.

# DESCRIZIONE DEL DATASET

Il software utilizzato per condurre l'analisi è "RStudio". Le librerie adoperate sono le seguenti: "bnlearn", "lattice", "BiocManager", "gridExtra", "gRain", "Rgraphviz" e "openxlsx".

Il set di dati a disposizione proviene dall' "Indagine europea sulla salute" condotta in Italia nel 2015. Esso è costituito da 2755 osservazioni su 9 variabili. Le variabili che costituiranno il dominio del problema sono:

- Sex (sesso), con modalità "male" o "female";
- Age (età), con modalità "young", "adult" e "old";
- Urbanization (livello di urbanizzazione), con modalità "low", "medium" e "high";
- Education (livello di istruzione), con modalità "high" e "low";
- Geographic area (area geografica), con modalità "north", "centre" e "south/islands";
- Allergy (allergie), con modalità "yes" e "no";
- Smoke (fumo), con modalità "yes" e "no";
- Sedentary (vita sedentaria), con modalità "yes" e "no";
- Asthma (asma), con modalità "yes" e "no".

```
> head(data)
     sex   age urbanization education geographic_area allergy smoke
1   male adult          low       low   south/islands     yes   yes
2 female   old          low       low   south/islands     yes    no
3 female adult         high      high          centre      no    no
4   male adult       medium       low   south/islands     yes    no
5 female adult          low      high           north      no    no
6 female adult       medium      high           north      no   yes
  sedentary asthma
1       yes    yes
2       yes    yes
3       yes    yes
4        no     no
5        no     no
6        no    yes

> dim(data)
[1] 2755    9
```

Tra le variabili si assume una struttura di dipendenza di tipo probabilistico.

Le variabili sono state trasformate tutte in "factor" e rinominate con lo stesso nome che verrà dato al corrispondente nodo nella rete bayesiana.

```
> str(data)
'data.frame':  2755 obs. of  9 variables:
 $ sex            : Factor w/ 2 levels "female","male": 2 1 1 2 1 1 1 2 2 1 ...
 $ age            : Factor w/ 3 levels "adult","old",..: 1 2 1 1 1 1 2 2 1 2 ..
.
 $ urbanization   : Factor w/ 3 levels "high","low","medium": 2 2 1 3 2 3 3 2 3
2 ...
 $ education      : Factor w/ 2 levels "high","low": 2 2 1 2 1 1 2 2 2 2 ...
```

```
 $ geographic_area: Factor w/ 3 levels "centre","north",..: 3 3 1 3 2 2 2 2 1 2
...
 $ allergy        : Factor w/ 2 levels "no","yes": 2 2 1 2 1 1 1 1 1 2 ...
 $ smoke          : Factor w/ 2 levels "no","yes": 2 1 1 1 1 2 1 2 1 1 ...
 $ sedentary      : Factor w/ 2 levels "no","yes": 2 2 2 1 1 1 2 2 2 2 ...
 $ asthma         : Factor w/ 2 levels "no","yes": 2 2 2 1 1 2 1 1 1 1 ...

> names(data)<-c("SEX","AGE","URB","EDU","GEO","ALG","SMK","SED","ASTHMA")
> names(data)
[1] "SEX"    "AGE"    "URB"    "EDU"    "GEO"    "ALG"    "SMK"
[8] "SED"    "ASTHMA"
```

Il dataset è risultato essere bilanciato rispetto ai valori della variabile risposta. Non si riscontrano missing values.

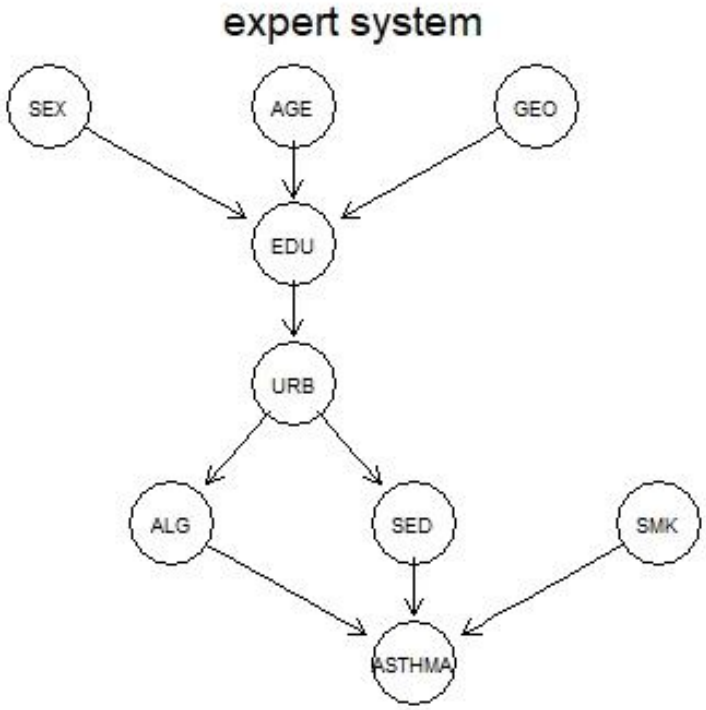
```
> table(data$ASTHMA)

  no  yes
1573 1182

> table(is.na.data.frame(data))

FALSE
24795
```

# COSTRUZIONE DELLA RETE BAYESIANA MULTINOMIALE

## 1. RAPPRESENTAZIONE GRAFICA DELLA RETE

### EXPERT SYSTEM

Sulla base delle conoscenze pregresse sulle relazioni di dipendenza e indipendenza tra le variabili del dominio (cioè sul problema da risolvere), si costruisce il seguente modello grafico per la rete:




Il sesso, l'età e l'area geografica di appartenenza sono caratteristiche dei soggetti sulle quali non può esserci influenza di alcun fattore. Le tre variabili sopracitate possono avere però un'influenza importante sul livello di istruzione del soggetto che a sua volta andrà ad influenzare i risvolti professionali dello stesso e conseguentemente la sua futura residenza (considerata anche l'elevata migrazione da Sud a Nord di molti studenti e lavoratori). Per tali ragioni la variabile relativa al livello di studio influenza il livello di urbanizzazione relativo alla residenza del soggetto. Da questa variabile può dipendere a sua volta la sedentarietà del soggetto: chi vive in aree densamente popolate potrebbe avere meno tempo libero da dedicare all'attività fisica a causa del tempo speso nel traffico per raggiungere il luogo di studio/lavoro, in aggiunta potrebbe anche influire la scarsità/lontananza delle aree a ciò dedicate.
Inoltre, vivere in un'area altamente urbanizzata sembrerebbe aumentare il rischio di

soffrire di varie forme di allergia.
Ultimo fattore preso in considerazione, completamente indipendente dagli altri, è la dipendenza da tabacco che è un fenomeno che riguarda l'intera popolazione a prescindere da sesso, residenza, livello di istruzione ecc.
Dipendenza da tabacco, sedentarietà e allergie sono fattori che, secondo la scienza, incidono fortemente sulla possibilità di soffrire di asma bronchiale.

In definitiva, la probabilità congiunta di avere l'asma viene fattorizzata secondo l'approccio "expert system" nel seguente modo:

$$P[SEX]\,P[AGE]\,P[GEO]\,P[SMK]\,[EDU|SEX:AGE:GEO]\,P[URB|EDU]\,P[ALG|URB]\,[SED|URB]\,P[ASTHMA|ALG:SMK:SED]$$

```
> modelstring(dag)
[1] "[SEX][AGE][GEO][SMK][EDU|SEX:AGE:GEO][URB|EDU][ALG|URB][SED|URB][ASTHMA|AL
G:SMK:SED]"
```

Il grafo contiene 9 nodi e 9 archi diretti, che determinano le relazioni di dipendenza tra le variabili.

```
> nodes(dag)
[1] "SEX"    "AGE"    "URB"    "EDU"    "GEO"    "ALG"    "SMK"    "SED"
[9] "ASTHMA"
> arcs(dag)
      from  to
 [1,] "SEX" "EDU"
 [2,] "AGE" "EDU"
 [3,] "GEO" "EDU"
 [4,] "EDU" "URB"
 [5,] "URB" "ALG"
 [6,] "URB" "SED"
 [7,] "SMK" "ASTHMA"
 [8,] "ALG" "ASTHMA"
 [9,] "SED" "ASTHMA"
```

A questo punto, si valuta il DAG costruito (e le dipendenze in essa tra le variabili) nella sua interezza, considerando la bontà di adattamento dell'intera rete tramite statistiche campionarie. Gli scores registrati sono:

- **AIC** (Akaike Information criterion) = -18106.08
- **BIC** (Bayesian Information criterion) = -18230.42
- **BDen** (Bayesian Dirichlet equivalent uniform) = -18197.26

Il valore degli scores risulta molto basso e ciò indica un buon livello di adattamento dei dati alla struttura di dipendenze rappresentata nel DAG.

## ADDESTRAMENTO AUTOMATICO CON L'ALGORITMO HILL CLIMBING

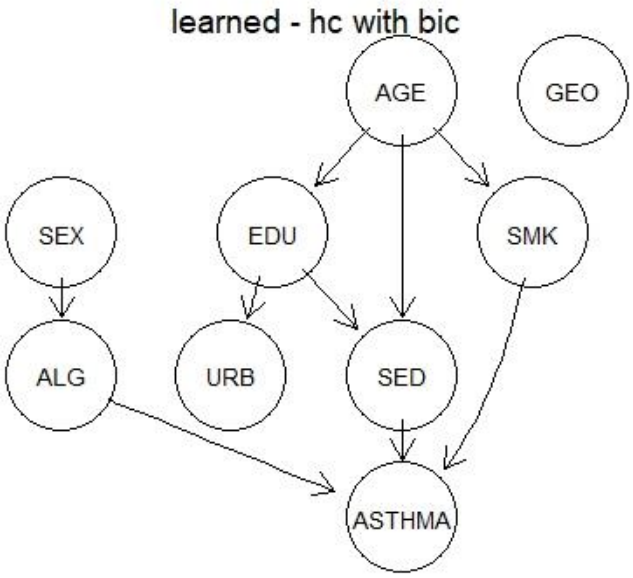
Il passo successivo è quello di utilizzare l'algoritmo Hill Climbing allo scopo di addestrare automaticamente la rete in modo da identificare fra tutti i possibili grafi (date le variabili del dominio) quello che si adatta meglio ai dati. In particolare, si considera un sottoinsieme dei possibili grafi, determinato dall'utilizzo di una "white list" e di una "black list" (rappresentanti le conoscenze teoriche di maggiore rilievo da tenere necessariamente in considerazione), che saranno funzionali all'addestramento automatico della rete.

```
> black_list<-matrix(c("EDU","ASTHMA",
+                      "ASTHMA","EDU",
+                      "ASTHMA","URB",
+                      "ASTHMA","SED",
+                      "ASTHMA","ALG",
+                      "ASTHMA","SMK",
+                      "SMK","SEX",
+                      "AGE","SEX",
+                      "GEO","SEX",
+                      "ALG","SEX",
+                      "URB","SEX",
+                      "EDU","SEX",
+                      "SED","SEX",
+                      "ASTHMA","SEX",
+                      "SMK","EDU",
+                      "ALG","EDU",
+                      "URB","EDU",
+                      "SED","EDU",
+                      "ASTHMA","EDU",
+                      "SMK","GEO",
+                      "AGE","GEO",
+                      "EDU","GEO",
+                      "ALG","GEO",
+                      "URB","GEO",
+                      "SEX","GEO",
+                      "SED","GEO",
+                      "ASTHMA","GEO",
+                      "SMK","AGE",
+                      "GEO","AGE",
+                      "EDU","AGE",
+                      "ALG","AGE",
+                      "URB","AGE",
+                      "SEX","AGE",
+                      "SED","AGE",
+                      "ASTHMA","AGE",
+                      "AGE","ALG",
+                      "EDU","ALG"),
+                    byrow=TRUE,ncol=2,
+                    dimnames=list(NULL,c("from","to")))

> white_list<-matrix(c("ALG","ASTHMA",
+                      "SMK","ASTHMA",
+                      "SED","ASTHMA"),
+                    byrow=TRUE,ncol=2,
+                    dimnames=list(NULL,c("from","to")))
```

La black list contiene le connessioni illogiche da non considerare nella costruzione del modello grafico, mentre la white list comprende quelle connessioni logiche che devono necessariamente presenti all'interno del DAG.

*BIC*

Con il criterio *BIC* si ottiene il seguente grafo:




```
> learned<-hc(data,score="bic",whitelist=white_list,blacklist=black_list)
```

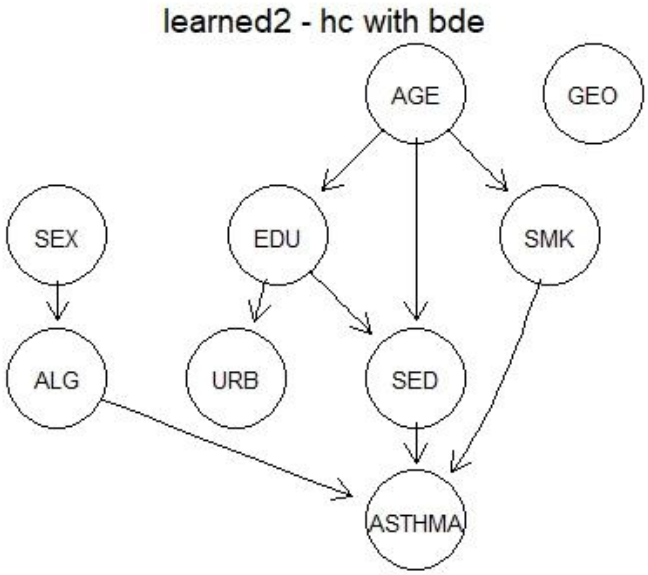
Gli archi presenti nella rete sono 9, così come i nodi. Di seguito, si riporta la fattorizzazione della probabilità congiunta ottenuta con tale criterio:

$$P[SEX]P[AGE]P[GEO]P[EDU|AGE]P[ALG|SEX]P[SMK|AGE]P[URB|EDU]P[SED|AGE:EDU]P[ASTHMA|ALG:SMK:SED]$$

In termini di bontà di adattamento della rete ai dati, si registra il seguente valore di score: -18032.05 (contro il -18230.42 registrato nel caso dell'Expert System).

*BDE*

Utilizzando il criterio *Bayesian Dirichlet equivalent uniform* si ottiene il seguente grafo:




```
> learned2<-hc(data,score="bde",whitelist=white_list,blacklist=black_list)
```

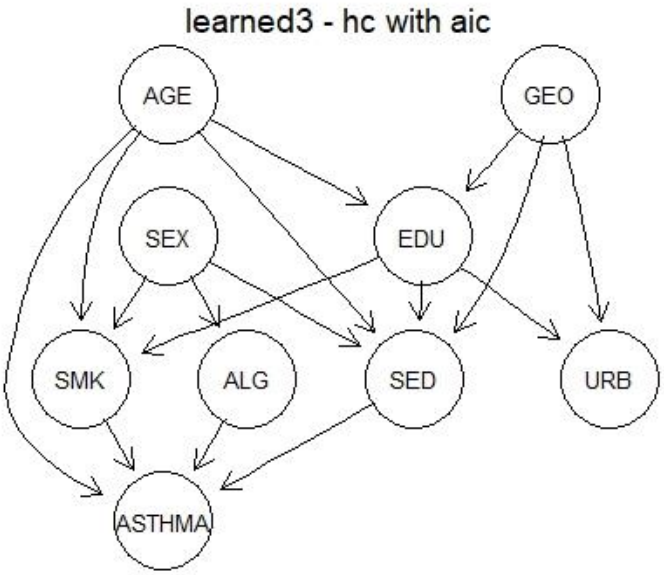
Anche qui il grafo presenta 9 nodi e 9 archi diretti. La probabilità congiunta viene in questo caso fattorizzata nel seguente modo:

$$P[SEX]P[AGE]P[GEO]P[EDU|AGE]P[ALG|SEX]P[SMK|AGE]P[URB|EDU]P[SED|AGE:EDU]P[ASTHMA|ALG:SMK:SED]$$

Lo score BDE registrato è pari a -18032.05 (contro il -18197.26 nel caso dell'Expert System).

*AIC*

Infine, utilizzando il criterio *AIC* si ottiene la seguente rete:




```
> learned3<-hc(data,score="aic",whitelist=white_list,blacklist=black_list)
```

Questa rete presenta 9 nodi e ben 16 archi diretti. La probabilità congiunta di soffrire d'asma è fattorizzata come segue:

$$P[SEX]P[AGE]P[GEO]P[EDU|AGE:GEO]P[ALG|SEX][URB|EDU:GEO]P[SMK|SEX:AGE:EDU]P[SED|SEX:AGE:EDU:GEO]P[ASTHMA|AGE:ALG:SMK:SED]$$

In termini di adattamento della rete, si registra il seguente valore di score: -17850.03 (contro il -18106.08 nel caso dell'Expert System).

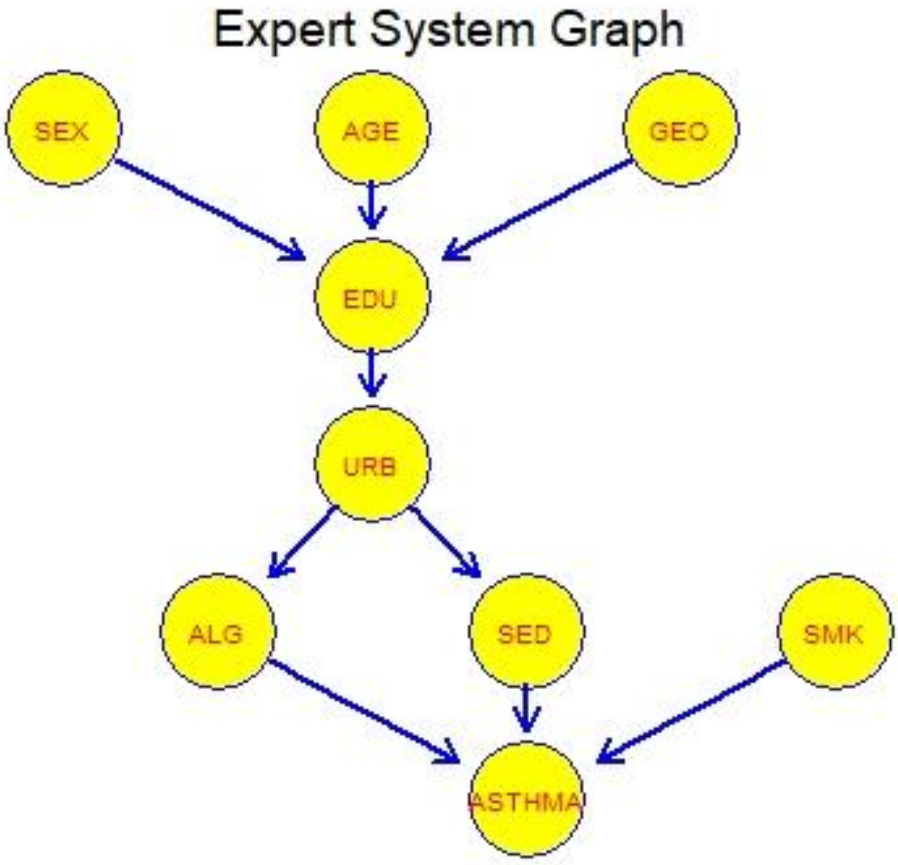
**CONFRONTO TRA I MODELLI GRAFICI**

La tabella che segue mette a confronto i valori dei tre criteri statistici (BIC, AIC e BDE) ottenuti nei due approcci considerati ("expert system" e apprendimento tramite algoritmo HC).

| | *EXPERT SYSTEM* | *HC* |
|---|---|---|
| *BIC* | **-18230.42** | -18019.94 |
| *AIC* | **-18106.08** | -17928.16 |
| *BDE* | **-18197.26** | -18032.05 |

In tutti e tre i casi, la rete basata sull'approccio "expert system" registra degli scores inferiori a quelli delle reti ottenute tramite l'algoritmo HC, e ciò significa che gode di un maggior adattamento ai dati, per cui verrà utilizzata per fare inferenza sulle probabilità.

Si riporta il DAG di riferimento:



## 2. RAPPRESENTAZIONE PROBABILISTICA DELLA RETE

Dopo aver scelto il grafo (DAG) ricavato tramite l'approccio Expert System, da utilizzare come componente qualitativa della rete bayesiana, si passa a considerarne la sua componente quantitativa: le tabelle di probabilità condizionate (CPT), ovvero le distribuzioni di probabilità locali, le quali, dal momento che non sono note, devono essere stimate.

L'approccio utilizzato per stimare i parametri (CPT) del modello di rete bayesiana, tramite i dati a disposizione, è quello bayesiano, che prevede il calcolo delle probabilità a posteriori. Il motivo per il quale si è optato per tale scelta riguarda il fatto che non abbiamo alcuna informazione sulla distribuzione delle probabilità a priori. Dunque, la distribuzione ipotizzata per esse è quella uniforme (ipotesi su cui si basa l'algoritmo utilizzato per la stima delle probabilità a posteriori). In tal modo, è possibile ottenere stime più robuste.

È stato fissato un valore piccolo di ISS (imaginary sample size) pari a 8, perché:

- non si ha completa certezza che l'ipotesi di distribuzione uniforme delle probabilità a priori sia valida
- il modello non è molto complesso (con poche variabili nel modello, è difficile che la distribuzione sia uniforme)
- il campione non è così piccolo

Tramite la funzione "bn.fit" sono state ottenute tutte le tabelle di probabilità condizionate stimate. Si riporta, a titolo di esempio, la distribuzione locale di URB (condizionata a EDU).

```
> bn.bayes$URB

  Parameters of node URB (multinomial distribution)

Conditional probability table:

        EDU
URB           high       low
  high   0.3725026 0.2664437
  low    0.2179285 0.3286511
  medium 0.4095689 0.4049052
```

A questo punto è possibile ragionare in termini probabilistici. Ad esempio, 0.3725 è la probabilità che un soggetto, che possiede un diploma o laurea, viva in un'area molto urbanizzata. Inoltre, è molto probabile che chi ha una bassa istruzione viva in un'area mediamente urbana.

Effettuata la stima dei parametri, è stata caratterizzata la rete bayesiana con le sue due componenti: rappresentazione grafica e probabilistica.

```
> bn<-custom.fit(dag,cpt)
> nparams(bn)
[1] 42
```

Il numero di parametri del modello di rete bayesiana considerato è pari a 42: 1+2+2+1*2*3*3+2*2+1*3+1*3+1+1*2*2*2. Senza fattorizzazione della probabilità congiunta (scomposizione di un problema complesso in più problemi di più semplice risoluzione) il numero di parametri sarebbe stato molto più grande, pari al prodotto del numero di livelli di ogni variabile: 2*3*3*2*3*2*2*2*2*2=1728.

Si propone un esempio di verifica di ipotesi di indipendenza (probabilistica) condizionale, utilizzando i dati: URB indipendente da SEX dato EDU. La connessione tra le 3 variabili considerate è di tipo seriale.

```
> ci.test("URB","SEX","EDU",test="mi",data=data)

        Mutual Information (disc.)

data:  URB ~ SEX | EDU
mi = 2.0994, df = 4, p-value = 0.7175
alternative hypothesis: true value is greater than 0


> ci.test("URB","SEX","EDU",test="x2",data=data)

        Pearson's X^2

data:  URB ~ SEX | EDU
x2 = 2.0981, df = 4, p-value = 0.7177
alternative hypothesis: true value is greater than 0
```

Sono state utilizzate la statistica test $G^2$ e la statistica test $X^2$. I test statistici portano entrambi a non rifiutare l'ipotesi nulla di indipendenza condizionale, quindi URB e SEX sono condizionalmente indipendenti: dato un valore di EDU, un valore di SEX non influenza la probabilità di URB.

I gradi di libertà delle due statistiche, per la specifica ipotesi di indipendenza condizionale considerata, sono pari a 4:
(numero di livelli di URB – 1) * (numero di livelli di SEX – 1) * (numero di livelli di EDU).

Sono state eseguite poi delle Conditional Independence Queries tramite la funzione "dsep", per valutare la d-separazione tra i nodi del DAG. In particolate, è stata testata l'indipendenza (grafica) condizionale in:

- una connessione seriale (con nodi di interesse: URB, SEX e EDU; stessa connessione testata precedentemente tramite le statistiche test $X^2$ e $G^2$)
- una connessione divergente (con nodi di interesse: ALG, SED e URB)
- una connessione convergente (con nodi di interesse: ALG, SMK e ASTHMA)

```
> dsep(dag,"URB","SEX","EDU")
[1] TRUE
> dsep(dag,"ALG","SED","URB")
[1] TRUE
> dsep(dag,"ALG","SMK")
[1] TRUE
> dsep(dag,"ALG","SMK","ASTHMA")
[1] FALSE
```

Come era possibile aspettarsi, per quanto concerne la connessione seriale e la connessione divergente, si ha separazione grafica tra i nodi (URB e SEX sono d-separati e saranno anche probabilisticamente indipendenti dato un valore di EDU; ALG e SED sono d-separati e saranno anch'essi probabilisticamente indipendenti dato un valore di URB); nel caso della connessione convergente, se non si ha alcuna evidenza su ASTHMA, ALG e SMK sono d-separati e di conseguenza probabilisticamente indipendenti, mentre se si ha una evidenza su ASTHMA, ALG e SMK non sono d-separati ma saranno condizionalmente dipendenti.

Si passa all'utilizzo del modello di rete bayesiana tramite procedure di inferenza statistica, in particolare eseguendo delle Conditional Probability Queries grazie all'utilizzo della libreria "gRain".

Nello specifico, nell'ambito del ragionamento probabilistico, è possibile interpretare e comparare le probabilità. Si considera, a tal proposito, la tabella di probabilità di ASTHMA condizionata a ALG e SED.

```
> querygrain(junction,nodes=c("ASTHMA","ALG","SED"),type="conditional")
, , SED = no

      ALG
ASTHMA        no       yes
   no  0.7535039 0.2755079
   yes 0.2464961 0.7244921

, , SED = yes

      ALG
ASTHMA        no       yes
   no  0.6751085 0.2445218
   yes 0.3248915 0.7554782
```

La probabilità che un soggetto non sedentario e senza allergie abbia l'asma è pari a 0.2464. I sedentari che hanno un'allergia presentano una maggiore probabilità (0.7554)

di avere l'asma rispetto agli attivi che hanno un'allergia (0.7244). Inoltre, le persone attive che non hanno l'asma presentano una maggiore probabilità (0.7535) di non avere un'allergia rispetto alle persone sedentarie che non soffrono di asma (0.6751).

Si riporta adesso la tabella di probabilità congiunta di ASTHMA e ALG.

```
> querygrain(junction,nodes=c("ASTHMA","ALG"),type="joint")
     ASTHMA
ALG           no       yes
  no  0.50141386 0.2217066
  yes 0.06971911 0.2071605
```

La configurazione di stati delle variabili ALG e ASTHMA più probabile (al 50%) è il non avere né asma e né allergia.

Si propone infine la tabella di probabilità marginale di ASTHMA.

```
> querygrain(junction,nodes="ASTHMA",type="marginal")
$ASTHMA
ASTHMA
      no      yes
0.571133 0.428867
```

Si osserva che la probabilità di non soffrire d'asma è maggiore (0.57) rispetto a quella di soffrirne (0.43).

Supponiamo, infine, di avere due nuove evidenze: i soggetti fumano e hanno un'allergia. In tal modo, avremo un aggiornamento di tutte le tabelle di probabilità condizionate.

```
> j1<-setEvidence(junction,nodes=c("ALG","SMK"),states=c("yes","yes"))
```

La domanda che ci si pone è la seguente: i soggetti che fumano e hanno un'allergia hanno una maggiore probabilità di soffrire di asma?

```
> querygrain(junction,nodes="ASTHMA",type="marginal")
$ASTHMA
ASTHMA
      no      yes
0.571133 0.428867

> querygrain(j1,nodes="ASTHMA",type="marginal")
$ASTHMA
ASTHMA
       no       yes
0.2435954 0.7564046
```

Si può osservare che la distribuzione marginale di probabilità di ASTHMA cambia completamente: in particolare, la probabilità di avere l'asma aumenta fortemente, passando dal 42.88% al 75.64%. Ciò vuol dire che lo stato di ALG e lo stato di SMK influenzano la probabilità di ASTHMA.

A titolo di esempio, si riporta la tabella di probabilità di ASTHMA condizionata a SED prima e dopo aver considerato le evidenze, notando che le probabilità si sono aggiornate.

```
> querygrain(junction,nodes=c("ASTHMA","SED"),type="conditional")
      SED
ASTHMA        no       yes
   no  0.6204933 0.5560702
   yes 0.3795067 0.4439298
> querygrain(j1,nodes=c("ASTHMA","SED"),type="conditional")
      SED
ASTHMA       no      yes
   no  0.171875 0.265625
   yes 0.828125 0.734375
```

# CONCLUSIONI

La rete bayesiana finale, rappresentante le relazioni di dipendenza tra le variabili che influenzano l'insorgere dell'asma bronchiale in Italia, è stata ricavata secondo un approccio "expert system". Il sistema esperto costruito sulla base di conoscenze e ipotesi pregresse offre performance migliori rispetto alle reti ottenute attraverso l'algoritmo di apprendimento automatico Hill Climbing.

Una rete di questo tipo può offrire un valido supporto nella pianificazione e gestione del sistema sanitario di tutto il territorio nazionale visto e considerato anche il forte impatto economico che tale patologia ha sulle casse delle singole regioni.