INFERTILITY

.

By Elka Segura

Anno Accademico 2023/2024

ABSTRACT

L'infertilità è un problema comune che colpisce le coppie. Può essere definita come l'incapacità di portare a termine una gravidanza dopo un ragionevole periodo di rapporti sessuali senza adottare misure contraccettive. Le cause dell'aumento della prevalenza dell'infertilità sono difficili da stabilire. Questo aumento potrebbe essere dovuto ad almeno quattro fattori: il rinvio della decisione di avere figli, le alterazioni della qualità dello sperma dovute ad abitudini come il fumo e l'alcol, i cambiamenti nel comportamento sessuale e l'eliminazione della maggior parte dei tabù.

Ci sono stati molti studi sul tema dell'infertilità. Nel caso di questo lavoro discuteremo di come gli aborti, sia spontanei che indotti, influenzano l'infertilità delle donne. Il set di dati studiato corrisponde a uno studio effettuato e pubblicato con i seguenti riferimenti Trichopoulos *et al* (1976) *Br. J. of Obst. and Gynaec.* **83**, 645–650.

Su questo articolo è stato studiato il ruolo degli aborti indotti (e spontanei) nell'eziologia della sterilità secondaria. Le storie ostetriche e ginecologiche sono state ottenute da 100 donne con infertilità secondaria ammesse al Primo Dipartimento di Ostetricia e Ginecologia della Facoltà di Medicina dell'Università di Atene e alla Divisione di Fertilità e Sterilità di quel Dipartimento. Per ogni paziente si è cercato di trovare due soggetti sani di controllo provenienti dallo stesso ospedale corrispondenti per età, parità e livello di istruzione. Per 83 pazienti indice sono stati trovati due soggetti di controllo ciascuno. Il rischio relativo di infertilità secondaria tra le donne con almeno un aborto indotto e senza aborti spontanei era 3,4 volte quello tra le donne senza aborti indotti o spontanei (intervallo di confidenza al 95% 1,38-8,37). La relazione era statisticamente significativa e indicava che in Grecia circa il 45% dei casi di infertilità secondaria può essere attribuibile a precedenti aborti indotti.

Il presente lavoro viene a determinare l'utilizzo delle rete neurali bayesiana, altro metodo diverso ultizzato nella investigazione referita, il ruolo degli aborti indotti (e spontanei) nell'eziologia della sterilità secondaria.

INTRODUZIONE

Uno degli scopi alla base dell'utilizzo delle reti bayesiane è descrivere e identificare le relazioni di dipendenza e rilevanza tra le variabili per scoprire conoscenza o classificare nuove osservazioni. Il compito precedente è di grande importanza perché fornisce nuova conoscenza in schemi di incertezza, facilitando a sua volta il processo decisionale e il ragionamento attraverso l'uso della teoria della probabilità.

In relazione a quanto sopra, una rete bayesiana descrive un modello probabilistico rappresentato da un grafo aciclico diretto (DAG), dove i nodi sono variabili casuali legate all'argomento affrontato e gli archi costituiscono le relazioni di dipendenza di un nodo sull'altro. In quest'ultima sono codificati i vincoli all'indipendenza condizionale inerenti alla distribuzione congiunta delle variabili. In questo senso la struttura della rete è costituita da nodi e dai rispettivi archi, mentre l'indipendenza condizionata di ciascuna variabile può essere dedotta dagli archi e dalla distribuzione di probabilità congiunta.

Ogni variabile o nodo nel modello è associato alla sua distribuzione di probabilità condizionale (CPT), dati i suoi nodi genitori nel grafico, ad eccezione dei nodi senza genitori che invece della CPT hanno la probabilità marginale ad essi associata. Pertanto, il CPT specifica la distribuzione di probabilità che assume ciascuna variabile, ciascuna assegnazione di valori che possono assumere le altre variabili che compaiono come nodi genitori. Tornando al nostro dataset, iniziamo con un semplice analisi delle variabili e dei dati con cui lavoreremo.

SVILUPPO

Il dataset è composto da 284 istanze e 6 variabili. (Si chiarisce che il set di dati è composto da poche istanze, il che significa che la costruzione della rete potrebbe non essere sufficientemente addestrata perche non disponiamo di più esempi, ma è comunque valida.) Tutte le variabili sono di tipo categoriale/fattoriale, tranne l'età, che è stata trasformata in tipologia categoriale, tenendo conto che il periodo di massima fecondità per una donna è compreso tra i 18 ei 35 anni di età. Le varaibile sono: EDUCATION, AGE, PARITY, INDUCED, CASE, _SPONTANEOUS.

 # [EDUCATION] Istruzione in anni,

. . **livello 0**: 0-5 years

. . **livello 1**: 6-11 years

. . **livello 2**: 12+ years

 # [AGE] Età

. . **[21-35]**: periodo di massima fertilità

. . **[piu35]**: periodo di non massima fertilità

 # [PARITY] Numero di volte che la donna è rimasta incinta

. . **[1]**: a donna è rimasta incinta una sola volta

. . **[>1]**: la donna è rimasta incinta più di una volta

 # [INDUCED] Numero di precedenti aborti indotti

. . **livello "no"**: non ha aborti indotti

. . **livello "yes"**: non ha aborti indotti

 # [SPONTANEOUS] Numero di precedenti aborti spontanei

. . **livello "no"**: non ha aborti spontanei

. . **livello "yes"**: ha aborti spontanei

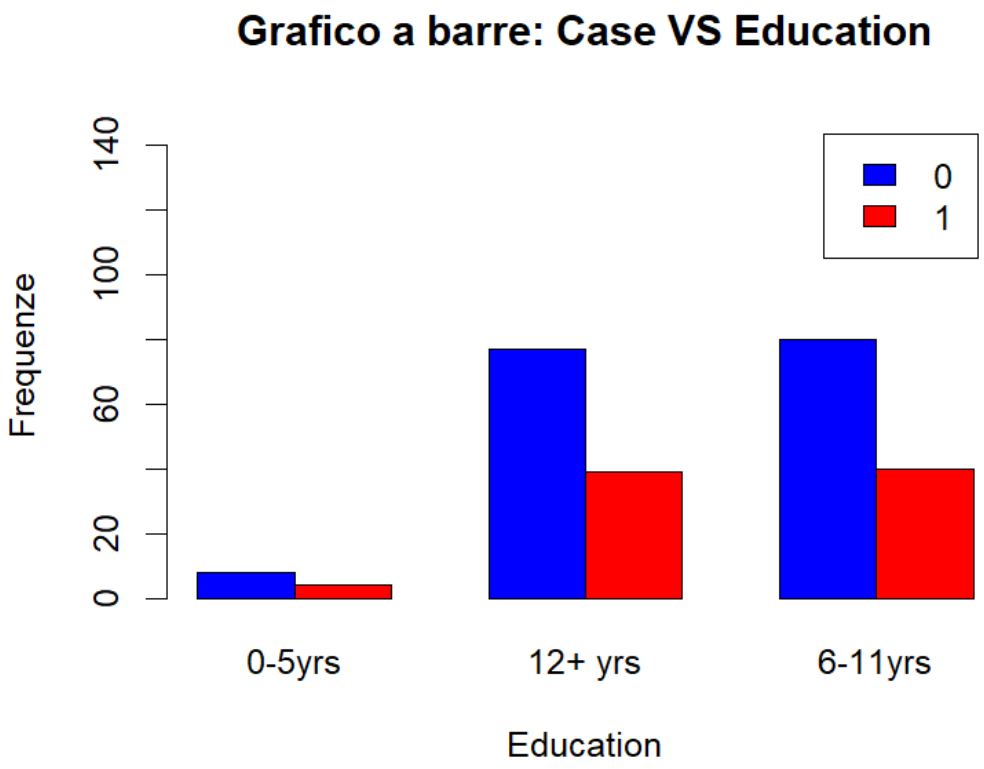
La variabile target è "**case**": la valutazione della "infertility". Questa variabile ha due categorie di valutazione: "infertility:1" e "no infertility:0. Questo studio arriva a valutare la probabilità che la donna sia fertile o meno, dato che ha avuto in precedenza aborti spontanei o indotti. Il tipo di dati delle variabili nel set di dati originale è carattere e numerici e per utilizzare questi dati dobbiamo convertirli in tipo fattoriale, poiché l'obiettivo è proporre una rete bayesiana in grado di modellare dati discreti

Il dataset mostra n=284 individui, di cui 165 classificati come “no infertility” e 83 “infertility”, considerando la variabile target.

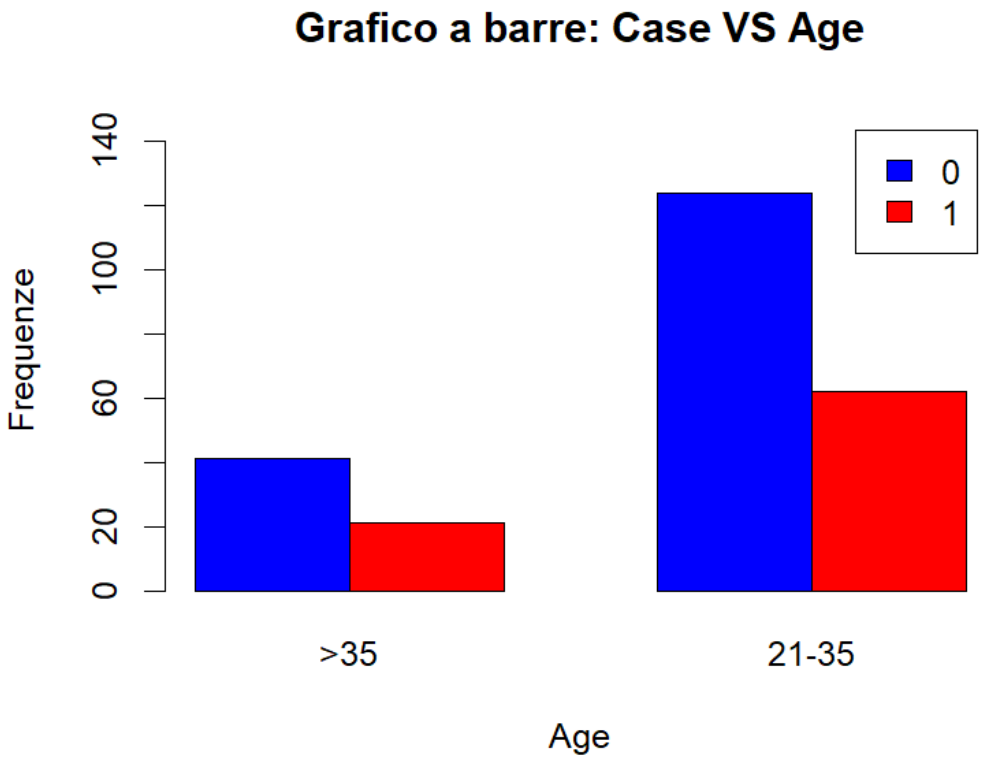
```
infertP <- read.csv("C:/Users/2davi/OneDrive/Desktop/2 Anno/Statistica Per la Azienda(COZZUCOLI)/Proggetto bayesiano 1/Infert/infertP.txt", stringsAsFactors=TRUE)
View(infertP)
#Elimino la columna de que cuenta la cantidad de instancias
infertP<-infertP[][-1]

#Funcion para convertir en factor

infertP[] <-
  lapply(infertP, function(x) {
    if (is.numeric(x) ||
        is.character(x)) {
      return(as.factor(x))
    } else {
      return(x)
    }
  })
```

Per descrivere il comportamento delle altre variabili faremo un'analisi incrociata con la variabile target.

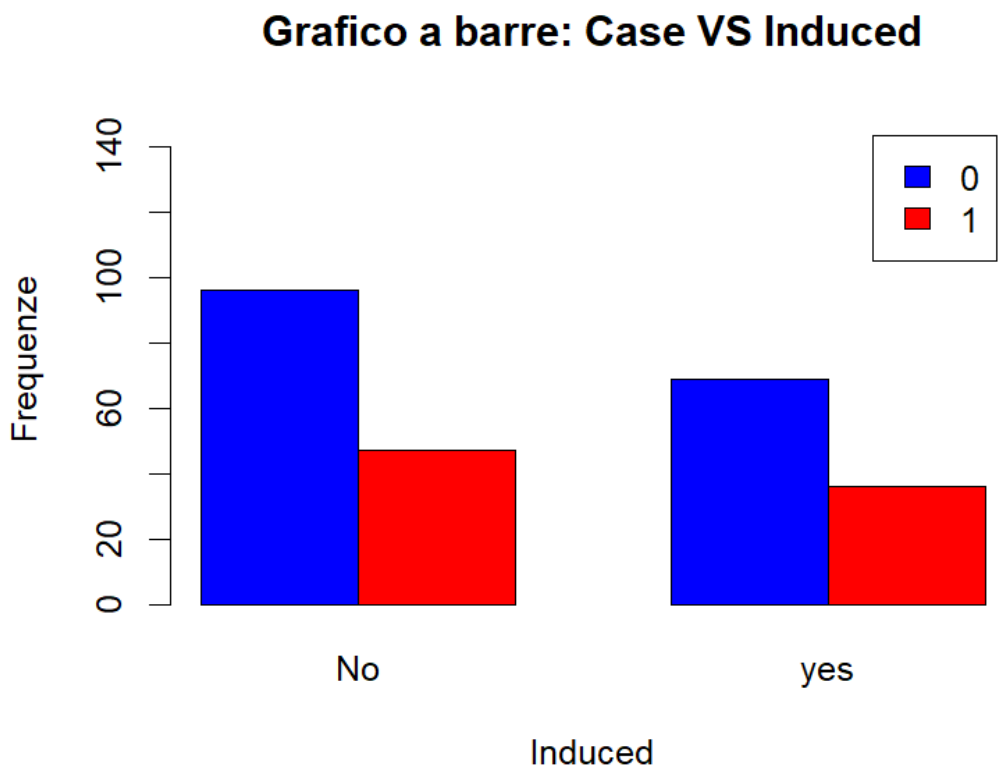
```
#Education
tab <- table(infertP$case, infertP$education)
barplot(tab,
        main="Grafico a barre: Case VS Education",
        xlab="Education",
        ylab="Frequenze",
        legend = rownames(tab),
        ylim = c(0, 150),
        col=c("blue", "red"),
        beside=TRUE)
```



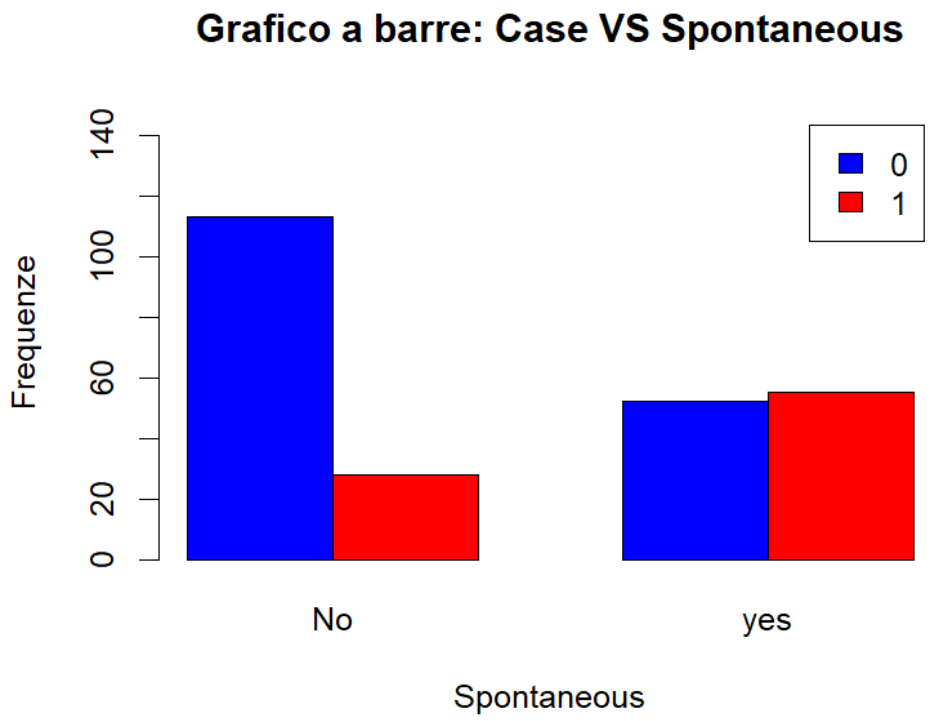

```r
#Age
tab1 <- table(infertP$case, infertP$ageT)
barplot(tab1,
        main="Grafico a barre: Case VS ",
        xlab="Age",
        ylab="Frequenze",
        legend = rownames(tab1),
        ylim = c(0, 150),
        col=c("blue", "red"),
        beside=TRUE)
```




```r
#Induced
tab2 <- table(infertP$case, infertP$inducedT)
barplot(tab2,
        main="Grafico a barre: Case VS Induced",
        xlab="Induced",
        ylab="Frequenze",
        legend = rownames(tab2),
        ylim = c(0, 150),
        col=c("blue", "red"),
        beside=TRUE)
```




```r
#Spontaneous
tab3 <- table(infertP$case, infertP$spontaT)
barplot(tab3,
        main="Grafico a barre: Case VS Spontaneous",
        xlab="case",
        ylab="Frequenze",
        legend = rownames(tab3),
        ylim = c(0, 150),
        col=c("blue", "red"),
        beside=TRUE)
```



Construzione della rete bayesiana

In generale i modelli costruiti come reti bayesiane hanno un significato intuitivo, ciò avviene poiché il grafico che li compone dà una nozione di causalità. Inoltre, le tabelle di probabilità per ciascun nodo aiutano a quantificare queste relazioni. Le teorie su cui si basano le reti bayesiane sono la teoria della probabilità e la teoria dei grafi. Per quanto sopra menzionato, le reti bayesiane hanno avuto successo in diversi ambiti applicativi quali: economia, diagnosi medica, diagnosi industriale, bioinformatica e molti altri.

Una rete bayesiana ha due componenti: una qualitativa e l'altra quantitativa. Il grafico è la parte qualitativa, definita da un insieme di nodi che rappresentano le variabili del modello. Questi nodi sono collegati tra loro in modo direzionale (one Direction) e non devono violare il principio di aciclicità. Deve avere un grafo che parte da un nodo ma non è possibile tornare allo stesso nodo o al nodo precedente. L'altra componente è la parte quantitativa, che stabilisce che per ciascun nodo corrisponderà a una distribuzione di probabilità condizionata. Un altro punto molto importante in questo tipo di rete è che si basa sulla condizione di Markov, che definisce che qualsiasi nodo (X) è condizionatamente indipendente dai suoi non discendenti (Z_{1j} , ...,Z_{nj}) dati i suoi genitori (U_1, ..., U_m). In modo semplice ed informale, la rete bayesiana (G, P) è definita da un grafo di tipo DAG (direzionato, aciclico), al quale è associata un insieme di distribuzione di probabilità congiunta, ovvero i parametri del modello, che soddisfano la condizione di Markov. Il vantaggio della rete bayesiana è la semplificazione, da un numero elevato di parametri da stimare ad un numero ridotto di parametri, garantendo l'equivalenza nel risultato.

Uno dei compiti principali che deve essere svolto quando si lavora con una rete bayesiana consiste nella costruzione di un DAG, attività che può essere sviluppata sulla base di conoscenze pregresse sull'argomento o attraverso l'applicazione di algoritmi sui suoi set di dati relativi al fenomeno o anche attraverso una combinazione di questi.

La costruzione del DAG sulla base delle conoscenze pregresse viene eseguita manualmente, il che può risultare problematico e confuso, mentre la sua costruzione tramite algoritmi di apprendimento applicati a dati che descrivono il fenomeno è molto più rapida e semplice. Quest'ultimo si riferisce all'identificazione del miglior DAG tenendo conto di un punteggio che qualifica i diversi modelli in base alla qualità con cui adattano i dati. L'apprendimento di una rete bayesiana, come abbiamo detto, può essere effettuato combinando conoscenze pregresse e l'applicazione di un algoritmo sul data set. In sostanza, l'informazione preventiva sul modello si configura come la certezza che abbiamo sulla presenza o meno di archi tra i nodi. L'inserimento di queste conoscenze avviene tramite argomenti noti come whitelist o blacklist. Il primo di questi garantisce la presenza di un arco tra due nodi, mentre il secondo assicura che l'arco non esista.

Carichiamo i pakages di R per la construzione de la rete bayesiana

```
library(lattice)
library(BiocManager)
library(gridExtra)
library(gRain)
library(Rgraphviz)
library(graph)
library(bnlearn)
```

Per prima cosa costruiamo un grafico vuoto con la definizione delle variabili del nostro argomento, nel quale se define como target la variabile “CASE” :

```
nodi<-c("education", "ageT", "parityT", "induced", "case", "spontaneous")
nodi

## [1] "education"   "ageT"        "parityT"     "inducedT"     "case"
## [6] "spontaT"

class(nodi)

## [1] "character"

dag<- empty.graph(nodes=nodi, num=1)
dag

##
##   Random/Generated Bayesian network
##
##   model:
##    [education][ageT][parityT][inducedT][case][spontaT]
##   nodes:                                 6
##   arcs:                                  0
##     undirected arcs:                     0
```

```
##     directed arcs:                          0
##   average markov blanket size:              0.00
##   average neighbourhood size:               0.00
##   average branching factor:                 0.00
##
##   generation algorithm:                     Empty
```

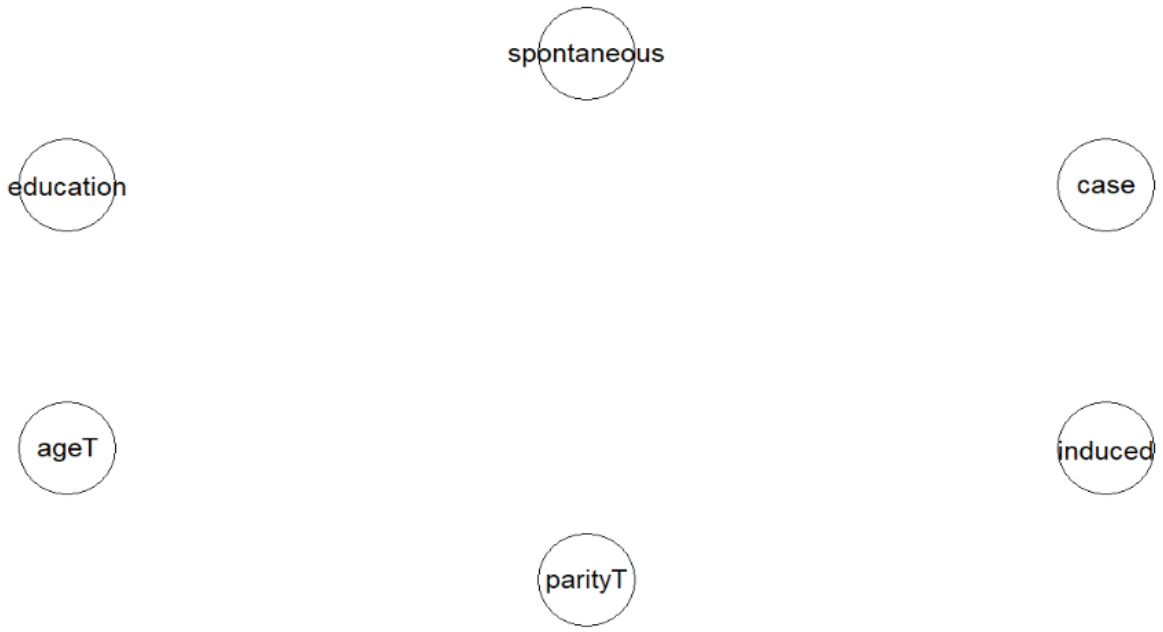
```
class(dag)
```

```
## [1] "bn"
```

```
dag
```

```
##
##   Random/Generated Bayesian network
##
##   model:
##    [education][ageT][parityT][inducedT][case][spontaT]
##   nodes:                                    6
##   arcs:                                     0
##     undirected arcs:                        0
##     directed arcs:                          0
##   average markov blanket size:              0.00
##   average neighbourhood size:               0.00
##   average branching factor:                 0.00
##
##   generation algorithm:                     Empty
```

```
plot(dag)
```



Algoritmo di apprendimento basato sul punteggio:Hill-Climbing

Il compito svolto da questi algoritmi è quello di trovare il miglior DAG, tenendo conto di un punteggio che qualifica i diversi modelli in base alla qualità con cui adattano i dati.

Considerato quanto sopra, l'apprendimento della rete bayesiana diventa un problema di massimizzazione del punteggio. Uno degli algoritmi ampiamente utilizzati è HILL-CLIMBING. L'algoritmo inizia con la selezione di una rete bayesiana, può essere una struttura vuota (senza alcun arco), costruita sulla base di conoscenze pregresse o una selezione casuale. Successivamente viene calcolato il punteggio iniziale di detta rete ed inizia un processo di iterazione in cui si tenta di migliorare tale valore eliminando, aggiungendo o reindirizzando un arco alla volta. Per ogni possibile modifica viene calcolato il punteggio della nuova rete per optare infine per la struttura che ha ottenuto il valore più alto. L'algoritmo termina quando non vi è alcun miglioramento nel punteggio.

Una volta selezionato l'algoritmo per la costruzione della rete, è fondamentale definire il punteggio che si vuole massimizzare tenendo conto della composizione del data set su cui si andrà a lavorare. I punteggi comunemente utilizzati sono l'Akaike Information Criterion (AIC), il BAyesian Information Criterion (BIC) e il Bayesian Dirichlet Equivalent (BDeu), che sono quelli che verranno utilizzati.

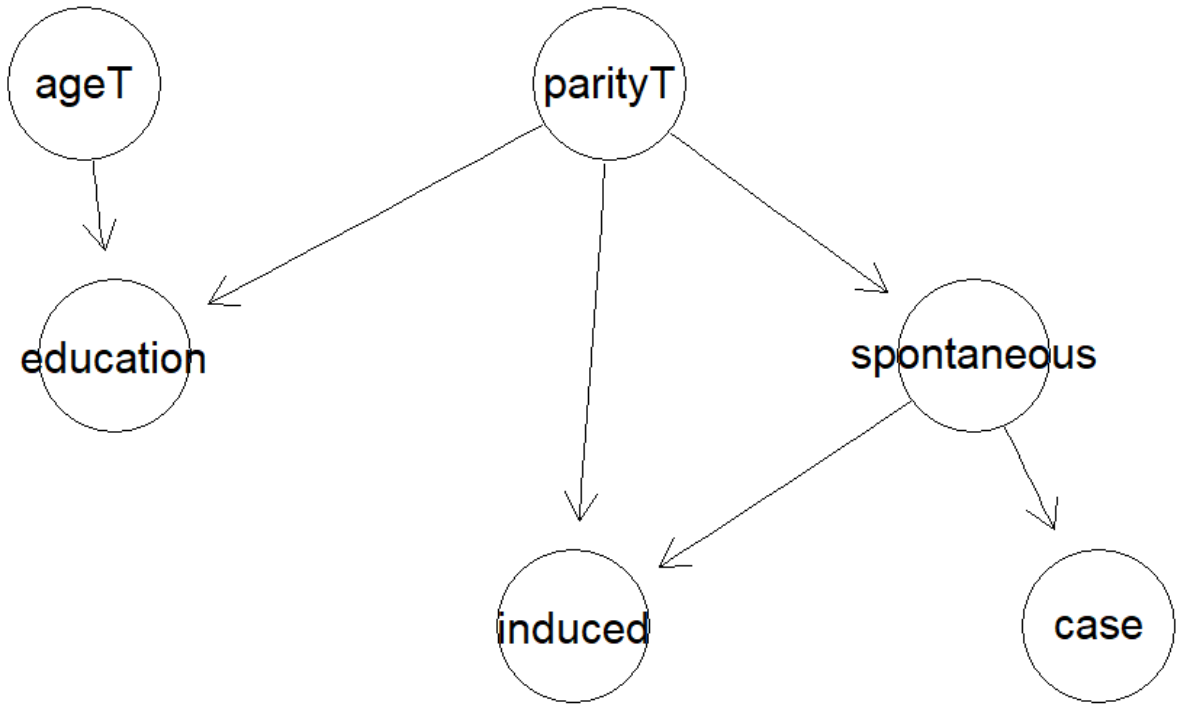
I primi criteri da definire sono la whitelist e la blacklist, in questo caso non abbiamo una conoscenza preliminare degli archi che devono esserci, ma sappiamo che nessun arco dovrebbe uscire dal nodo della *case* poiché è la variabile target. Troviamo anche l'attributo age, è un valore intrinseco che non dipende da nessun altro attributo, quindi è considerato un genitore a cui non deve entrare nessun arco.

Blacklist

```
cond1<-data.frame(from=c("case"), to=c("ageT","parityT", "inducedT", "spontaT
", "education"))
cond2<-data.frame(from=c("parityT", "induced", "spontaT", "education"), to=c(
"ageT"))
blacklist=data.frame(rbind(cond1, cond2))
```

Rete di apprendimento con HC criterio BIC

```
learned <- hc(infertP, blacklist=blacklist)
graphviz.plot(learned, fontsize = 20)
```




```
score(learned, data=infertP, type="bic")
```

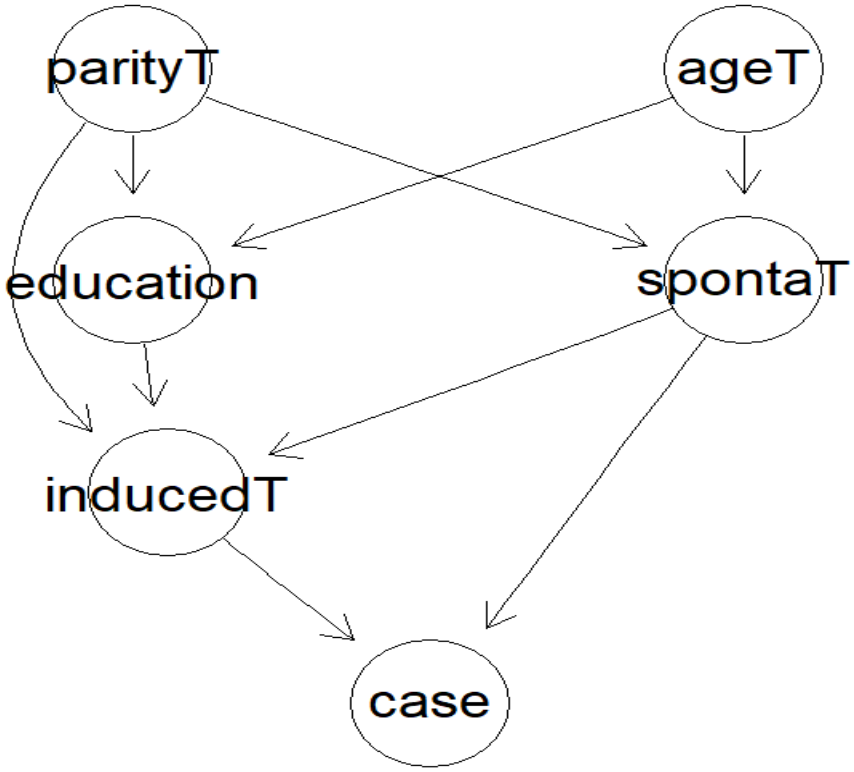
```
## [1] -998.3877
```

```
arc.strength(learned, data = infertP, criterion = "x2")
```

```
##      from        to     strength
## 1 spontaT      case 1.849545e-07
## 2 parityT  inducedT 2.594708e-08
## 3 spontaT  inducedT 3.305043e-06
## 4    ageT education 1.661995e-06
## 5 parityT education 1.497933e-04
## 6 parityT   spontaT 1.098977e-02
```

Rete di apprendimento con HC criterio AIC

```
learned1 <- hc(infertP, score="aic", blacklist=blacklist)
graphviz.plot(learned1, fontsize = 20)
```




```
score(learned1, data=infertP, type="aic")
```

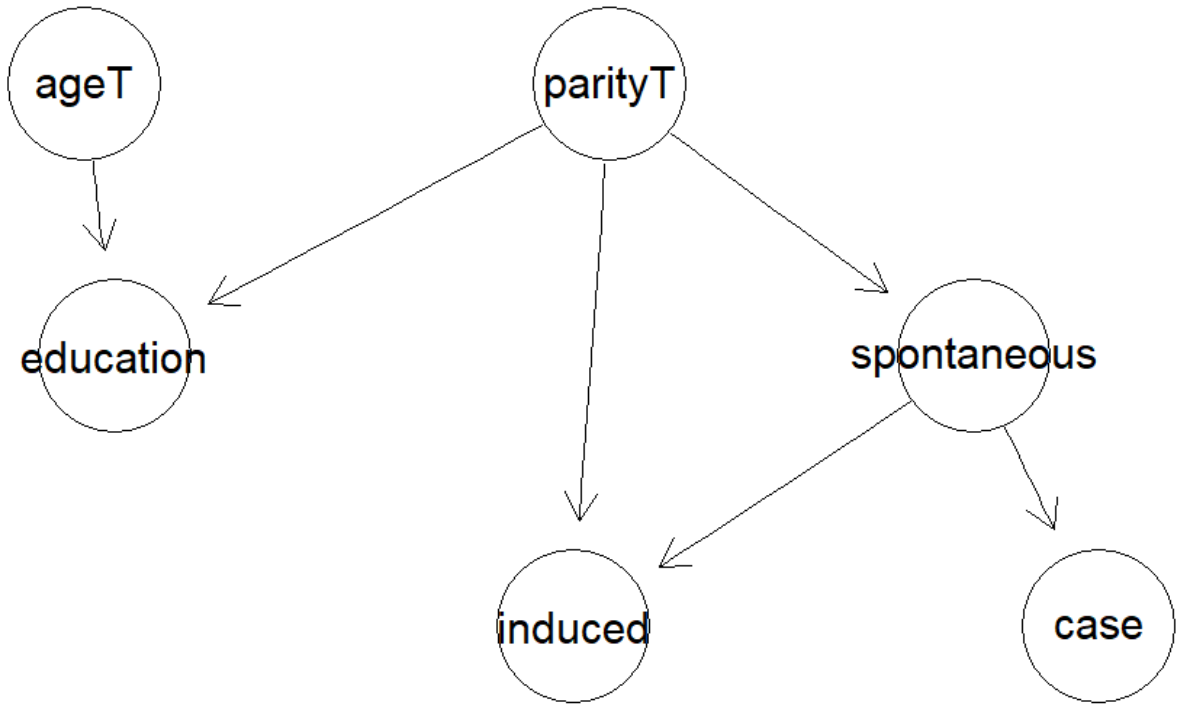
```
## [1] -958.4604
```

```
arc.strength(learned1, data = infertP, criterion = "x2")
```

```
##         from        to     strength
## 1    spontaT      case 8.338237e-08
## 2    parityT  inducedT 1.058571e-07
## 3    spontaT  inducedT 2.351311e-07
## 4  education  inducedT 1.366016e-03
## 5       ageT education 1.661995e-06
## 6    parityT education 1.497933e-04
## 7    parityT   spontaT 5.788313e-03
## 8   inducedT      case 2.648418e-02
## 9       ageT   spontaT 1.341710e-01
```

Rete di apprendimento con HC criterio BDeu

```
learned2 <- hc(infertP, score="bde", blacklist=blacklist)
graphviz.plot(learned2, fontsize = 20)
```




```
score(learned2, data=infertP, type="bde")

## [1] -1001.729

arc.strength(learned2, data = infertP, criterion = "x2")

##      from        to     strength
## 1 spontaT      case 1.849545e-07
## 2 parityT  inducedT 2.594708e-08
## 3 spontaT  inducedT 3.305043e-06
## 4    ageT education 1.661995e-06
## 5 parityT education 1.497933e-04
## 6 parityT   spontaT 1.098977e-02
```

Abbiamo anche utilizzato la funzione arc.strength per misurare quanto sono significativi gli archi della rete.

	BIC	AIC	BDeu
Score	-998.877	-958.4604	-1001.729
Average Markov blanket Size	2.67	3.67	2.33

Dai modelli stimati con l'algoritmo Hill-Climbing viene selezionato quello stimato con il punteggio Bdeu. Come sappiamo, l'AIC prende in considerazione una penalità inferiore

rispetto agli altri punteggi, ma effettuando il confronto in termini di DAG, i modelli con criterio BIC e Bdeu coincidono con la stessa struttura. Possiamo notare anche che il Average Blankov Market del *learned2* è il minore di tutti.

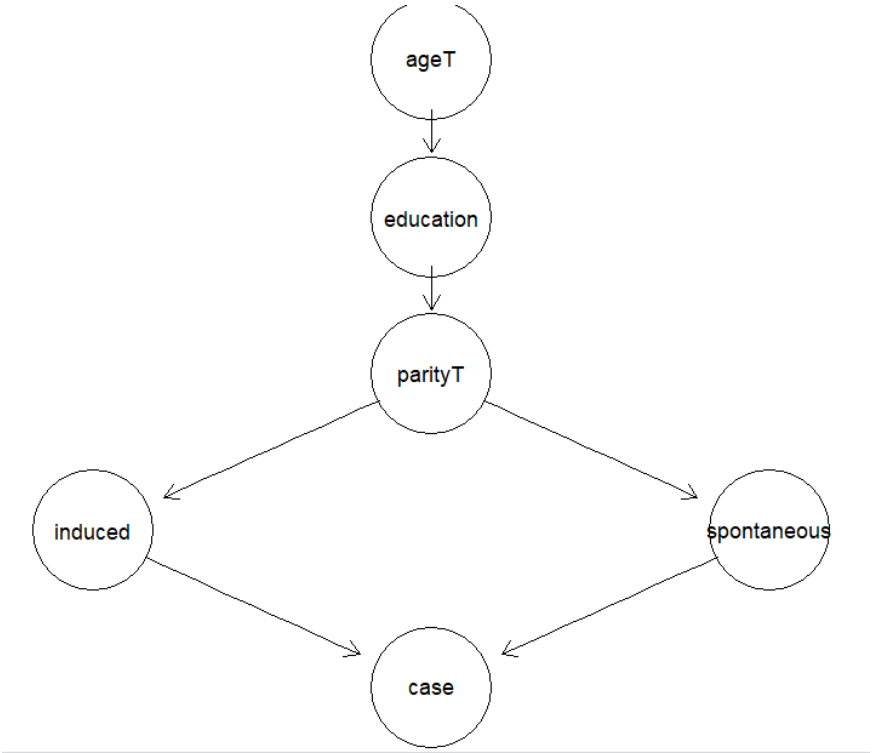
Modifica la rete selezionata

Dopo aver ottenuto i modelli applicando l'algoritmo HILL-CLIMBING, selezioniamo la rete ottenuta con il punto Bdeu. In questa parte introdurremo modifiche alla rete selezionata per rendere più logicamente le relazioni stabilite negli archi tra i nodi. Biologicamente una donna può avere aborti spontanei o indotti a seconda del numero di gravidanze precedenti, ma l'aborto è indotto o spontaneo. Osservando le reti BIC e BDeu, vediamo la relazione *inducedT-spontaT*, secondo cui se una donna ha avuto in precedenza aborti indotti, c'è una probabilità che avrà aborti spontanei che può essere dato da diversi motivi, tra questi, lesioni all'utero o alla cervice, ma secondo gli studi, anche dopo l'aborto indotto e la sua esecuzione errata aumenta il rischio che la infertilità della donna venga compromessa e anche è probabile che si sia verificato un nuovo aborto, ma questa volta involontario.

Sulla base di questo ragionamento procederemo ad eliminare l'arco tra *spontaT* e *inducedT*, e aggiungeremo un arco tra *inducedT* e *case*. Otterremo una connessione convergente in cui il nodo caso avrà come genitore *inducedT* e lo *spontaT*. L'altra modifica è invertire l'arco tra *education* e *parityT*.

Bdeu Modifica I

```
learned2<- reverse.arc(learned2, from="parityT", to="education")
learned2<- drop.arc(learned2, from="spontaT", to="inducedT")
learned2<- set.arc(learned2, from="inducedT", to="case")
graphviz.plot(learned2, fontsize = 25)
```




```
score(learned2, data=infertP, type="bde")
```

```
## [1] -1120.835
```

```
arc.strength(learned2, data = infertP, criterion = "x2")
```

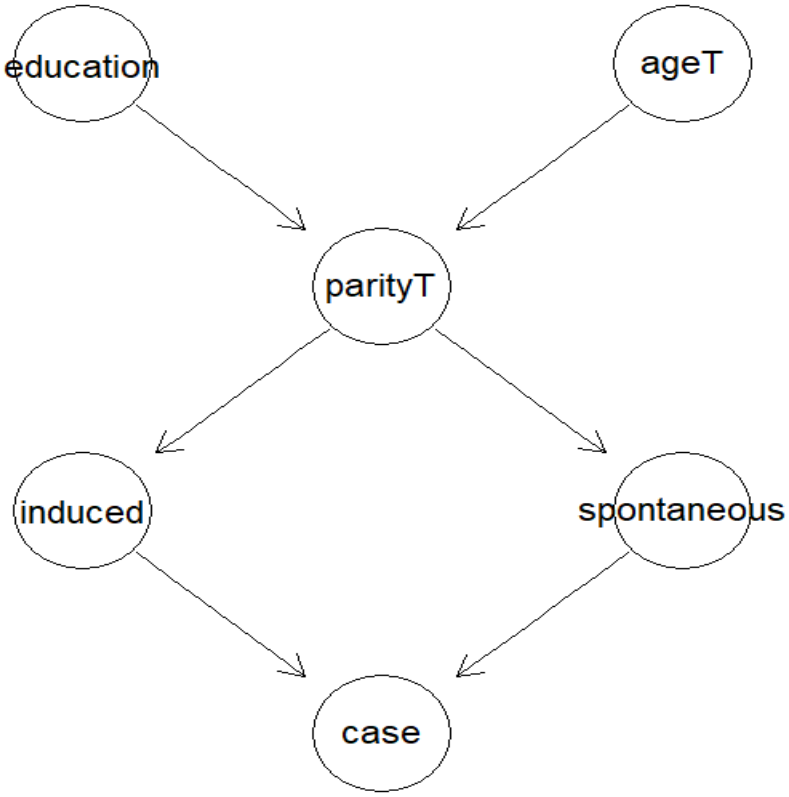
```
##         from        to     strength
## 1   spontaT      case 8.338237e-08
## 2   parityT  inducedT 6.911248e-07
## 3      ageT education 8.166134e-04
## 4   parityT   spontaT 1.098977e-02
## 5 education   parityT 1.079570e-01
## 6  inducedT      case 2.648418e-02
```

Come vediamo i risultati di questa nuova rete, il punteggio è peggiore di quello ottenuto da learned2 (ma non significativamente lontana), da -1001.729 a -1120.835. Lo average mark ov blanket size rimane a 2.33. Lo strength education-parityT è il meno significativi con p-va lue di 0.1079.

Bdeu Modifica II

Seguendo un ragionamento logico, i livelli educativi ed economici più bassi hanno un tasso di prevalenza più elevato di gravidanze, spesso nell'adolescenza. Molte indagini di questo tipo presentano questa affermazione con alta probabilità. Detto questo, abbiamo introdotto le seguenti modifiche: eliminare l'arco tra *ageT* e *education*, anche perché potremmo considerare l'attributo *education* in questo caso come attributo intrinseco alla persona e aggiungere un arco tra *ageT* e *parityT.*

```
learned2<- drop.arc(learned2, from="ageT", to="education")
learned2<- set.arc(learned2, from="ageT", to="parityT")
graphviz.plot(learned2, fontsize = 25)
```




```
score(learned2, data=infertP, type="bde")

## [1] -1119.781

arc.strength(learned2, data = infertP, criterion = "x2")

##         from       to     strength
## 1   spontaT     case 8.338237e-08
## 2   parityT inducedT 6.911248e-07
## 3   parityT  spontaT 1.098977e-02
## 4 education  parityT 1.497933e-04
## 5  inducedT     case 2.648418e-02
## 6      ageT  parityT 1.950170e-04
```

	Bdeu_modifica I	Bdeu_modifica II
Score	-1120.835	-1119.781
Average Blankov Market Size	2.33	2.67

Come vediamo il punteggio non è migliorato tanti ma è megliore dal primo modificato. Nonostante lo average markov blanket size aumenta. Sulla base di quest'ultima rete ottenuta verranno effettuate le analisi.

Verifica delle dipendenze del DAG

```
#Criterio x2
A<-ci.test("parityT", "ageT","education", test = "x2", data = infertP)
B<-ci.test("parityT", "education","ageT", test = "x2", data = infertP)
C<-ci.test("inducedT", "parityT", test = "x2", data = infertP)
D<-ci.test("spontaT", "parityT", test = "x2", data = infertP)
E<-ci.test("case", "induced", "spontaT", test="x2", data=infertP)
F<-ci.test("case", "spontaT", "inducedT", test="x2", data=infertP)

nameR<-c(A$data.name,B$data.name, C$data.name, D$data.name, E$data.name, F$da
ta.name)
p_value<-c(A$p.value, B$p.value, C$p.value, D$p.value, E$p.value, F$p.value)
colN<-c("name", "p-value")

testChi<-cbind(nameR, p_value)
testChi

##      nameR                          p_value
## [1,] "parityT ~ ageT | education" "0.000195017036608122"
## [2,] "parityT ~ education | ageT" "0.000149793293014544"
## [3,] "inducedT ~ parityT  "       "6.91124827533478e-07"
## [4,] "spontaT ~ parityT  "        "0.0109897671346661"
## [5,] "case ~ inducedT | spontaT"  "0.0264841763438041"
## [6,] "case ~ spontaT | inducedT"  "8.33823732135854e-08"

#Criterio Mi
G<-ci.test("parityT", "ageT","education", test = "mi", data = infertP)
H<-ci.test("parityT", "education","ageT", test = "mi", data = infertP)
I<-ci.test("inducedT", "parityT", test = "mi", data = infertP)
J<-ci.test("spontaT", "parityT", test = "mi", data = infertP)
K<-ci.test("case", "inducedT", "spontaT", test="mi", data=infertP)
L<-ci.test("case", "spontaT", "inducedT", test="mi", data=infertP)



nameT<-c(G$data.name,H$data.name, I$data.name, J$data.name, K$data.name, L$da
ta.name)
p_value1<-c(G$p.value, H$p.value, I$p.value, J$p.value, K$p.value, L$p.value)

testMi<-cbind(nameT, p_value1)
testMi

##      nameT                          p_value1
## [1,] "parityT ~ ageT | education" "9.19345211867413e-05"
## [2,] "parityT ~ education | ageT" "4.29435063036491e-05"
## [3,] "inducedT ~ parityT  "       "4.24270567491643e-07"
```

```
## [4,] "spontaT ~ parityT  "        "0.0105235815985642"
## [5,] "case ~ inducedT | spontaT"  "0.022703372693251"
## [6,] "case ~ spontaT | inducedT"  "2.665777611541e-08"
```

Come si vede, le dipendenze del DAG risultano significative tutte gli archi.

Il modello selezionato:

Learned2= P[education]P[ageT]P[parityT|education: ageT]P[induced|parityT]P[spontaneous|parityT]P[case|induced:spontaneous]

I parametri (CPT) della rete sono stimati dall'utilizzo dell'approccio bayesiano utilizzando la funzione bn.fit(). Tale approccio garantisce di non avere-probabilità uguali a 0

```
bn.bayes <- bn.fit(learned2, data = infertP, method = "bayes",iss = 10)
bn.bayes
```

```
##
##   Bayesian network parameters
##
##   Parameters of node education (multinomial distribution)
##
## Conditional probability table:
##      0-5yrs    12+ yrs    6-11yrs
## 0.05943152 0.46253230 0.47803618
##
##   Parameters of node ageT (multinomial distribution)
##
## Conditional probability table:
##        >35     21-35
## 0.2596899 0.7403101
##
##   Parameters of node parityT (multinomial distribution)
##
## Conditional probability table:
##
## , , ageT = >35
##
##        education
## parityT    0-5yrs   12+ yrs   6-11yrs
##      >1 0.5000000 0.7946429 0.4631148
##      1  0.5000000 0.2053571 0.5368852
##
## , , ageT = 21-35
##
##        education
## parityT    0-5yrs   12+ yrs   6-11yrs
##      >1 0.8913043 0.4850993 0.7358871
##      1  0.1086957 0.5149007 0.2641129
```

```
##
##
##   Parameters of node inducedT (multinomial distribution)
##
## Conditional probability table:
##
##         parityT
## inducedT        >1         1
##      No  0.4512987 0.7548077
##      yes 0.5487013 0.2451923
##
##   Parameters of node case (multinomial distribution)
##
## Conditional probability table:
##
## , , spontaT = No
##
##     inducedT
## case        No       yes
##    0 0.8812950 0.7091503
##    1 0.1187050 0.2908497
##
## , , spontaT = yes
##
##     inducedT
## case        No       yes
##    0 0.4745223 0.5149254
##    1 0.5254777 0.4850746
##
##
##   Parameters of node spontaT (multinomial distribution)
##
## Conditional probability table:
##
##        parityT
## spontaT        >1         1
##     No  0.5032468 0.6586538
##     yes 0.4967532 0.3413462
```

Dal punto di vista strutturale e anche probabilistico, la rete trovata dall'applicazione dell'algoritmo Hill Climbing è abbastanza semplice. Una struttura con connessione convergente, divergente e seriale. In questo senso proviamo a fare un'inferenza su Conditional Indenpendency Query e d-separazione, in cui confrontiamo alcuni nodi

```
dsep(learned2, x = "education", y = "ageT")
```

```
## [1] TRUE
```

```
#Conessione serial
dsep(learned2, x = "ageT", y = "inducedT") # nessuna evidenza su parityT
```

```
## [1] FALSE
```

```
dsep(learned2, x = "ageT", y = "inducedT", z="parityT") #  evidenza su parity
T
```

```
## [1] TRUE
```

```
#Conessione divergente
dsep(learned2, x = "inducedT", y = "spontaT") #  nessuna evidenza su parityT
```

```
## [1] FALSE
```

```
dsep(learned2, x = "inducedT", y = "spontaT", z="parityT") #  evidenza su par
ityT
```

```
## [1] TRUE
```

```
#Conessione convergente
dsep(learned2, x = "ageT", y = "education", z="parityT") #  evidenza su parit
yT
```

```
## [1] FALSE
```

```
dsep(learned2, x = "inducedT", y = "spontaT", z="case") #  evidenza su case
```

```
## [1] FALSE
```

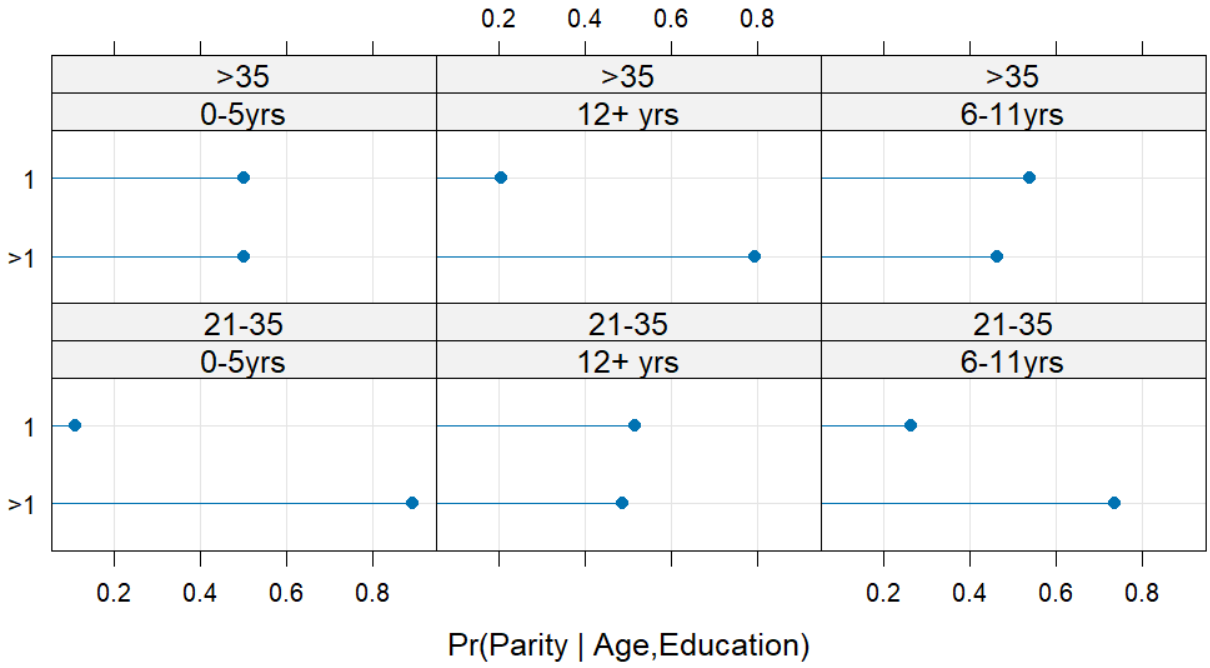
In questo caso tra i nodi *ageT, parityT* e *inducedT* esiste una connessione seriale. È stato applicato il test d-separazione che dimostra che i nodi *ageT* e *inducedT* non sono d-separati avendo evidenza del nodo *parityT*, quindi sono condizionatamente indipendenti dato *parityT*.

Come connessione divergente abbiamo la relazione *inducedT* , *spontaT* e *parityT*. In cui si dimostra che i nodi *induced* e *spontaT* sono d-separati data *parityT*, cioè sono condizionatamente indipendenti dato parityT.

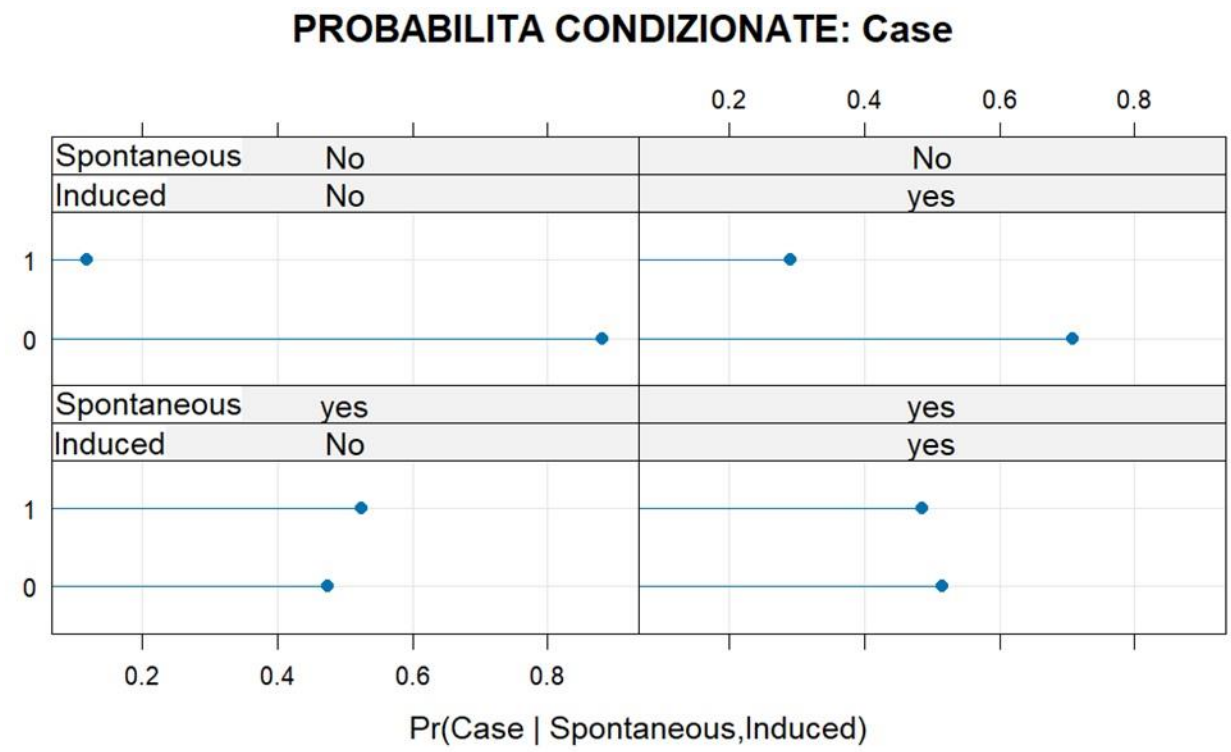
Abbiamo che nel caso de connessioni convergente i genitori (le casuse *ageT e education*) sono marginalemente indipendenti tra di loro, ma se si ha qualche evidenza sul figlio (*parityT* effetto comune), allora sono condizionalmente dipendenti, *ageT* è condizionalmente dipendenti da *education* data la evidenza sul figlio *parityT*. Anche l'atro caso convergente in quale si dimostra *inducedT* è condizionalmente dipendenti da *spontaT* data la evidenza sul figlio *case.*

CPT

```
bn.fit.dotplot(bn.bayes$parityT, main = "PROBABILITA CONDIZIONATE: Parity",xl
ab = "Pr(Parity | Age,Education)", ylab = "")
```




```
bn.fit.dotplot(bn.bayes$case, main = "PROBABILITA CONDIZIONATE: Case",xlab =
"Pr(Case | Spontaneous,Induced)", ylab = "")
```



In questo passo, facciamo inferenza su Conditional Indenpendency Query sul modello e iniziamo a porre domande al modello, stabilendo evidenze o meno.

```
library(gRain)
junction = compile(as.grain(bn.bayes))
junction

## Independence network: Compiled: TRUE Propagated: FALSE Evidence: FALSE
```

Qual è la probabilità che la donna abbia problemi di infertilità, quando quando hai avuto aborti indotti?

```
#senza envidenze
querygrain(junction, nodes="case", type="marginal")

## $case
## case
##         0         1
## 0.6708457 0.3291543

#con envidenze
dom1<-setEvidence(junction, nodes="inducedT", states = "yes")
querygrain(dom1, nodes="case", type="marginal")

## $case
## case
##       0       1
## 0.61902 0.38098
```

Qual è la probabilità che la donna abbia problemi di infertilità, quando quando ha avuto aborti spontanei?

```
#senza envidenze
querygrain(junction, nodes="case", type="marginal")

## $case
## case
##         0         1
## 0.6708457 0.3291543

#con envidenze
dom2<-setEvidence(junction, nodes="spontaT", states = "yes")
querygrain(dom2, nodes="case", type="marginal")

## $case
## case
##         0         1
## 0.4931268 0.5068732
```

Qual è la probabilità che la donna abbia problemi di infertilità, quando quando ha avuto aborti spontanei e aborti indotti?

```
#senza envidenze
querygrain(junction, nodes="case", type="marginal")
```

```
## $case
## case
##         0         1
## 0.6708457 0.3291543
```

```
#con envidenze
dom3<-setEvidence(junction, nodes=c("spontaT","inducedT"), states = c("yes","
yes"))
querygrain(dom3, nodes="case", type="join")
```

```
## case
##         0         1
## 0.5149254 0.4850746
```

Qual è la probabilità che la donna maggiore di 35 anni e rimasta incinta più di una volta, abbia problemi di infertilità, quando quando ha avuto aborti spontanei?

```
#senza envidenze
querygrain(junction, nodes="case", type="marginal")
```

```
## $case
## case
##         0         1
## 0.6708457 0.3291543
```

```
#con envidenze
dom4<-setEvidence(junction, nodes=c("ageT","parityT","spontaT"), states = c("
>35",">1","yes"))
querygrain(dom4, nodes="case", type="join")
```

```
## case
##         0         1
## 0.4966915 0.5033085
```

Qual è la probabilità che una donna con un'istruzione di 0 a 5 anni abbia avuto più di una gravidanza?

```
#senza envidenze
querygrain(junction, nodes="parityT", type="marginal")
```

```
## $parityT
## parityT
##        >1         1
## 0.6264056 0.3735944
```

```
#con envidenze
dom5<-setEvidence(junction, nodes="education", states = "0-5yrs")
querygrain(dom5, nodes="parityT", type="marginal")

## $parityT
## parityT
##        >1         1
## 0.7896866 0.2103134
```

CONCLUSIONE

I risultati ottenuti dalla costruzione della rete bayesiana si basano sulla probabilità condizionata che possiamo leggere. In questo caso, come avevamo detto prima, i dati per costruire la rete sono pochi, quindi la fiducia non è la migliore possibile, ma non è sbagliata. Confrontando i risultati della ricerca che precede questo insieme di dati, si dice che la probabilità che una donna abbia problemi di infertilità se ha avuto in precedenza aborti provocati è del 45%, con l'applicazione della rete bayesiana arriviamo al 38%. Non sono valori così distanti e la causa di questa differenza è che per la costruzione della rete non abbiamo tenuto conto di altre variabili che in questo caso potrebbero contribuire al 45% ottenuto dall'indagine iniziale. La ricerca risale al 1976, molti sono stati i progressi in campo medico. Le cause dell'infertilità possono essere diverse, ma per quanto riguarda gli aborti indotti, una procedura eseguita in modo inadeguato implica grandi implicazioni per la gravidanza successiva.

RIFERIMENTI BIBLIOGRAFICI

Trichopoulos D, Handanos N, Danezis J, Kalandidi A, Kalapothaki V. Induced abortion and secondary infertility. Br J Obstet Gynaecol. 1976 Aug;83(8):645-50. doi: 10.1111/j.1471-0528.1976.tb00904.x. PMID: 952796.