# Un'analisi della Sindrome dell'Ovaio Policistico mediante Reti Neurali e Reti Bayesiane

Francesca Bellissimo

## Contents

# 1 Sommario

Questo studio si concentra sull'analisi della Sindrome dell'Ovaio Policistico (PCOS), una delle patologie endocrine più comuni nelle donne in età fertile, caratterizzata da una serie di sintomi che includono irregolarità mestruali, iperandrogenismo e ovaie policistiche.

Il lavoro si articola in due fasi: nella prima parte, viene utilizzata una rete neurale per sviluppare un classificatore capace di predire la tipologia di PCOS di una paziente sulla base di variabili cliniche e anamnestiche, sfruttando la capacità delle reti neurali di riconoscere e modellare pattern complessi. Nella seconda parte, viene costruita una rete bayesiana che permette di esplorare e analizzare le relazioni di dipendenza tra le diverse variabili.

L'obiettivo finale è quello di ottenere una comprensione più completa e integrata della PCOS, utilizzando entrambe le tecniche per analizzare il problema da diverse prospettive.

## 2 Introduzione

La sindrome dell'ovaio policistico (PCOS, dall'inglese *Polycystic Ovary Syndrome*) è un disturbo endocrino-metabolico molto diffuso tra le donne in età fertile; la prevalenza complessiva risulta essere del 6% (95% IC: 5%-8%) secondo i criteri diagnostici del *National Institute of Health* (NIH), del 10% (95% IC: 8%-13%) secondo i criteri della *Androgen Excess & PCOS Society*, e del 10% (95% IC: 7%-13%) secondo i criteri di Rotterdam [1].

È importante precisare che, a dispetto del nome, la presenza di cisti ovariche non è un requisito indispensabile per la diagnosi di PCOS. In molti casi, le donne con PCOS presentano numerosi piccoli follicoli che non completano il processo di maturazione, piuttosto che cisti vere e proprie. La PCOS, comunque, è associata a numerosi sintomi, eterogenei e non sempre tutti presenti. Uno dei tratti maggiormente distintivi è il verificarsi di disfunzioni ovariche, manifestate tramite la presenza di oligomennorea ($35 <$ durata del ciclo mestruale in giorni $< 90$) o amenorrea (durata del ciclo mestruale in giorni $> 90$). Tali condizioni derivano, in genere, da anovulazione cronica e determinano spesso infertilità. Va specificato, tuttavia, che il 60% delle donne con PCOS sono fertili, sebbene il tempo necessario al concepimento risulti maggiore [2]. Le donne con PCOS presentano frequentemente un rapporto LH:FSH (ormone luteinizzante/ormone follicolo-stimolante) elevato, una condizione che contribuisce all'aumento della produzione di androgeni, responsabili a loro volta di sintomi come irsutismo, acne e alopecia androgenetica. L'insulino-resistenza è un altro tratto molto comune nelle donne con PCOS; in presenza di tale condizione, le cellule risultano scarsamente sensibili all'insulina e, di conseguenza, il pancreas ne produce quantità maggiori. Questo porta a iperinsulinemia, che a sua volta può contribuire all'accumulo di tessuto adiposo e impedisce il completo sviluppo dei follicoli ovarici. Inoltre, la *International Diabetes Federation* ha identificato la PCOS come un fattore di rischio significativo e non modificabile associato al Diabete di tipo 2 [3]. Oltre agli aspetti biologici, la sindrome è associata anche a problemi psicologici, come disturbi d'ansia e depressione, che si verificano con maggiore frequenza nelle donne con PCOS, soprattutto tra quelle non fertili, rispetto alla popolazione generale [4].

Il trattamento della PCOS si concentra principalmente sulla gestione dei sintomi e sulla correzione dei disturbi ormonali e metabolici. In prima linea, si consiglia una sostanziale modifica dello stile di vita. È possibile migliorare la sensibilità all'insulina e ripristinare un normale funzionamento mestruale tramite l'esercizio fisico e una modesta perdita di peso [5]. Il trattamento farmacologico, invece, prevede principalmente l'utilizzo di contraccettivi orali o di sensibilizzanti all'insulina (come la Metformina), sebbene tali metodi non debbano sostituire l'adozione di uno stile di vita più sano.

# 3 Materiali

Ho utilizzato il *Polycystic Ovary Sindrome Dataset* [6], disponibile su Kaggle. Il dataset include informazioni relative a 541 pazienti di dieci diversi ospedali dello stato del Kerala, in India. Tra queste, 177 (il 33% circa) presentano la sindrome. Il dataset contiene 42 variabili, sia cliniche che relative allo stile di vita; tutte le variabili sono riportate nella tabella sottostante, con relative statistiche descrittive.

Table 1: Statistiche descrittive delle variabili utilizzate

| | Tipo | Descrizione | mean | sd | min | max |
|---|---|---|---|---|---|---|
| PCOS | binaria | Presenza di PCOS (0 = No, 1 = Sì) | 0.33 | 0.47 | 0.00 | 1.00 |
| Age | discreta | Età(anni) | 31.43 | 5.41 | 20.00 | 48.00 |
| Weight | continua | Peso(kg) | 59.64 | 11.03 | 31.00 | 108.00 |
| Height | continua | Altezza(cm) | 156.48 | 6.03 | 137.00 | 180.00 |
| BMI | continua | Indice di massa corporea(kg/mˆ2) | 24.32 | 4.05 | 12.42 | 38.90 |
| Blood.Group | discreta | Gruppo sanguigno | 13.80 | 1.84 | 11.00 | 18.00 |
| Pulse.rate | continua | Frequenza cardiaca(bpm) | 73.46 | 2.69 | 70.00 | 82.00 |
| RR | continua | Intervallo RR(respiri/min) | 19.24 | 1.69 | 16.00 | 28.00 |
| Hb | continua | Emoglobina(g/dl) | 11.16 | 0.87 | 8.50 | 14.80 |
| Cycle | binaria | Ciclo mestruale irregolare (0 = No, 1 = Sì) | 0.28 | 0.45 | 0.00 | 1.00 |
| Cycle.length | discreta | Lunghezza della mestruazione(giorni) | 4.94 | 1.49 | 0.00 | 12.00 |
| Marriage.status | continua | Anni di matrimonio | 7.68 | 4.80 | 0.00 | 30.00 |
| Pregnant | binaria | Essere incinta (0 = No, 1 = Sì) | 0.38 | 0.49 | 0.00 | 1.00 |
| Aborptions | discreta | Numero di aborti | 0.29 | 0.69 | 0.00 | 5.00 |
| I.beta.HCG | continua | Prima betaHCG(mIU/mL) | 664.55 | 3348.92 | 1.30 | 32460.97 |
| II.beta.HCG | continua | Seconda betaHCG(mIU/mL) | 238.23 | 1603.83 | 0.99 | 25000.00 |
| FSH | continua | Ormone follicolo-stimolante(mIU/mL) | 14.60 | 217.02 | 0.21 | 5052.00 |
| LH | continua | Ormone luteinizzante(mIU/mL) | 6.47 | 86.67 | 0.02 | 2018.00 |
| LHFSH | continua | Rapporto FSH/LH | 0.55 | 0.45 | 0.00 | 4.35 |
| Hip | continua | Circonferenza dei fianchi(pollici) | 37.99 | 3.97 | 26.00 | 48.00 |
| Waist | continua | Circonferenza della vita(pollici) | 33.84 | 3.60 | 24.00 | 47.00 |
| WHR | continua | Rapporto vita/fianchi | 0.89 | 0.05 | 0.76 | 0.98 |
| TSH | continua | Ormone tireostimolante(mIU/mL) | 2.98 | 3.76 | 0.04 | 65.00 |
| AMH | continua | Ormone anti-mulleriano(ng/mL) | 5.62 | 5.88 | 0.10 | 66.00 |
| PRL | continua | Prolattina(ng/mL) | 24.32 | 14.97 | 0.40 | 128.24 |
| Vit.D3 | continua | Vitamina D3(ng/mL) | 49.92 | 346.21 | 0.00 | 6014.66 |
| PRG | continua | Progesterone(ng/mL) | 0.61 | 3.81 | 0.05 | 85.00 |
| RBS | continua | Glicemia (mg/dl) | 99.84 | 18.56 | 60.00 | 350.00 |
| Weight.gain | binaria | Aumento di peso (0 = No, 1 = Sì) | 0.38 | 0.49 | 0.00 | 1.00 |
| Hair.growth | binaria | Irsutismo (0 = No, 1 = Sì) | 0.27 | 0.45 | 0.00 | 1.00 |
| Skin.darkening | binaria | Iperpigmentazione (0 = No, 1 = Sì) | 0.31 | 0.46 | 0.00 | 1.00 |
| Hair.loss | binaria | Alopecia (0 = No, 1 = Sì) | 0.45 | 0.50 | 0.00 | 1.00 |
| Pimples | binaria | Acne (0 = No, 1 = Sì) | 0.49 | 0.50 | 0.00 | 1.00 |
| Fast.food | binaria | Dieta non salutare (0 = No, 1 = Sì) | 0.52 | 0.50 | 0.00 | 1.00 |
| Reg.Exercise | binaria | Esercizio regolare (0 = No, 1 = Sì) | 0.25 | 0.43 | 0.00 | 1.00 |
| BP.systolic | continua | Pressione sistolica(mmHg) | 114.66 | 7.38 | 12.00 | 140.00 |
| BP.diastolic | continua | Pressione diastolica (mmHg) | 76.93 | 5.57 | 8.00 | 100.00 |
| FolliclesL | discreta | Numero di follicoli nell'ovaio sinistro | 6.13 | 4.23 | 0.00 | 22.00 |
| FolliclesR | discreta | Numero di follicoli nell'ovaio destro | 6.64 | 4.44 | 0.00 | 20.00 |
| L.size | continua | Dimensione media dei follicoli nell'ovaio sinistro | 15.02 | 3.57 | 0.00 | 24.00 |

| | Tipo | Descrizione | mean | sd | min | max |
|---|---|---|---|---|---|---|
| R.size | continua | Dimensione media dei follicoli nell'ovaio destro | 15.45 | 3.32 | 0.00 | 24.00 |
| Endometrium | continua | Dimensione dell'endometrio(mm) | 8.48 | 2.17 | 0.00 | 18.00 |

Nel corso dell'analisi, ho aggiunto alcune variabili derivate da quelle già presenti nel dataset. Ne riporto una breve descrizione di seguito.

- Follicles: è una variabile discreta che rappresenta il massimo tra il numero di follicoli nell'ovaia sinistra e nell'ovaia destra. La scelta di utilizzare il massimo, piuttosto che la media, permette di evidenziare l'ovaio con la maggiore disfunzione, evitando che eventuali anomalie in un lato vengano attenuate.
- F.size: è una variabile continua che rappresenta il minimo fra la dimensione media dei follicoli nell'ovaio sinistro e nell'ovaio destro. La scelta di utilizzare il minimo, piuttosto che la media, permette di focalizzarsi sull'ovaio con la dimensione follicolare più ridotta, evitando che la maggiore dimensione di un lato possa mascherare eventuali anomalie in un altro, rendendo così più evidente la gravità della disfunzione ovarica.
- PCOStype: è una variabile categoriale di tre categorie (*None*, *Anovulatory*, *Ovulatory*), definite secondo il livello dell'ormone AMH.La categorizzazione si basa sul fatto che, nelle donne con PCOS, l'AMH è spesso elevato, poiché riflette la presenza di follicoli ovarici immaturi. Un livello di AMH superiore a 5 ng/mL indica una condizione di anovulazione (mancanza di ovulazione), che è tipica delle forme più gravi di PCOS (categoria *Anovulatory*). Al contrario, livelli inferiori a 5 ng/mL suggeriscono una situazione di ovulazione regolare, pur in presenza della sindrome (categoria *Ovulatory*). La categoria *None* indica l'assenza di PCOS. PCOStype è la variabile target nell'analisi, e la sua introduzione ha determinato l'eliminazione della variabile *PCOS* dal dataset, poiché ridondante.

Sono stati rilevati dei valori mancanti in alcune variabili. Per garantire la completezza dei dati, i valori mancanti sono stati sostituiti con la mediana della variabile corrispondente, una scelta basata sulla robustezza della mediana in presenza di valori estremi.

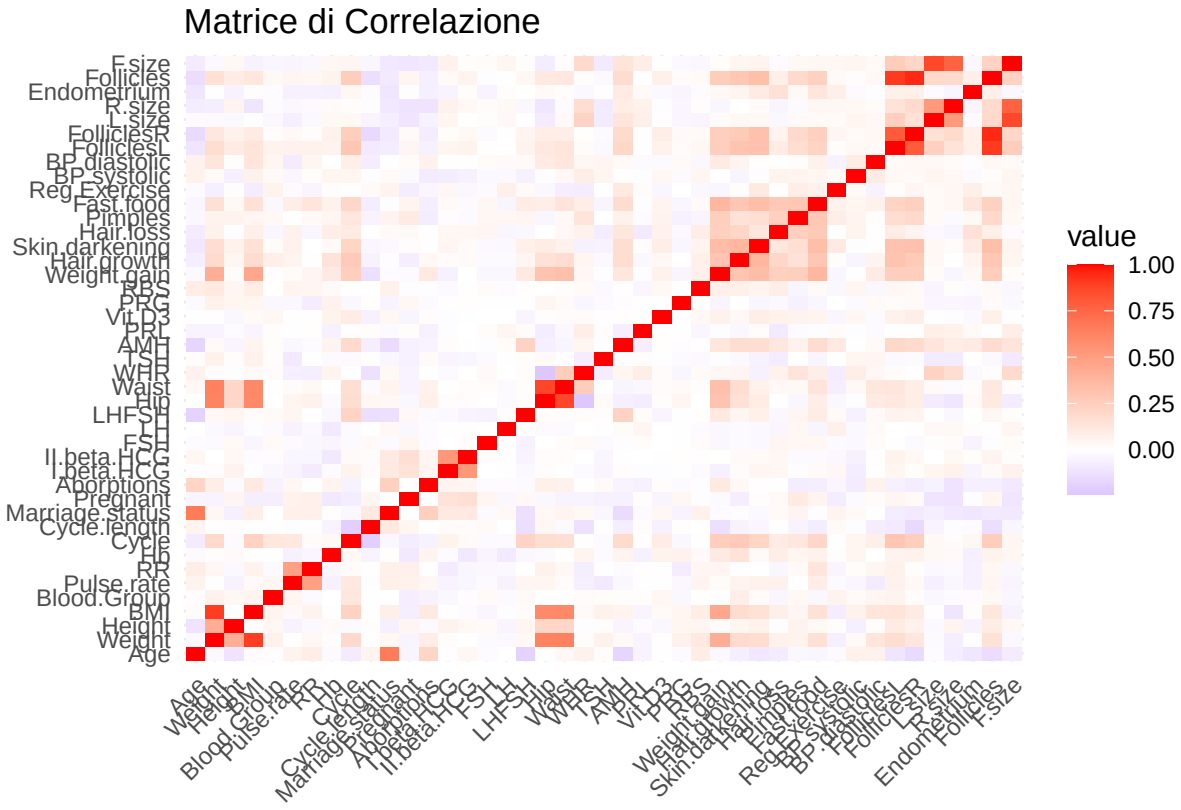
Per garantire la coerenza nella scala delle variabili e migliorare le performance del modello, ho applicato la normalizzazione min-max a tutte le variabili numeriche del dataset, riportandole nell'intervallo [0, 1]:

$$x' = \frac{x - min}{max - min}.$$

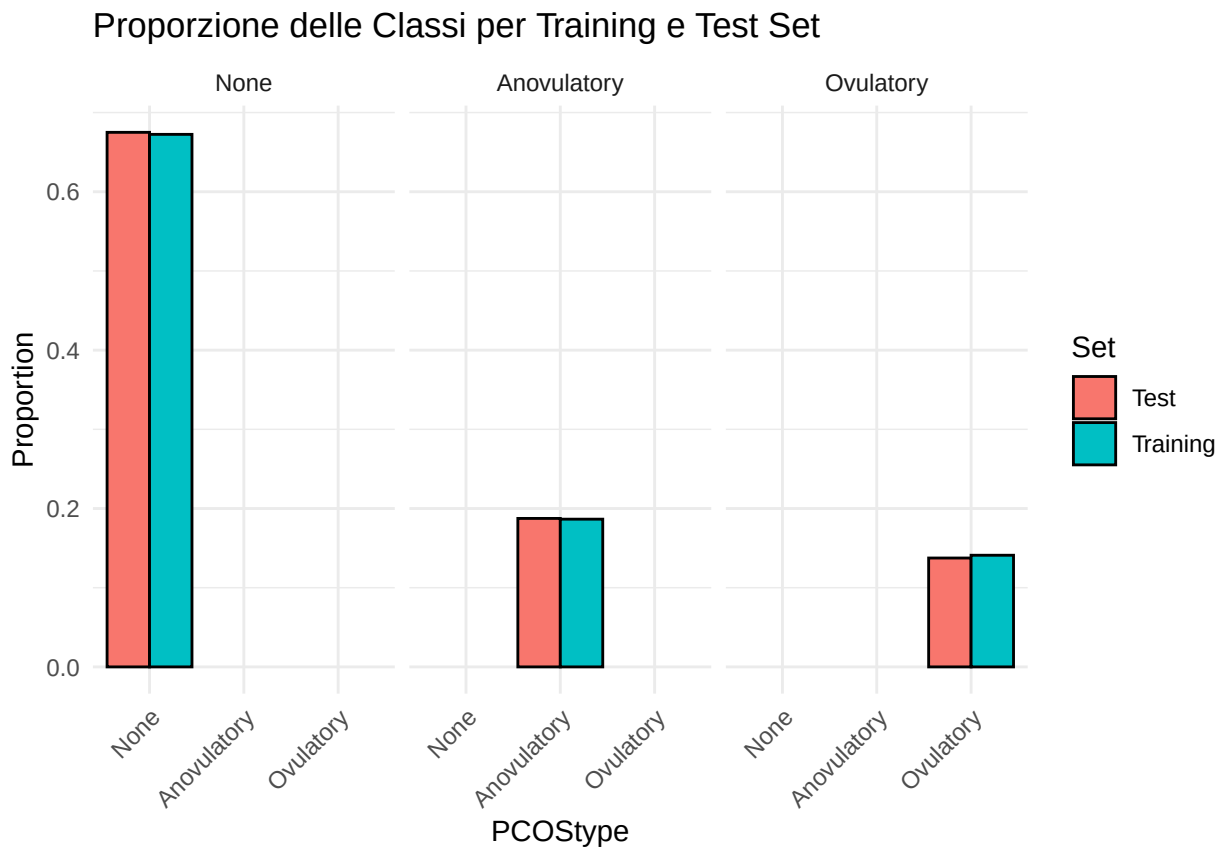
Alcune variabili sono state eliminate dal dataset per evitare ridondanze o per la presenza di correlazioni elevate con altre variabili. In particolare, sono state rimosse le seguenti:

- *Weight* e *Height*, mantenendo il sintentico indicatore *BMI*;
- *Marriage Status*, poiché altamente correlata ad *Age*;
- *FSH* e *LH*, mantenendo il più informativo rapporto *FSH/LH*, che fornisce una misura più adeguata dell'equilibro ormonale;

- *Waist* e *Hip*, in favore del più rappresentativo rapporto WHR (Waist to Hip Ratio), che sintetizza meglio la distribuzione del grasso corporeo;
- *L.size* e *R.size*, poiché direttamente legate a *F.size*, che rappresenta la minima dimensione tra i follicoli dell'ovaio sinistro e dell'ovaio destro;
- *FolliclesL* e *FolliclesR* poiché direttamente legate a *Follicles*, che rappresenta il numero massimo di follicoli tra le due ovaie.




Per allenare il modello, ho suddiviso il dataset in un training set (85%) e un test set (15%). La scelta di stabilire una proporzione molto sbilanciata verso il training set è stata dettata dalla piccola dimensione del dataset utilizzato. I due sottoinsiemi sono stati costruiti facendo sì che le proporzioni delle diverse classi della variabile target *PCOStype* fossero le medesime.




# 4 Costruzione di un classificatore mediante le reti neurali

## 4.1 Reti neurali con un solo strato nascosto

Inizialmente, ho sviluppato delle reti neurali per mezzo della libreria *neuralnet*, che permette di costruire reti con un solo strato nascosto. Ho testato tutte le possibili combinazioni dei seguenti parametri:

- *treshold* (il valore soglia utilizzato durante la minimizzazione della funzione di errore): 0.5, 0.1, 0.001;
- numero di neuroni nello strato nascosto: 16, 8, 5;
- *decay*: 1, 0.5, 0.1, 0.001.

In particolare, il parametro *decay* rappresenta un termine di penalizzazione aggiunto alla funzione di perdita, che ha lo scopo di penalizzare i pesi troppo grandi. Questo aiuta a ridurre il rischio di overfitting, un problema particolarmente rilevante quando si lavora con un dataset di dimensioni ridotte, come in questo caso.

Al termine della fase di allenamento, sono state effettuate le previsioni sul test set ed è stata selezionata la rete alla quale corrisponde il più alto valore di F1-score:

$$F1score = \frac{2 \cdot precision \cdot recall}{precision + recall},$$

dove la *precision* è la percentuale di previsioni positive corrette sul totale di tutte le previsioni positive

$$precision = \frac{TP}{TP + FP}$$

e la *recall* è la percentuale di veri positivi che sono stati correttamente identificati rispetto al totale dei positivi

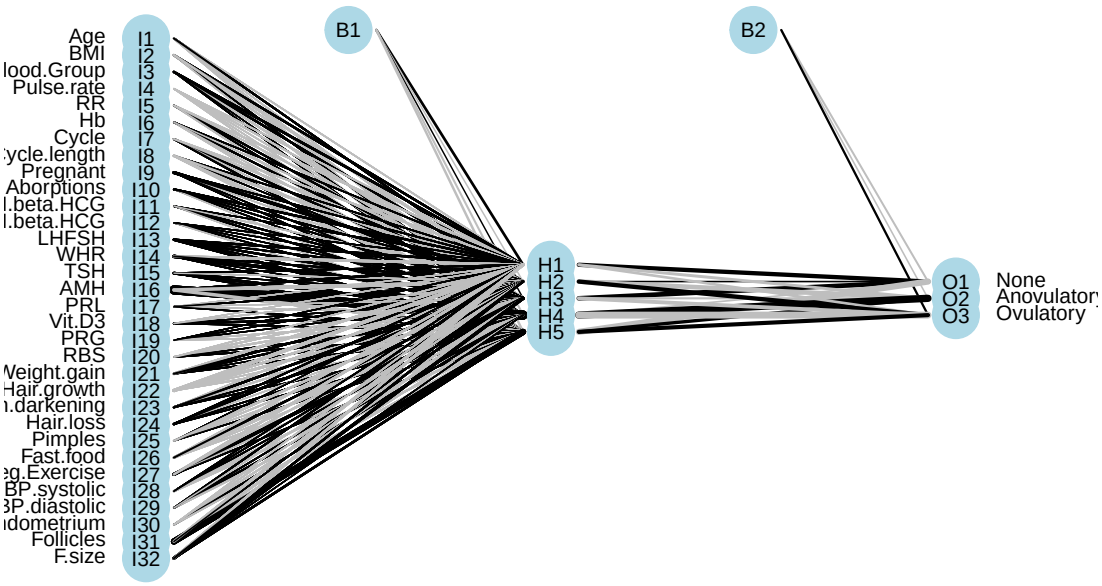
$$recall = \frac{TP}{TP + FN}.$$

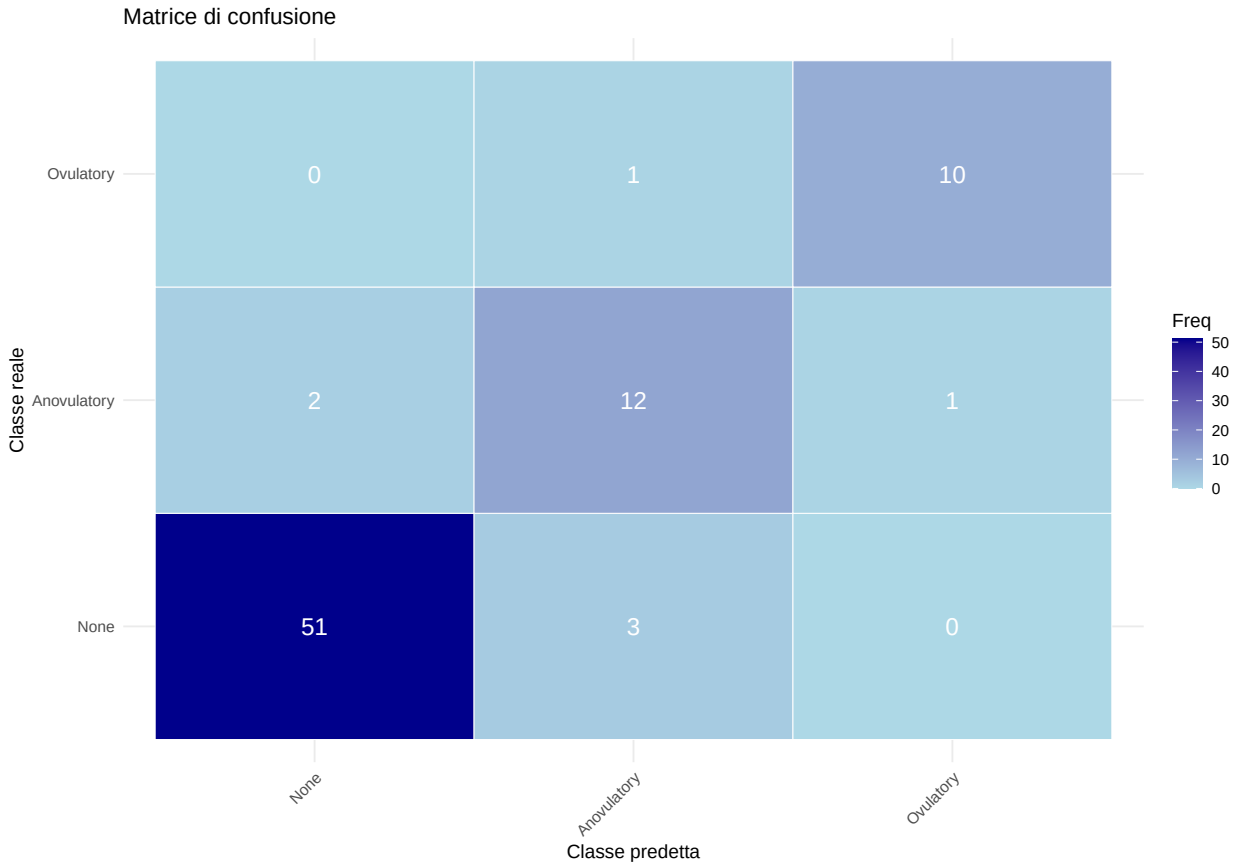
Per un problema di classificazione multi-classe, l'F1-Score complessivo può essere ottenuto attraverso diverse metodologie. Una delle più comuni consiste nel calcolarlo come una sintesi delle misure F1-Score per ciascuna classe, pesata in base al numero di osservazioni per classe. In questo caso, ogni F1-Score di classe contribuisce alla media finale proporzionalmente alla frequenza di quella classe nel dataset. Questa metrica è stata preferita in quanto è particolarmente adatta a gestire situazioni in cui le classi della variabile target molto sbilanciate, come in questo caso, in cui le proporzioni dei vari livelli di *PCOStype* sono circa 67.25%, 18.66% e 14.10%. La rete selezionata è stata poi valutata anche in termini di *accuratezza*, definita come la percentuale di previsioni corrette rispetto al totale delle osservazioni.

La rete neurale ottimale individuata da questa procedura ha mostrato un valore di F1score pari superiore al 91%. I parametri ottimali, le metriche di valutazione, la struttura della rete e la matrice di confusione sono mostrati di seguito.

Table 2: Best Model Parameters

| Metric | Value |
|---|---|
| F1 Score | 0.9136 |
| Accuracy | 0.9125 |
| Threshold | 0.5000 |
| Neurons in Hidden Layer | 5.0000 |
| Decay | 0.1000 |


Age
BMI
lood.Group
Pulse.rate
RR
Hb
Cycle
ycle.length
Pregnant
Aborptions
l.beta.HCG
l.beta.HCG
LHFSH
WHR
TSH
AMH
PRL
Vit.D3
PRG
RBS
Weight.gain
Hair.growth
n.darkening
Hair.loss
Pimples
Fast.food
g.Exercise
BP.systolic
BP.diastolic
ndometrium
Follicles
F.size
I1
I2
I3
I4
I5
I6
I7
I8
I9
I10
I11
I12
I13
I14
I15
I16
I17
I18
I19
I20
I21
I22
I23
I24
I25
I26
I27
I28
I29
I30
I31
I32
B1
B2
H1
H2
H3
H4
H5
O1
O2
O3
None
Anovulatory
Ovulatory




## 4.2 Reti neurali con più di uno strato nascosto

Per consentire la costruzione di reti neurali più profonde, ho utilizzato la libreria RSNNS, che permette di implementare architetture con più strati nascosti.

Ho testato due configurazioni:

- una rete neurale con due strati nascosti (10 e 5 neuroni, rispettivamente)
- una rete neurale con tre strati nascosti (20, 10, 5 neuroni, rispettivamente).

Nello strato di output, è stata adottata la funzione di attivazione softmax, adatta alla classificazione multi-classe.

Di seguito, sono riportati i valori di accuratezza e F1-score per ciascuna rete:

Table 3: Rete con 2 strati nascosti

| Metric | Value |
|---|---|
| F1 Score | 0.8251 |
| Accuracy | 0.825 |

| Metric | Value |
|---|---|
| Threshold | 0.01 |
| Neurons in Hidden Layer | 10, 5 |

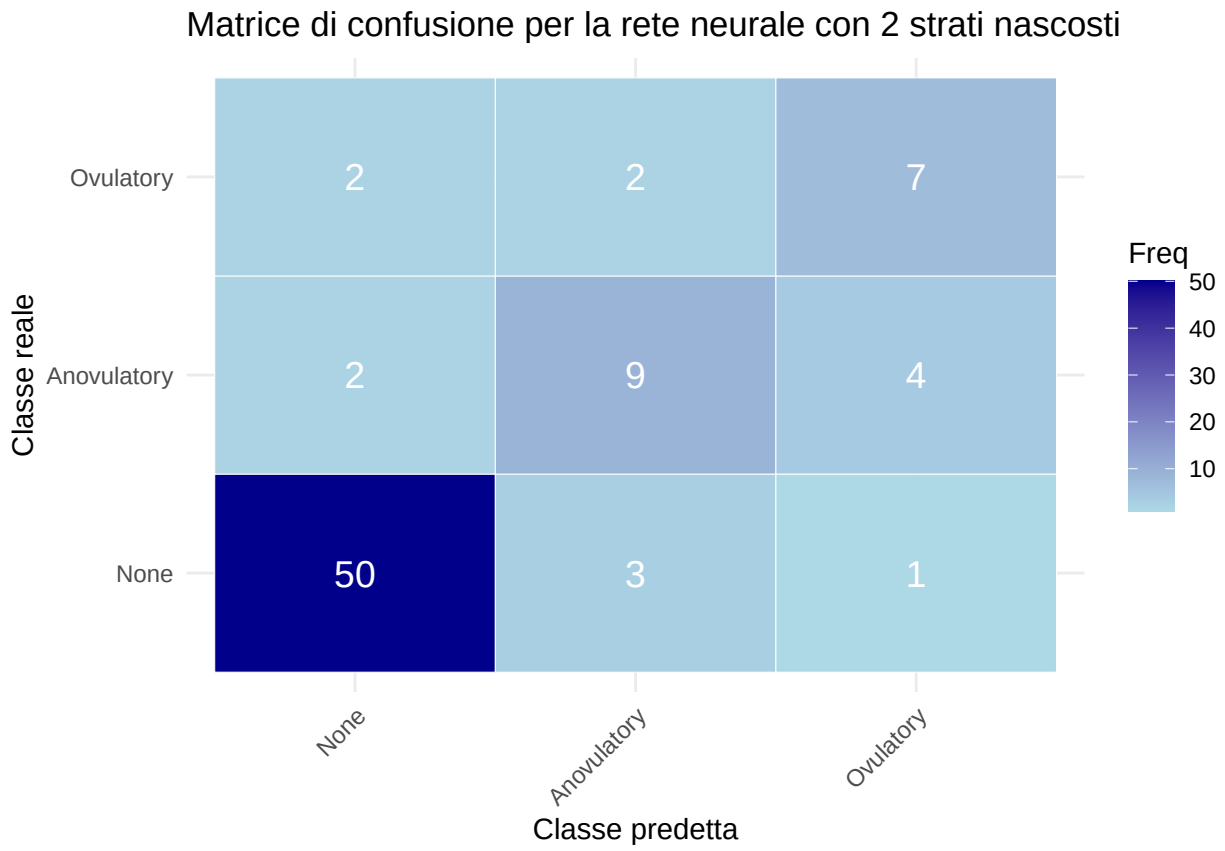
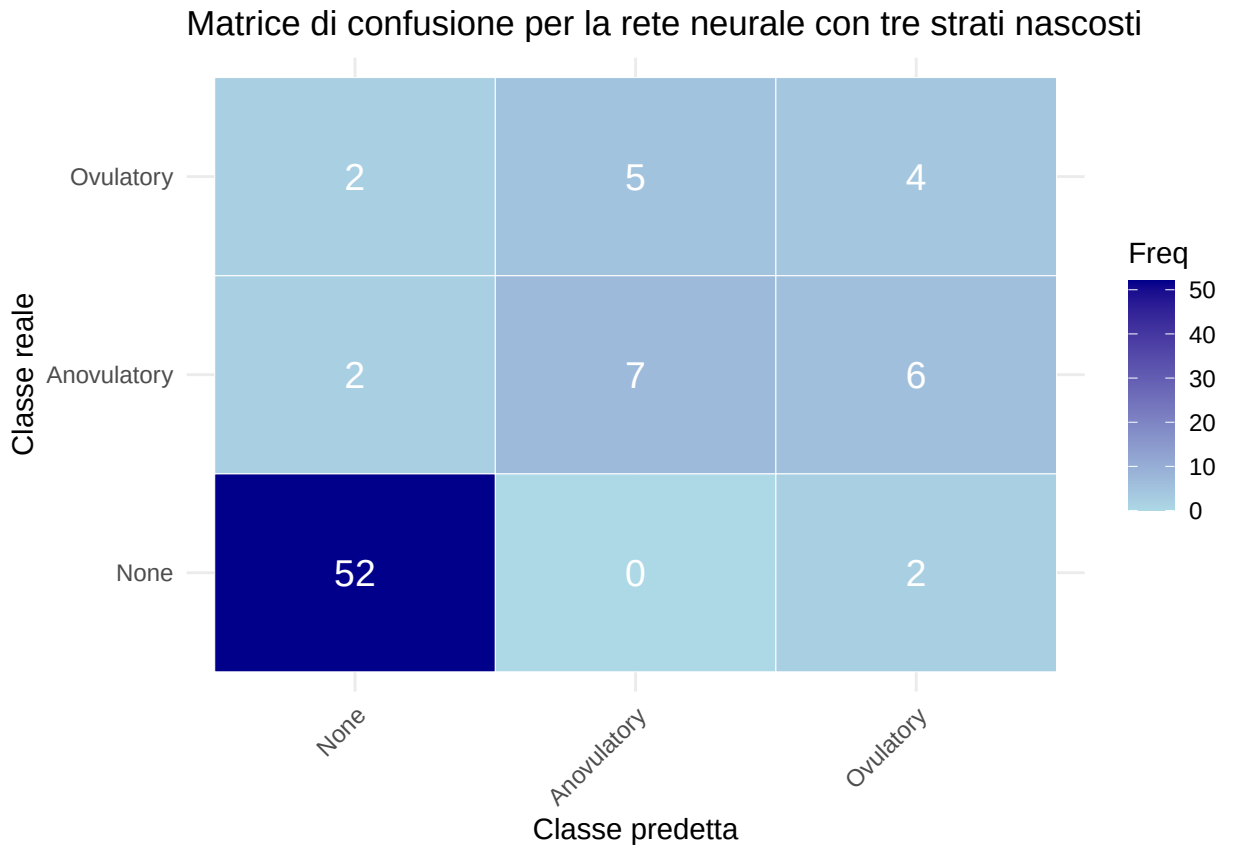



Table 4: Rete con 3 strati nascosti

| Metric | Value |
|---|---|
| F1 Score | 0.7832 |
| Accuracy | 0.7875 |
| Threshold | 0.01 |
| Neurons in Hidden Layer | 20, 10, 5 |



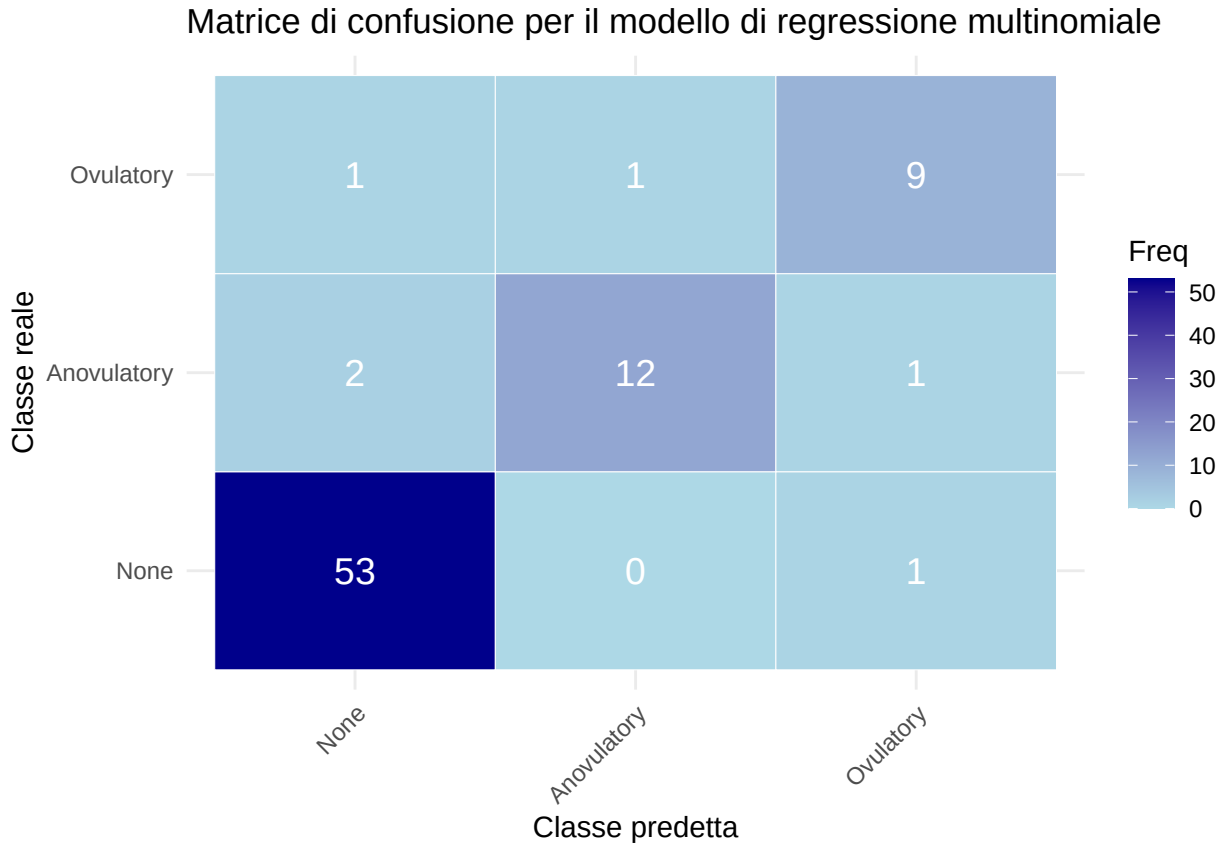

I risultati della rete con tre strati nascosti sono nettamente peggiori rispetto alle reti più semplici. Questo potrebbe essere dovuto all'overfitting: con un dataset relativamente piccolo, l'aggiunta di ulteriori strati aumenta la complessità del modello, che finisce per adattarsi troppo ai dati di training e perde la capacità di generalizzare sui dati di test.

### 4.3 Modello di regressione multinomiale

Ho deciso di testare un approccio tradizionale per la classificazione, costruendo un modello di regressione multinomiale, che rappresenta una soluzione consolidata ed efficace per i problemi di classificazione con più di due classi. Il modello ha evidenziato buone prestazioni, offrendo un valido confronto con le reti neurali utilizzate precedentemente.

Table 5: Modello di regressione multinomiale

| Metric | Value |
|---|---|
| F1 Score | 0.9237 |
| Accuracy | 0.9250 |


Matrice di confusione per il modello di regressione multinomiale
Ovulatory
Anovulatory
None
Classe reale
1
1
9
2
12
1
53
0
1
None
Anovulatory
Ovulatory
Classe predetta
Freq
50
40
30
20
10
0

### 4.4 Risultati

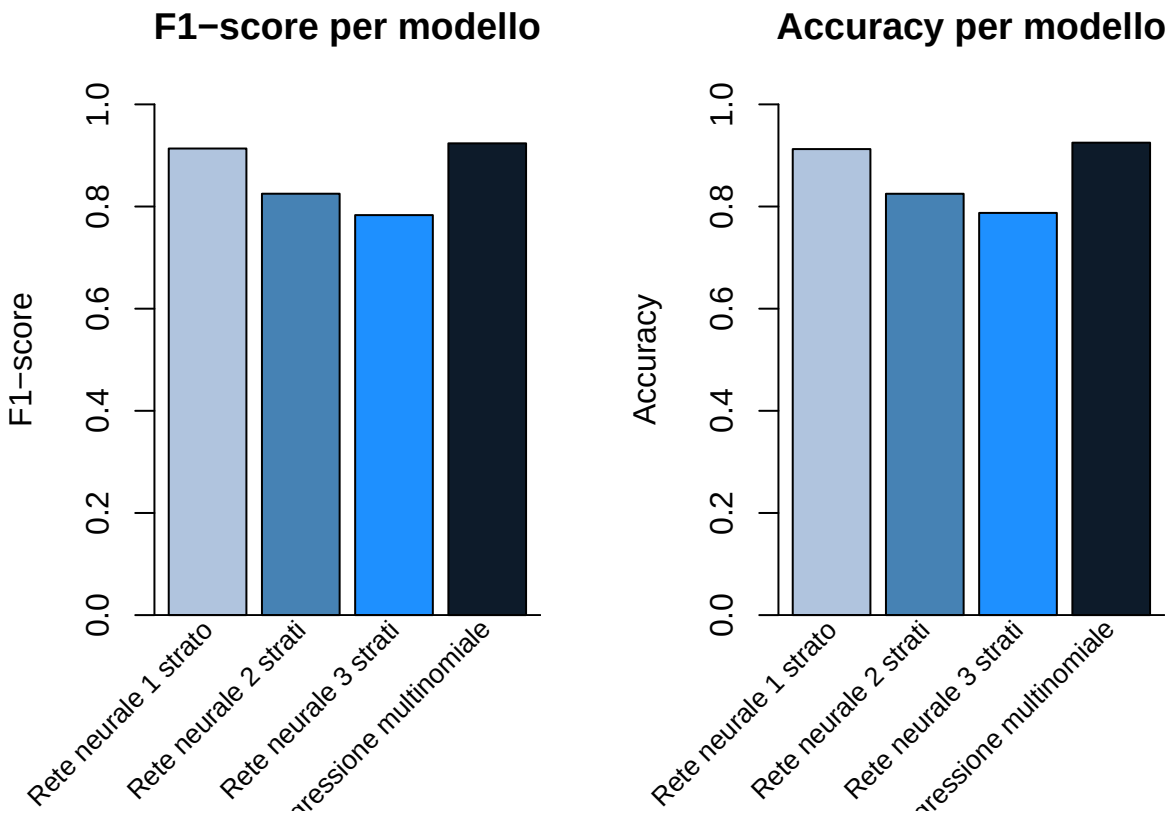



Table 6: Confronto tra i modelli in base ad Accuratezza e F1-Score

| Model | F1score | Accuracy |
|---|---|---|
| **Rete neurale 1 strato** | **0.9136** | **0.9125** |
| Rete neurale 2 strati | 0.8251 | 0.825 |
| Rete neurale 3 strati | 0.7832 | 0.7875 |
| **Regressione multinomiale** | **0.9237** | **0.925** |

I risultati mostrano che il modello di regressione multinomiale ha ottenuto le migliori performance complessive, con valori di F1-score e accuratezza leggermente superiori rispetto a tutte le reti neurali testate. In particolare, la rete neurale con un solo strato nascosto, che inizialmente sembrava essere il modello più efficace, mostra prestazioni molto simili alla regressione multinomiale, ma leggermente inferiori.

Le reti neurali più complesse, invece, con due o tre strati nascosti, registrano un calo significativo delle performance, con una riduzione evidente sia dell'F1-score che dell'accuratezza. Questo suggerisce che, per questo dataset, l'uso di modelli più sofisticati non porta necessariamente a un miglioramento delle prestazioni e che una soluzione più semplice, come la regressione multinomiale, può risultare altrettanto, se non più, efficace.

# 5 Sviluppo di una rete bayesiana

## 5.1 Categorizzazione delle variabili

Poiché si vuole sviluppare una rete bayesiana discreta, è necessario che tutte le variabili vengano categorizzate, nel caso in cui siano numeriche. Inoltre, per semplicità, ho selezionato un sottoinsieme di variabili, descritte di seguito.

- Age: è una variabile categoriale di tre categorie (*Giovane*, *Adulta*, *Matura*);
- BMIlevel: è una variabile categoriale di quattro categorie (*Underweight*, *Normal Weight*, *Overweight*, *Obese*), costruite rispettando le soglie segnalate dall'OMS [7].
- HighFHLSH: è una variabile binaria che vale 1 se la paziente ha un rapporto FH:LSH elevevato ($> 2$), 0 altrimenti.
- TSH_Category: è una variabile categoriale con tre categorie (*Ipotiroidismo*, *Normale*, *Ipertiroidismo*), dipendentemente dal livello dell'ormone tireo-stimolante (TSH).
- AMH_Category: è una variabile categoriale con tre categorie (*Bassa riserva ovarica*, *Normale*, *Alta riserva ovarica*), definite dipendentemente dal livello dell'ormone anti-mulleriano (AMH).
- Hyperglycemia: è una variabile binaria che vale 1 se la paziente presenta un elevato valore di glicemia, 0 altrimenti.
- Abnormal_Endometrium: è una variabile binaria che vale 1 se lo spessore dell'endometrio è fuori dalla norma, 0 altrimenti.
- High.Follicles: è una variabile binaria che vale 1 se il numero dei follicoli è elevato, 0 altrimenti.
- Immature.follicles: è una variabile binaria che vale 1 se i follicoli sono reputati immaturi, 0 altrimenti, dipendentemente dalla dimensione media dei follicoli.

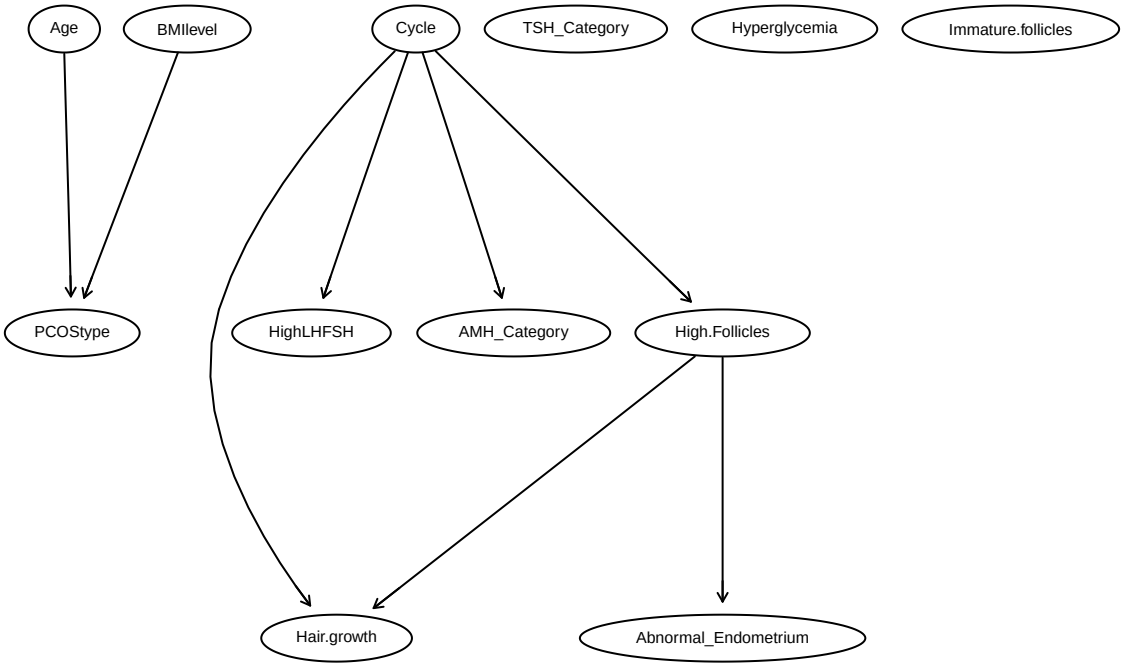
Inoltre, sono state mantenute le già definite *Cycle*, *Hair.growth* e *PCOStype*. Il dataset finale, dunque, contiene 12 variabili.

## 5.2 Costruzione del dag

Per la costruzione della struttura del DAG, ho seguito un approccio che combinasse vincoli teorici con metodi di apprendimento strutturale. In primo luogo, ho definito una blacklist per evitare che *PCOStype* fosse collegata come genitore ad altre variabili, in quanto rappresenta la variabile target del modello. Inoltre, ho impostato che nessuna variabile potesse essere genitore di *Age* e *BMIlevel*.

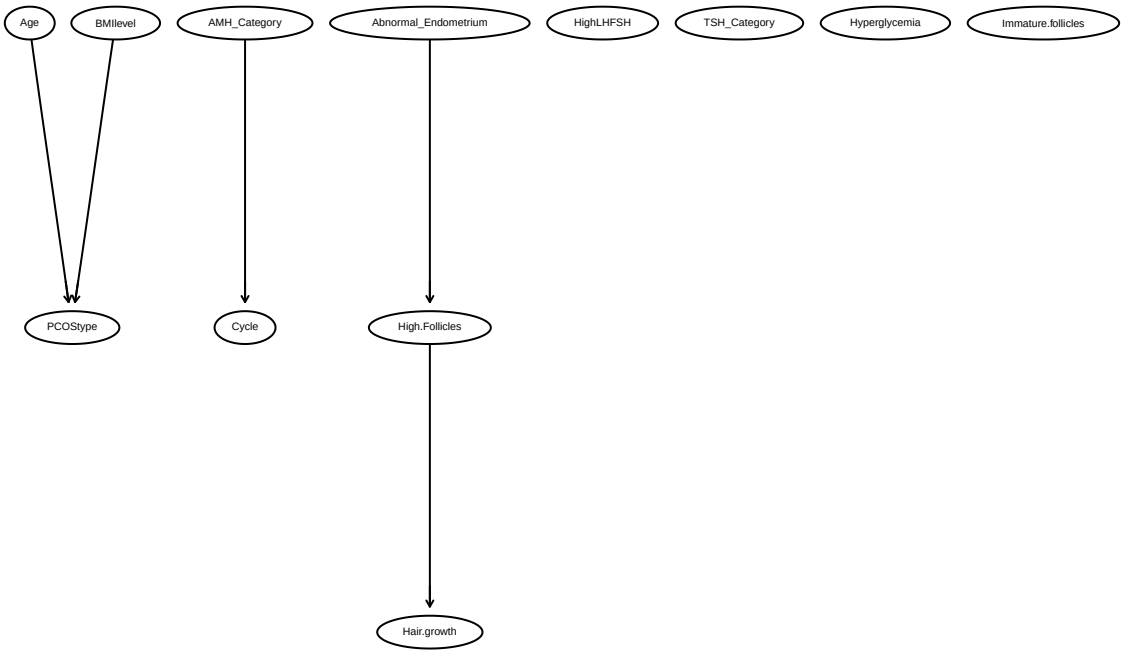
Per incorporare la conoscenza a priori, ho creato una whitelist in modo che *Age* e *BMIlevel* fossero considerati genitori diretti di *PCOStype*, come ampiamente documentato in letteratura. Successivamente, ho applicato l'algoritmo di Hill-Climbing con il criterio di selezione BIC, ottenendo così un grafo iniziale.

## Grafo iniziale

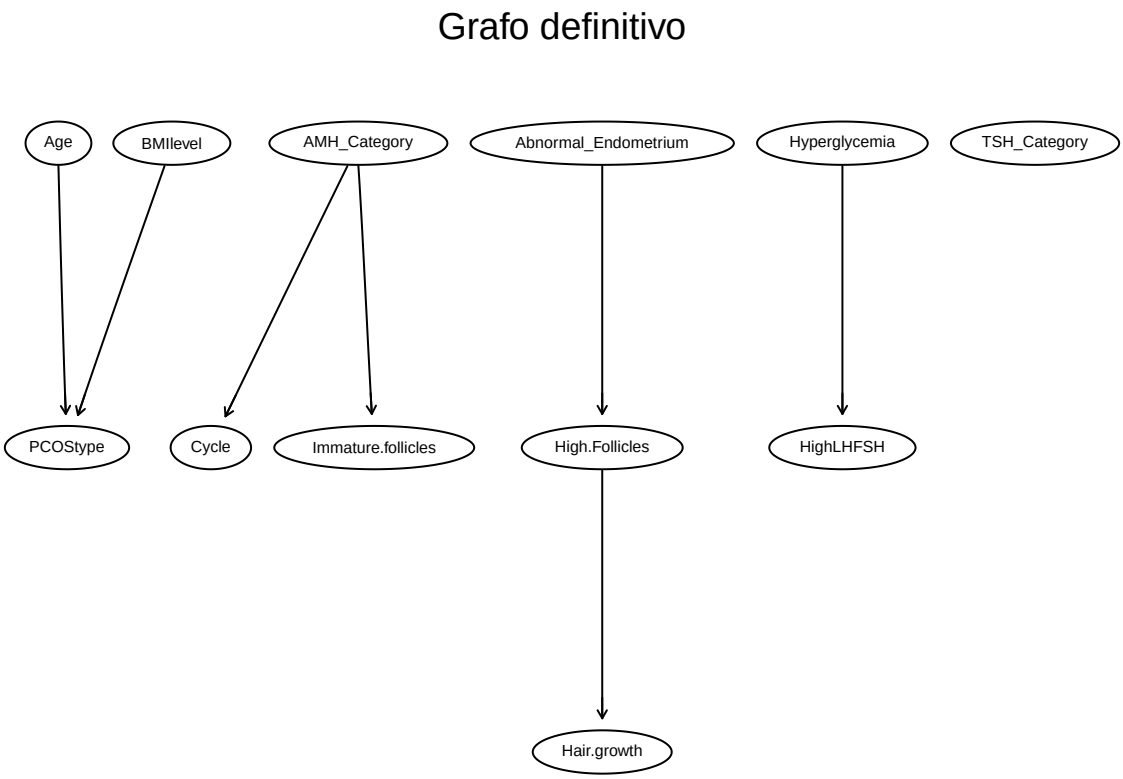



Per valutare la robustezza degli archi, ho calcolato la strength di ciascuno, e per quelli che risultavano troppo deboli, ho eseguito test di indipendenza condizionale rispetto a tutte le altre variabili. Gli archi non validi sono stati eliminati e aggiunti alla blacklist.

## Secondo grafo




Infine, ho integrato alcune conoscenze teoriche specifiche, inserendo nella whitelist archi considerati cruciali, come la connessione tra *Hyperglycemia* e *HighLHFSH*, e quella tra *AMH_Category* e *Immature.Follicles.* Questo ha portato alla costruzione del grafo finale, che riflette sia le evidenze empiriche che le connessioni teoriche necessarie per il modello.




Nonostante le variazioni nei valori di BIC tra i diversi grafi, i risultati si stabilizzano attorno a valori simili. Per questo motivo, ritengo sia stato raggiunto un buon equilibrio tra l'integrazione delle conoscenze teoriche del fenomeno e l'adattamento ai dati, garantendo una rappresentazione coerente e robusta del sistema studiato.

Table 7: Valori di BIC per i diversi grafi

| Grafo | BIC |
|---|---|
| Grafo iniziale | -3517.791 |
| Grafo 2 | -3530.323 |
| Grafo finale | -3538.893 |

L'espressione generale del modello, dunque, risulta essere la seguente:

```
## [Age] [BMIlevel] [TSH_Category] [AMH_Category] [Hyperglycemia] [Abnormal_Endometrium]
##  [Cycle | AMH_Category] [PCOStype | Age:BMIlevel] [HighLHFSH | Hyperglycemia]
##  [High.Follicles | Abnormal_Endometrium] [Immature.follicles | AMH_Category]
##  [Hair.growth |   High.Follicles]
```

## 5.3 Stima delle distribuzioni di probabilità

Per la stima delle distribuzioni di probabilità, è stato utilizzato il metodo bayesiano, che combina le frequenze campionarie con una distribuzione uniforme. Questa scelta permette di ottenere stime più stabili rispetto a un approccio puramente frequentista, soprattutto in presenza di un dataset di dimensioni ridotte.

In particolare, le probabilità sono state calcolate come media pesata tra la distribuzione uniforme delle unità campionarie ISS (di numerosità pari a 10) e le frequenze osservate nei dati.

## 5.4 Inferenza

Una volta stimati i parametri della rete bayesiana, è possibile procedere alla fase di inferenza, che consente di ottenere informazioni sul fenomeno modellato interrogando la rete. In particolare, sono state eseguite sia query prognostiche, per valutare la probabilità di un esito dato un insieme di condizioni osservate, sia query diagnostiche, per inferire la probabilità di una causa nota la manifestazione di un determinato effetto.

### Rapporto tra il BMI e la PCOS

Ci si chiede, a questo punto, se il BMI ha un impatto sulla probabilità di sviluppare PCOS. Per rispondere a questa domanda, è necessario la distribuzione generale di PCOS con quella condizionata ai diversi livelli di BMI.

```
querygrain(junction_tree, nodes = c("PCOStype", "BMIlevel"), type = "conditional")
```

```
##              BMIlevel
## PCOStype      Underweight Normal Weight Overweight     Obese
##   Anovulatory   0.2056273     0.1340243  0.2505587 0.2975929
##   None          0.6759058     0.7365834  0.6065725 0.4356831
##   Ovulatory     0.1184668     0.1293922  0.1428688 0.2667240
```

```
querygrain(junction_tree, nodes = c("PCOStype"), type = "marginal")
```

```
## $PCOStype
## PCOStype
## Anovulatory        None   Ovulatory
##   0.1919785   0.6634907   0.1445308
```

Dall'osservazione della distribuzione marginale di PCOStype, è possibile concludere che, nella popolazione generale, la maggior parte delle donne non presenta la sindrome, con una probabilità del 66.3%. Tra coloro che invece ne soffrono, il 19.2% ha la forma anovulatoria, mentre il 14.5% ha la forma ovulatoria.

Esaminando la distribuzione condizionata al BMI, emergono differenze interessanti. Le donne normopeso sembrano essere quelle più protette dalla PCOS: la probabilità che abbiano la forma anovulatoria scende

al 13.4%, un valore decisamente inferiore rispetto alla media della popolazione, e la percentuale di coloro che non hanno PCOS sale al 73.6%, suggerendo che il normopeso possa effettivamente ridurre il rischio complessivo di sviluppare la sindrome. Al contrario, tra le donne sovrappeso e obese si osserva un forte aumento del rischio. La probabilità di avere PCOS anovulatoria cresce progressivamente con l'aumento del peso, raggiungendo quasi il 30% tra le donne obese. Anche la forma ovulatoria segue un andamento simile: mentre nelle donne normopeso la sua incidenza rimane in linea con la distribuzione generale, tra le donne obese si registra un netto aumento, con una probabilità che quasi raddoppia rispetto alla media.

Alla luce di questi risultati, è possibile concludere che il BMI è un fattore determinante nella predisposizione alla PCOS. Il fatto che la distribuzione condizionata al BMI sia così diversa da quella marginale è un'indicazione forte del fatto che il peso corporeo deve essere considerato un elemento chiave nella valutazione del rischio di PCOS.

Su questo argomento, è possibile condurre anche un'analisi di tipo diagnostico, cioè verificare la distribuzione di probabilità di *BMIlevel* noto, ad esempio, che la paziente soffre della forma anovulatoria di PCOS.

```
querygrain(setEvidence(junction_tree, nodes = "PCOStype", states = "Anovulatory"),
           nodes = "BMIlevel", type = "conditional")
```

```
## BMIlevel
##   Underweight Normal Weight    Overweight         Obese
##    0.07484063    0.35539583    0.44175768    0.12800586
```

```
querygrain(junction_tree, nodes = c("PCOStype"), type = "marginal")
```

```
## $PCOStype
## PCOStype
## Anovulatory        None   Ovulatory
##   0.1919785   0.6634907   0.1445308
```

I dati suggeriscono che la maggior parte delle donne con questa forma di PCOS ha un indice di massa corporea elevato. Il gruppo più rappresentato, infatti, è quello delle donne sovrappeso, che costituiscono il 44.2% del totale.

Un altro aspetto interessante è che la percentuale di donne obese (12.8%) è inferiore a quella delle sovrappeso, il che potrebbe suggerire che, sebbene l'obesità sia un fattore di rischio noto per la PCOS, il sovrappeso sia più comune tra le pazienti con questa forma della sindrome.

D'altra parte, la probabilità di essere sottopeso tra le donne con PCOS Anovulatoria è solo del 7.5%, un valore molto basso. Questo conferma che il sottopeso non è una condizione tipica tra chi soffre di questa variante della PCOS, e suggerisce che l'eccesso di peso potrebbe giocare un ruolo più rilevante nello sviluppo di questa condizione.

Dalle analisi condotte rispetto alla forma ovulatoria della sindrome (considerata meno grave), si evidenzia che la categoria più rappresentata tra le donne con questo tipo di PCOS è quella normopeso, con una

percentuale del 45.6%. Tale valore è decisamente più alto rispetto a quanto osservato tra le donne con PCOS anovulatoria, tra le quali il normopeso rappresentava solo il 35.5%. Questo suggerisce che la PCOS ovulatoria non è così fortemente associata all'eccesso di peso come la forma anovulatoria e che molte pazienti con questa variante della sindrome rientrano in una fascia di BMI considerata normale.

```
querygrain(setEvidence(junction_tree, nodes = "PCOStype", states = "Ovulatory"),
           nodes = "BMIlevel", type = "conditional")
```

```
## BMIlevel
##   Underweight Normal Weight    Overweight         Obese
##    0.05727242    0.45575256    0.33458327    0.15239175
```

```
querygrain(junction_tree, nodes = c("PCOStype"), type = "marginal")
```

```
## $PCOStype
## PCOStype
## Anovulatory        None   Ovulatory
##   0.1919785   0.6634907   0.1445308
```

Anche il gruppo delle donne sovrappeso, pur essendo ancora numeroso (33.5%), è meno rappresentato rispetto a quello delle donne con PCOS anovulatoria, dove superava il 44%.

L'obesità, invece, si mantiene su livelli simili: il 15.2% delle donne con PCOS ovulatoria è obeso, una percentuale leggermente superiore a quella osservata tra le donne con PCOS anovulatoria (12,8%). Questo suggerisce che, sebbene l'eccesso di peso sia un fattore importante nella PCOS, le donne obese possono sviluppare sia la variante anovulatoria che quella ovulatoria, senza una prevalenza netta per una delle due.

La percentuale di donne con BMI molto basso suggerisce che il sottopeso non è una condizione tipica tra le pazienti con PCOS, indipendentemente dalla variante della sindrome.

In sintesi, questi risultati evidenziano una distinzione tra le due varianti della PCOS in relazione al BMI. La PCOS ovulatoria è più frequente tra le donne normopeso rispetto alla forma anovulatoria, mentre il sovrappeso, pur essendo presente, ha una minore incidenza. Questo potrebbe indicare che il BMI gioca un ruolo più forte nello sviluppo della PCOS anovulatoria, mentre la PCOS ovulatoria può presentarsi in una gamma più ampia di BMI, con una distribuzione più vicina alla popolazione generale.

Dal punto di vista clinico, questi risultati suggeriscono che il controllo del peso potrebbe essere una strategia particolarmente importante per le pazienti con PCOS anovulatoria, mentre nelle donne con PCOS ovulatoria il BMI potrebbe non essere un fattore altrettanto determinante. Tuttavia, la presenza di un numero non trascurabile di donne in sovrappeso e obese in entrambe le varianti della sindrome rafforza l'idea che il peso corporeo sia comunque un elemento chiave da considerare nella gestione della PCOS.

**Influenza dell'età sulla PCOS**

Per verificare se l'appartenenza a una determinata fascia d'età modifica la distribuzione delle diverse forme di PCOS, si devono confrontare la distribuzione marginale della PCOS e quella condizionata all'età.

```
querygrain(junction_tree, nodes = c("PCOStype", "Age"), type = "conditional")
```

```
##              Age
## PCOStype          Adulta   Giovane     Matura
##   Anovulatory 0.1270511 0.2858893 0.15299870
##   None         0.7133605 0.5702020 0.78856507
##   Ovulatory    0.1595884 0.1439087 0.05843623
```

```
querygrain(junction_tree, nodes = c("PCOStype"), type = "marginal")
```

```
## $PCOStype
## PCOStype
## Anovulatory        None   Ovulatory
##   0.1919785   0.6634907   0.1445308
```

Le donne più giovani mostrano una probabilità più elevata di avere la forma anovulatoria della PCOS rispetto alla media della popolazione. Infatti, tra le giovani, la percentuale di PCOS anovulatoria sale al 28.6%, molto più alta rispetto alla media generale del 19.2%. Questo aumento è accompagnato da una riduzione della percentuale di donne senza PCOS, che scende al 57%, suggerendo che la giovane età possa rappresentare un fattore di rischio per lo sviluppo della sindrome, specialmente nella forma anovulatoria.

Al contrario, le donne adulte sembrano essere soggette a un rischio inferiore. La probabilità di PCOS anovulatoria scende al 12.7%, ben al di sotto della media della popolazione. Parallelamente, la percentuale di donne che non soffrono di PCOS sale al 71.3%, un dato che suggerisce che l'età adulta possa essere associata a un minor rischio di sviluppare la sindrome rispetto alle fasce più giovani. Per quanto riguarda la forma ovulatoria della PCOS, la sua distribuzione nelle donne adulte è in linea con la media generale, indicando che l'età potrebbe non avere un impatto significativo su questa variante specifica della sindrome.

Nelle donne più mature, con un'età pari o superiore ai 40 anni, la probabilità di non avere PCOS è ancora più alta, raggiungendo il 78.9%. Inoltre, la PCOS anovulatoria è meno comune rispetto alla media, con una probabilità del 15.3%, mentre la PCOS ovulatoria è decisamente meno frequente in questa fascia d'età, con una probabilità di appena il 5.8%, molto inferiore rispetto alla media del 14.5%. Questo dato suggerisce che con l'avanzare dell'età, il rischio di PCOS in generale, e in particolare della sua forma ovulatoria, tende a diminuire.

**Squilibri ormonali e iperglicemia**

Questa analisi è utile per comprendere la relazione tra iperglicemia e livelli elevati di LH rispetto a FSH, un aspetto cruciale nella sindrome dell'ovaio policistico.

```r
querygrain(setEvidence(junction_tree, nodes = "Hyperglycemia", states = "1"),
           nodes = "HighLHFSH", type = "conditional")
```

```
## HighLHFSH
##         0         1
## 0.8214286 0.1785714
```

```r
querygrain(junction_tree, nodes = c("HighLHFSH"), type = "marginal")
```

```
## $HighLHFSH
## HighLHFSH
##          0          1
## 0.97640653 0.02359347
```

Osservando la distribuzione marginale di HighLHFSH, che rappresenta la probabilità complessiva di avere un rapporto LH/FSH alterato senza condizionare su alcuna variabile, emerge un dato interessante: nella popolazione generale, la probabilità che una donna presenti questo squilibrio ormonale è estremamente bassa, appena il 2.3%.

Quando però si introduce l'evidenza che la paziente soffre di iperglicemia, il quadro cambia in modo significativo. Tra le donne con glicemia alta, la probabilità di avere un rapporto LH/FSH alterato sale al 17.9%. Questo indica che la presenza di iperglicemia è fortemente associata a uno squilibrio ormonale tra LH e FSH, rafforzando l'ipotesi che l'insulino-resistenza possa essere un meccanismo chiave nel determinare alterazioni endocrine nelle pazienti con PCOS.

Un aspetto altrettanto rilevante di questa osservazione è che, sebbene il rischio di avere un rapporto ormonale elevato aumenti nelle donne con iperglicemia, la maggior parte delle pazienti continua comunque ad avere un rapporto LH/FSH nella norma. Infatti, anche tra le donne con iperglicemia, l'82.1% mantiene livelli ormonali equilibrati, il che significa che l'iperglicemia non è l'unico fattore coinvolto nello sviluppo di questa alterazione ormonale. Probabilmente, sono necessarie altre condizioni concomitanti, come una predisposizione genetica o altre disfunzioni endocrine, affinché l'iperinsulinemia porti effettivamente a un aumento significativo del rapporto LH/FSH.

# 6 Conclusioni

I risultati ottenuti con l'analisi delle reti neurali e delle reti bayesiane hanno permesso di ottenere una comprensione più approfondita della PCOS.

Dal punto di vista della classificazione, la rete neurale con un singolo strato nascosto ha mostrato ottime performance, raggiungendo una F1-score del 91.36% e un'accuratezza del 91.25%, dimostrando la capacità delle reti neurali di individuare pattern complessi nei dati. Sebbene la regressione multinomiale abbia ottenuto valori leggermente superiori (F1-score 92.37%, accuratezza 92.50%), la differenza tra i due modelli è minima, il che suggerisce che entrambi rappresentano soluzioni efficaci per questo dataset. Al contrario, le reti neurali con più strati nascosti hanno mostrato una riduzione delle performance, confermando che una maggiore complessità architetturale non sempre si traduce in un miglioramento dei risultati.

Dall'altro lato, la rete bayesiana sviluppata ha permesso di esplorare le relazioni di dipendenza tra le variabili, evidenziando il ruolo centrale del BMI e dell'età nel determinare il rischio di sviluppare la PCOS, con una maggiore incidenza della forma anovulatoria tra le donne giovani e in sovrappeso. Inoltre, l'analisi bayesiana ha confermato un'associazione tra iperglicemia e squilibri ormonali, rafforzando l'ipotesi di un legame tra insulino-resistenza e alterazioni endocrinologiche nella PCOS.

Complessivamente, l'utilizzo di questi due approcci ha consentito non solo di sviluppare un modello predittivo accurato, ma anche di offrire una chiave di lettura più approfondita per ottimizzare le strategie diagnostiche e terapeutiche nella gestione della PCOS.

# 7 Bibliografia

1. Bozdag G, Mumusoglu S, Zengin D, Karabulut E, Yildiz BO. The prevalence and phenotypic features of polycystic ovary syndrome: A systematic review and meta-analysis. Human reproduction. 2016;31:2841–55.

2. Brassard M, AinMelk Y, Baillargeon J-P. Basic infertility including polycystic ovary syndrome. Medical Clinics of North America. 2008;92:1163–92.

3. Alberti KGMM, Zimmet P, Shaw J. International diabetes federation: A consensus on type 2 diabetes prevention. Diabetic Medicine. 2007;24:451–63.

4. Deeks AA, Gibson-Helm ME, Teede HJ. Anxiety and depression in polycystic ovary syndrome: A comprehensive investigation. Fertility and sterility. 2010;93:2421–3.

5. Huber-Buchholz M-M, Carey D, Norman R. Restoration of reproductive potential by lifestyle modification in obese polycystic ovary syndrome: Role of insulin sensitivity and luteinizing hormone. The Journal of Clinical Endocrinology & Metabolism. 1999;84:1470–4.

6. Kottarathil P. Polycystic ovary syndrome (PCOS). 2020.

7. WHO Consultation on Obesity (1999: Geneva S, Organization WH. Obesity : Preventing and managing the global epidemic : Report of a WHO consultation. 2000;252 p.