POLITECNICO DI TORINO

Corso di Laurea Magistrale in Ingegneria Biomedica

Sviluppo e Implementazione di un algoritmo di Machine Learning per il riconoscimento di pazienti affetti da Malattie Sessualmente Trasmissibili

Relatrice

Prof.ssa Gabriella OLMO

Correlatrice

Ing. Federica SEMPRONI

Candidata

Virginia PASTORINI

Ottobre 2021

Abstract

Le Malattie Sessualmente Trasmissibili (MST) sono uno dei problemi più diffusi nel mondo e spesso la loro trasmissione è dovuta alla mancanza di informazione: infatti esse coinvolgono particolarmente adolescenti o giovani adulti, i quali frequentemente affrontano con imbarazzo questi temi.
Si intende quindi creare, con il progetto Bridge, un portale in grado di fornire affidabili informazioni relative a questa tematica, del quale uno degli elementi principali è un questionario, che intende concentrarsi sulle MST più diffuse in Italia, cioè Clamidia, Condiloma ano-genitale, Gonorrea, Herpes genitale, HIV e Sifilide. Lo scopo del questionario è informativo, non diagnostico: si intende infatti fornire all'utente il grado di rischio di aver contratto una MST e una serie di informazioni utili relative al singolo caso, in modo da incoraggiare innanzitutto una visita medica e successivamente alcune soluzioni e/o informazioni sulla base delle risposte fornite.
Il presente lavoro di tesi si pone come obiettivo principale l'ottimizzazione e il miglioramento dell'efficacia del questionario e il riconoscimento dei pazienti che hanno contratto una MST. A tal fine sono stati raccolti i dati, ovvero le risposte al questionario, dai pazienti che si sono recati nei consultori di 13 regioni d'Italia e che hanno deciso di contribuire al progetto. A causa dell'attuale pandemia da Covid-19 e la conseguente riduzione di affluenza di pazienti nei consultori, si è deciso di lavorare temporaneamente su un dataset simulato, ovvero con risposte inviate dai medici e/o specializzandi, ai quali viene chiesto di immedesimarsi in un paziente con una specifica patologia o un paziente sano.
Il fine ultimo dell'analisi dei dati svolta consiste nell' implementazione di un algoritmo, che utilizza la clusterizzazione mediante reti neurali SOM (Self-Organizing Map) per distinguere inizialmente i pazienti sani dai pazienti affetti dalle patologie indicate e, contemporaneamente, un Algoritmo Genetico per la selezione delle domande del questionario necessarie e sufficienti per il raggiungimento di soddisfacenti prestazioni della rete neurale precedente. Le domande selezionate in quest'ultima fase sono state confrontate con i risultati di una ricerca bibliografica volta a identificare il valore della domanda per la distinzione di ogni singola patologia. Un'ulteriore indagine è stata svolta per l'analisi delle singole patologie nei clusters ottenuti, in modo tale da verificare un possibile riconoscimento successivo della singola patologia del soggetto.
Il lavoro di tesi presentato rappresenta un'analisi in fase preliminare, in quanto il numero di dati del dataset simulato conta 314 pazienti, di cui 90 pazienti sani. Inoltre, in una fase successiva, l'algoritmo dovrà essere validato mediante l'utilizzo

del dataset costituito dai dati reali dei pazienti, che risultano ad oggi essere in numero troppo esiguo.
In conclusione, la procedura di selezione delle domande ha prodotto risultati soddisfacenti per questa fase preliminare: si registra infatti un miglioramento nella clusterizzazione a seguito della riduzione delle domande utilizzate dall'algoritmo, comportando un miglioramento nella distinzione dei pazienti sani dai pazienti malati.

Ringraziamenti

Desidero innanzitutto ringraziare la mia relatrice, la Prof.ssa Gabriella Olmo, per la sua disponibilità e professionalità dimostratami in questi mesi.

Un sentito ringraziamento a tutti i membri del team Bridge, per la possibilità di partecipare a questo progetto e per il supporto durante questi mesi di lavoro, in particolare all'Ing. Federica Semproni, mia correlatrice, per la sua immensa pazienza e per i suoi indispensabili consigli durante i mesi di lavoro e la stesura di questo elaborato.

Ringrazio infinitamente la mia famiglia, in particolare i miei genitori, che mi hanno sempre sostenuto, appoggiando ogni mia decisione, fin dalla scelta del mio percorso di studi.
Ringrazio anche i miei amici, per il supporto durante il mio percorso, soprattutto nei momenti più difficili.

Alla mia famiglia

Indice

Acronimi **xvii**

1 Introduzione **1**
1.1 Il progetto Bridge . . . 1
1.2 Descrizione del team Bridge . . . 2
1.3 Scopo del progetto . . . 3

2 Malattie Sessualmente Trasmissibili **4**
2.1 HIV . . . 5
2.1.1 Questionario . . . 5
2.2 Condiloma ano-genitale . . . 6
2.2.1 Questionario . . . 7
2.3 Sifilide . . . 8
2.3.1 Sintomi . . . 8
2.3.2 Questionario . . . 9
2.4 Herpes genitale . . . 11
2.4.1 Questionario . . . 11
2.5 Clamidia . . . 12
2.5.1 Sintomi . . . 13
2.5.2 Complicanze . . . 13
2.5.3 Questionario . . . 13
2.6 Gonorrea . . . 15
2.6.1 Sintomi . . . 15
2.6.2 Complicanze . . . 16
2.6.3 Questionario . . . 16

3 Classificatore **18**
3.1 Raccolta ed elaborazione dei dati . . . 18
3.1.1 Raccolta dati . . . 19
3.1.2 Questionario . . . 19
3.1.3 Trattamento dei dati . . . 22
3.1.4 Elaborazione e codifica . . . 23

3.2 Algoritmo di clusterizzazione 28
3.2.1 Metodo 31
3.2.2 Risultati 32
3.2.3 Distribuzione delle patologie nei clusters 51
3.2.4 Conclusioni 53
3.3 Feature Selection 54
3.3.1 Metodi 55
3.3.2 Analisi dei risultati dell'Algoritmo Genetico 57
3.3.3 Risultati delle prove con due classi 59
3.3.4 Risultati delle prove con patologie distinte 63
3.3.5 Scelta del metodo e analisi delle soluzioni 69
3.4 Clusterizzazione a seguito della Feature Selection 71
3.4.1 Risultati 71
3.4.2 Risultati delle ripetizioni 82

4 Discussione dei risultati **94**

5 Conclusioni e sviluppi futuri **101**

Bibliografia **103**

Elenco delle figure

3.1 Domanda “Indica il tuo sesso”. 19
3.2 Domanda “Indica la tua etnia”. 20
3.3 Domanda “Quando hai svolto il Pap test l’ultima volta?”. 20
3.4 Domanda “Nel passato, ti è mai stato detto da un operatore sanitario di avere le seguenti MST?”. 21
3.5 Domanda “Negli ultimi 6 mesi hai avuto rapporti...”. 21
3.6 Domanda relativa alla masturbazione e all’uso di sex toys. 22
3.7 Domanda relativa all’eritema a macchie rosa tipico del secondo stadio della Siflide. 22
3.8 Esempio di valori per i vari tagli e legenda 24
3.9 Confusion Matrix corrispondente al taglio. 24
3.10 Esempio dei valori ottenuti con la funzione “categorical”. 26
3.11 Parte della matrice “Answer”. 27
3.12 Variabilità intra-cluster del dataset con 116 pazienti. 27
3.13 Variabilità intra-cluster del dataset con 139 pazienti. 27
3.14 Distanza inter-cluster del dataset con 116 pazienti. 28
3.15 Distanza inter-cluster del dataset con 139 pazienti. 28
3.16 Confronto clusterizzazione per il dataset con 139 pazienti, che riporta la differenza tra percentuale di pazienti sani calcolata rispetto al totale (25) e la percentuale di pazienti presenti all’interno del cluster calcolata rispetto al totale (139). 29
3.17 Confronto clusterizzazione con percentuale di pazienti sani rispetto al totale nel cluster (139 pazienti). 29
3.18 Grafico che mostra la distribuzione delle classi dei 314 pazienti simulati. 30
3.19 Confusion matrix relativa alla sezione “intero questionario”, rete 4x4, taglio a 2. 33
3.20 Mappa della rete SOM relativa alla sezione “intero questionario”, rete 4x4, taglio a 2, con evidenziati i neuroni relativi ai clusters. . . 33
3.21 Confusion matrix relativa alla sezione “intero questionario”, rete 5x5, taglio a 2. 34

3.22 Mappa della rete SOM relativa alla sezione “intero questionario”, rete 5x5, taglio a 2, con evidenziati i neuroni relativi ai clusters. . . 34
3.23 Confusion matrix relativa alla sezione “intero questionario”, rete 6x6, taglio a 2. 35
3.24 Mappa della rete SOM relativa alla sezione “intero questionario”, rete 6x6, taglio a 2, con evidenziati i neuroni relativi ai clusters. . . 35
3.25 Confusion matrix relativa alla sezione “intero questionario”, rete 6x6, taglio a 3. 35
3.26 Mappa della rete SOM relativa alla sezione “intero questionario”, rete 6x6, taglio a 3, con evidenziati i neuroni relativi ai clusters. . . 35
3.27 Confusion matrix relativa alla sezione “intero questionario”, rete 7x7, taglio a 2. 36
3.28 Mappa della rete SOM relativa alla sezione “intero questionario”, rete 7x7, taglio a 2, con evidenziati i neuroni relativi ai clusters. . . 36
3.29 Confusion matrix relativa alla sezione “intero questionario”, rete 8x8, taglio a 2. 37
3.30 Mappa della rete SOM relativa alla sezione “intero questionario”, rete 8x8, taglio a 2, con evidenziati i neuroni relativi ai clusters. . . 37
3.31 Confusion matrix relativa alla sezione “vita sessuale”, rete 4x4, taglio a 2. 38
3.32 Mappa della rete SOM relativa alla sezione “vita sessuale”, rete 4x4, taglio a 2, con evidenziati i neuroni relativi ai clusters. 38
3.33 Confusion matrix relativa alla sezione “vita sessuale”, rete 4x4, taglio a 3. 38
3.34 Mappa della rete SOM relativa alla sezione “vita sessuale”, rete 4x4, taglio a 3, con evidenziati i neuroni relativi ai clusters. 38
3.35 Confusion matrix relativa alla sezione “vita sessuale”, rete 5x5, taglio a 2. 40
3.36 Mappa della rete SOM relativa alla sezione “vita sessuale”, rete 5x5, taglio a 2, con evidenziati i neuroni relativi ai clusters. 40
3.37 Confusion matrix relativa alla sezione “vita sessuale”, rete 6x6, taglio a 2. 40
3.38 Mappa della rete SOM relativa alla sezione “vita sessuale”, rete 6x6, taglio a 2, con evidenziati i neuroni relativi ai clusters. 40
3.39 Confusion matrix relativa alla sezione “vita sessuale”, rete 6x6, taglio a 3. 41
3.40 Mappa della rete SOM relativa alla sezione “vita sessuale”, rete 6x6, taglio a 3, con evidenziati i neuroni relativi ai clusters. 41
3.41 Confusion matrix relativa alla sezione “vita sessuale”, rete 7x7, taglio a 2. 42

3.42 Mappa della rete SOM relativa alla sezione “vita sessuale”, rete 7x7, taglio a 2, con evidenziati i neuroni relativi ai clusters. 42
3.43 Confusion matrix relativa alla sezione “vita sessuale”, rete 8x8, taglio a 2. 42
3.44 Mappa della rete SOM relativa alla sezione “vita sessuale”, rete 8x8, taglio a 2, con evidenziati i neuroni relativi ai clusters. 42
3.45 Confusion matrix relativa alla sezione “vita sessuale e sintomi”, rete 4x4, taglio a 2. 43
3.46 Mappa della rete SOM relativa alla sezione “vita sessuale e sintomi”, rete 4x4, taglio a 2, con evidenziati i neuroni relativi ai clusters. . . 43
3.47 Confusion matrix relativa alla sezione “vita sessuale e sintomi”, rete 5x5, taglio a 2. 44
3.48 Mappa della rete SOM relativa alla sezione “vita sessuale e sintomi”, rete 5x5, taglio a 2, con evidenziati i neuroni relativi ai clusters. . . 44
3.49 Confusion matrix relativa alla sezione “vita sessuale e sintomi”, rete 6x6, taglio a 2. 45
3.50 Mappa della rete SOM relativa alla sezione “vita sessuale e sintomi”, rete 6x6, taglio a 2, con evidenziati i neuroni relativi ai clusters. . . 45
3.51 Confusion matrix relativa alla sezione “vita sessuale e sintomi”, rete 6x6, taglio a 4. 46
3.52 Mappa della rete SOM relativa alla sezione “vita sessuale e sintomi”, rete 6x6, taglio a 4, con evidenziati i neuroni relativi ai clusters. . . 46
3.53 Confusion matrix relativa alla sezione “vita sessuale e sintomi”, rete 7x7, taglio a 2. 47
3.54 Mappa della rete SOM relativa alla sezione “vita sessuale e sintomi”, rete 7x7, taglio a 2, con evidenziati i neuroni relativi ai clusters. . . 47
3.55 Confusion matrix relativa alla sezione “vita sessuale e sintomi”, rete 7x7, taglio a 3. 47
3.56 Mappa della rete SOM relativa alla sezione “vita sessuale e sintomi”, rete 7x7, taglio a 3, con evidenziati i neuroni relativi ai clusters. . . 47
3.57 Confusion matrix relativa alla sezione “vita sessuale e sintomi”, rete 7x7, taglio a 6. 48
3.58 Mappa della rete SOM relativa alla sezione “vita sessuale e sintomi”, rete 7x7, taglio a 6, con evidenziati i neuroni relativi ai clusters. . . 48
3.59 Confusion matrix relativa alla sezione “vita sessuale e sintomi”, rete 8x8, taglio a 2. 49
3.60 Mappa della rete SOM relativa alla sezione “vita sessuale e sintomi”, rete 8x8, taglio a 2, con evidenziati i neuroni relativi ai clusters. . . 49
3.61 Confusion matrix relativa alla sezione “vita sessuale e sintomi”, rete 8x8, taglio a 4. 50

3.62 Mappa della rete SOM relativa alla sezione “vita sessuale e sintomi”, rete 8x8, taglio a 4, con evidenziati i neuroni relativi ai clusters. . . 50
3.63 Esempio di tabella per le patologie (rete 4x4 della sezione “Intero questionario”) 52
3.64 Grafico relativo alla qualità delle distribuzioni delle patologie relativamente alle tre diverse sezioni analizzate. 53
3.65 Esempio di rappresentazione dei criteri di confronto per le soluzioni dell’Algoritmo Genetico relativo al metodo con distanza euclidea e patologie unite, con pc=0,89 (pc=probabilità di crossover). 60
3.66 Riassunto delle probabilità di crossover (pc) che presentano soluzioni che soddisfano almeno due criteri (evidenziati in tabella) e le probabilità di crossover che presentano tutte le soluzioni uguali. . . 61
3.67 Grafico che mostra la composizione della soluzione con pc=0,95 per il metodo con due classi e utilizzo della distanza euclidea. 62
3.68 Esempio di rappresentazione dei criteri di confronto per le soluzioni dell’Algoritmo Genetico relativo al metodo con distanza euclidea e patologie distinte, con pc=0,88 (pc=probabilità di crossover). . . . 64
3.69 Riassunto delle probabilità di crossover (pc) che presentano soluzioni che soddisfano almeno due criteri per il metodo 3 (criteri soddisfatti evidenziati in tabella). 64
3.70 Grafico che mostra la composizione della soluzione con pc=0,82 con patologie separate. 65
3.71 Grafico che mostra la composizione della soluzione con pc=0,85 con patologie separate. 66
3.72 Riassunto delle probabilità di crossover (pc) che presentano soluzioni che soddisfano almeno due criteri per il Metodo 4. 67
3.73 Grafico che mostra la composizione della soluzione con pc=0,84 con patologie separate e distanza euclidea. 68
3.74 Grafico che mostra la composizione della soluzione con pc=0,93 con patologie separate e distanza euclidea. 68
3.75 Tabelle che illustrano le *features* selezionate dalle due soluzioni dell’Algoritmo Genetico. 70
3.76 Confusion matrix relativa alla soluzione 1, rete 4x4, taglio a 2. . . 72
3.77 Mappa della rete SOM relativa alla soluzione 1, rete 4x4, taglio a 2, con evidenziati i neuroni relativi ai clusters. 72
3.78 Confusion matrix relativa alla soluzione 1, rete 5x5, taglio a 2. . . 73
3.79 Mappa della rete SOM relativa alla soluzione 1, rete 5x5, taglio a 2, con evidenziati i neuroni relativi ai clusters. 73
3.80 Confusion matrix relativa alla soluzione 1, rete 6x6, taglio a 2. . . 74
3.81 Mappa della rete SOM relativa alla soluzione 1, rete 6x6, taglio a 2, con evidenziati i neuroni relativi ai clusters. 74

3.82 Confusion matrix relativa alla soluzione 1, rete 7x7, taglio a 2. . . 74
3.83 Mappa della rete SOM relativa alla soluzione 1, rete 7x7, taglio a 2, con evidenziati i neuroni relativi ai clusters. 74
3.84 Confusion matrix relativa alla soluzione 1, rete 8x8, taglio a 2. . . 75
3.85 Mappa della rete SOM relativa alla soluzione 1, rete 8x8, taglio a 2, con evidenziati i neuroni relativi ai clusters. 75
3.86 Confusion matrix relativa alla soluzione 2, rete 4x4, taglio a 2. . . 76
3.87 Mappa della rete SOM relativa alla soluzione 2, rete 4x4, taglio a 2, con evidenziati i neuroni relativi ai clusters. 76
3.88 Confusion matrix relativa alla soluzione 2, rete 4x4, taglio a 3. . . 77
3.89 Mappa della rete SOM relativa alla soluzione 2, rete 4x4, taglio a 3, con evidenziati i neuroni relativi ai clusters. 77
3.90 Confusion matrix relativa alla soluzione 2, rete 5x5, taglio a 2. . . 78
3.91 Mappa della rete SOM relativa alla soluzione 2, rete 5x5, taglio a 2, con evidenziati i neuroni relativi ai clusters. 78
3.92 Confusion matrix relativa alla soluzione 2, rete 6x6, taglio a 2. . . 78
3.93 Mappa della rete SOM relativa alla soluzione 2, rete 6x6, taglio a 2, con evidenziati i neuroni relativi ai clusters. 78
3.94 Confusion matrix relativa alla soluzione 2, rete 7x7, taglio a 2. . . 79
3.95 Mappa della rete SOM relativa alla soluzione 2, rete 7x7, taglio a 2, con evidenziati i neuroni relativi ai clusters. 79
3.96 Confusion matrix relativa alla soluzione 2, rete 8x8, taglio a 2. . . 80
3.97 Mappa della rete SOM relativa alla soluzione 2, rete 8x8, taglio a 2, con evidenziati i neuroni relativi ai clusters. 80
3.98 Confusion matrix relativa alla soluzione 2, rete 8x8, taglio a 3. . . 81
3.99 Mappa della rete SOM relativa alla soluzione 2, rete 8x8, taglio a 3, con evidenziati i neuroni relativi ai clusters. 81
3.100 Valori della variabilità intra-cluster relativa alla sezione “intero questionario”, per ogni dimensione delle reti neurali SOM. 82
3.101 Valori della distanza inter-cluster relativa alla sezione “intero questionario”, per ogni dimensione delle reti neurali SOM. 83
3.102 Boxplot della variabilità intra-cluster nelle prime prove (5 valori) e dopo le ripetizioni (25 valori) per la sezione “intero questionario”. . 83
3.103 Boxplot della distanza inter-cluster nelle prime prove (5 valori) e dopo le ripetizioni (25 valori) per la sezione “intero questionario”. . 83
3.104 Tabella con media e deviazione standard per entrambi i parametri per la sezione “intero questionario”. 84
3.105 Valori della variabilità intra-cluster relativa alla sezione “vita sessuale”, per ogni dimensione delle reti neurali SOM. 84
3.106 Valori della distanza inter-cluster relativa alla sezione “vita sessuale”, per ogni dimensione delle reti neurali SOM. 85

3.107 Boxplot della variabilità intra-cluster nelle prime prove (5 valori) e dopo le ripetizioni (25 valori) per la sezione “vita sessuale”. 85
3.108 Boxplot della distanza inter-cluster nelle prime prove (5 valori) e dopo le ripetizioni (25 valori) per la sezione “vita sessuale”. 85
3.109 Tabella con media e deviazione standard per entrambi i parametri per la sezione “vita sessuale”. 86
3.110 Valori della variabilità intra-cluster relativa alla sezione “vita sessuale e sintomi”, per ogni dimensione delle reti neurali SOM. . . . 86
3.111 Valori della distanza inter-cluster relativa alla sezione “vita sessuale e sintomi”, per ogni dimensione delle reti neurali SOM. 87
3.112 Boxplot della variabilità intra-cluster nelle prime prove (5 valori) e dopo le ripetizioni (25 valori) per la sezione “vita sessuale e sintomi”. 87
3.113 Boxplot della distanza inter-cluster nelle prime prove (5 valori) e dopo le ripetizioni (25 valori) per la sezione “vita sessuale e sintomi”. 87
3.114 Tabella con media e deviazione standard per entrambi i parametri per la sezione “vita sessuale e sintomi”. 88
3.115 Valori della variabilità intra-cluster relativa alla soluzione dell’AG, per ogni dimensione delle reti neurali SOM. 88
3.116 Valori della distanza inter-cluster relativa alla soluzione dell’AG, per ogni dimensione delle reti neurali SOM. 89
3.117 Boxplot della variabilità intra-cluster nelle prime prove (5 valori) e dopo le ripetizioni (25 valori) per la soluzione dell’AG. 89
3.118 Boxplot della distanza inter-cluster nelle prime prove (5 valori) e dopo le ripetizioni (25 valori) per la soluzione dell’AG. 89
3.119 Tabella con media e deviazione standard per entrambi i parametri per la soluzione dell’AG. 90
3.120 Tabella relativa al primo indicatore della clusterizzazione. 90
3.121 Boxplot relativo al primo indicatore della clusterizzazione. 91
3.122 Tabella relativa al secondo indicatore della clusterizzazione. 91
3.123 Boxplot relativo al secondo indicatore della clusterizzazione. 92
3.124 Tabella che indica quale sezione è maggiormente utile alla distinzione di ogni patologia per ogni serie di prove (“intero” indica la sezione “intero questionario”, “2” indica la sezione “vita sessuale”, “2+3” indica la sezione “vita sessuale e sintomi”, “GA” indica la soluzione dell’Algoritmo Genetico). 92
3.125 Grafico che indica la distribuzione delle patologie relative alle 5 ripetizioni. 93

4.1 Tabella con media e deviazione standard per entrambi i parametri per: (a) la sezione “intero questionario”, (b) la sezione “vita sessuale”, (c) la sezione “vita sessuale e sintomi”, (d) la soluzione dell’Algoritmo Genetico. 95
4.2 Grafico che riporta le strutture con variabilità intra-cluster diminuita, di cui si riporta anche la distribuzione nelle varie sezioni. . . 96
4.3 Grafico che riporta le strutture con distanza inter-cluster diminuita, di cui si riporta anche la distribuzione nelle varie sezioni. 96
4.4 Tabella che riporta la dimensione delle reti selezionate per i due indicatori della clusterizzazione, qualora la rete fosse la stessa è stata riportata una sola volta. 97
4.5 Grafico a torta che mostra la distribuzione dei tagli del dendrogramma nelle 4 prove per ognuna delle 5 ripetizioni. 98
4.6 Grafici a torta che mostrano i collegamenti individuati tra le singole patologie. 99
4.7 Risultato del confronto tra la soluzione ottenuta con la Feature Selection e la ricerca bibliografica. 99

Acronimi

AG
Algoritmo Genetico

HIV
Human Immunodeficiency Virus

HPV
Human Papilloma Virus

IST
Infezioni Sessualmente Trasmissibili

MIP
Malattia Infiammatoria Pelvica

MST
Malattie Sessualmente Trasmissibili

pc
probabilità di crossover

SOM
Self Organizing Map

Capitolo 1

Introduzione

1.1 Il progetto Bridge

Il progetto Bridge si propone di aiutare le persone nella gestione di potenziali Malattie e Infezioni Sessualmente Trasmissibili (MST) (IST) mediante la compilazione di un questionario che ha l'obiettivo di definire il grado di rischio di aver contratto una MST e, successivamente, fornire diverse strategie da adottare per iniziare ad affrontare l'eventuale condizione patologica.
Con l'utilizzo della telemedicina, inizialmente, lo scopo vedeva la progettazione di una mobile-app per fornire uno strumento che potesse rappresentare un supporto concreto per coloro che sospettino di aver contratto una MST, garantendo quelli che sono i valori principali di Bridge: privacy e anonimato. Successivamente, a seguito di sondaggi somministrati alla popolazione più giovane, l'idea e evoluta in una web-app, simile alla precedente versione per smartphone in quanto a funzionalità, ma con il vantaggio di essere piu facilmente accessibile tramite internet, evitando il download e l'eventuale accesso tramite account.
Tre sono stati i motivi principali che hanno spinto i membri del team a portare avanti l'idea nel corso dei mesi e a intraprendere una collaborazione con gli studenti laureandi del Politecnico di Torino. Innanzitutto, la realizzazione di uno strumento divulgativo di affidabili informazioni, utile ed efficace per la società odierna, che possa quindi ricoprire i gap presenti in questo campo della medicina. In secondo luogo, non meno importante per la buona riuscita del progetto, l'affiatamento del gruppo, l'abilita mostrata nella suddivisione dei compiti e l'entusiasmo nel condividere le conoscenze e le esperienze individuali in diversi ambiti. Infine, la conferma e il forte interesse da parte di medici e possibili utenti circa l'utilità effettiva dello strumento.

1.2 Descrizione del team Bridge

Il team Bridge si e formato durante un bootcamp di prototipizzazione chiamato *SEI INVENTOR* (Maggio 2020), con l'aggiunta di un quinto membro a seguito della stesura della propria tesi di Laurea Magistrale presso il Politecnico di Torino ("Progettazione e Sviluppo di una Mobile App per affrontare le Malattie Sessualmente Trasmissibili", Marzo 2021). Il bootcamp si e svolto presso la School of Entrepreneurship Innovation di Torino, scuola che ha l'intento di promuovere lo spirito imprenditoriale degli studenti universitari di varie discipline tramite momenti di formazione in aula, esperienze operative e testimonianze di imprenditori, investitori e vari professionisti.
Il team Bridge e costituito quindi da cinque membri, con forte propensione per la ricerca e l'innovazione, e con differenti background, che lo rendono solido ed eterogeneo.

Federica Semproni, 27 anni. Laurea Magistrale in Ingegneria Biomedica (Politecnico di Torino) con specializzazione in Strumentazione Medica. Attualmente dottoranda in BioRobotica, ricerca nell'ambito dei Soft Robots e Organi Artificiali, presso la Scuola Superiore Sant'Anna di Pisa. Nove mesi di esperienza nella ricerca a Boston (USA) presso il laboratorio di Motion Analysis dello Spaulding Rehabilitation Hospital, ospedale universitario per il programma di medicina fisica e riabilitazione della Harvard Medical School.

Andrea Fumero, 26 anni. Laurea Magistrale in Pianificazione territoriale, Urbanistica e Paesaggistico - Ambientale (Politecnico di Torino) e Laurea Magistrale in Pianificazione e progettazione urbana sostenibile (KTH Royal Institute of Technology in Stockholm). Un anno di esperienza lavorativa presso Ramboll Studio Dreiseitl (studio di architettura del paesaggio in Germania). Attualmente lavora come libero professionista presso lo studio LAND (studio di architettura del paesaggio di Milano).

Vincenzo Spinello, 27 anni. Laurea Magistrale in Business and Administration presso SSA-Scuola di management di Torino. Attualmente iscritto alla Scuola di Programmazione, 42 Roma Luiss.

Giuseppe Missale, 23 anni. Laurea Triennale in Ingegneria Biomedica (Politecnico di Torino), formazione precedente presso l'ITIS Archimede di Catania. Attualmente iscritto al corso di Laurea Magistrale in Ingegneria Biomedica (Politecnico di Torino), indirizzo Strumentazione Medica.

Maria Giovanna Longo, 25 anni. Laurea Magistrale in Ingegneria Biomedica presso il Politecnico di Torino, specializzazione in eHealth. Competenze nella progettazione di software medicali, sistemi per la telemedicina, sistemi di supporto alle decisioni cliniche. Attualmente lavora come Application Development Analyst presso Accenture a Milano.

1.3 Scopo del progetto

Spesso i pazienti che contraggono una MST sono adolescenti o giovani adulti, a causa di comportamenti scorretti dovuti a una scarsa consapevolezza e informazione e a causa dell'imbarazzo che questo argomento può suscitare nell'affrontare la tematica.
La web-app Bridge nasce con lo scopo di avvicinare e sensibilizzare l'utente alle MST e alla loro prevenzione fornendo la possibilità di accedere anonimamente ad affidabili informazioni, superando eventuali barriere dovute alla condizione sociale o familiare, o semplicemente alla giovane età.
Uno dei servizi che il team Bridge intende offrire è un questionario in grado di fornire indicazioni relative al grado di rischio di aver contratto una MST. Nonostante l'analisi venga svolta con lo scopo di distinguere i pazienti sani da pazienti affetti da un gruppo di patologie specifiche (Clamidia, Condiloma ano-genitale, Gonorrea, Herpes e Sifilide) non verrà mai effettuata una diagnosi, in quanto lo scopo della web-app è l'utilizzo della telemedicina come strumento di integrazione della sanità tradizionale, al fine di ampliarne il raggio d'azione senza sovrapporsi o sostituirsi ad essa.
Lo scopo del progetto di tesi è quello di creare un classificatore che, mediante i dati attualmente disponibili, permetta di distinguere i pazienti sani dai pazienti malati e migliorare il più possibile la distinzione delle singole patologie, al fine di aumentare l'accuratezza delle informazioni fornite all'utente relative al rischio di aver contratto un'infezione o malattia.
Parte fondamentale di questo progetto è quindi l'ottimizzazione del questionario, costituito inizialmente da 37 domande, innanzitutto modificandone la forma, in modo da migliorarne la completezza e la comprensibilità, successivamente selezionando ed eliminando le domande che non portano informazione. Queste modifiche miglioreranno anche l'esperienza dell'utente ed eviteranno che il suo interesse diminuisca durante la compilazione.

Capitolo 2

Malattie Sessualmente Trasmissibili

Conseguentemente al lavoro di tesi precedente [1], seguito dal gruppo Bridge, è stata svolta una ricerca bibliografica volta a raccogliere ulteriori informazioni sulle MST, in particolare modo relative alle domande presenti nel questionario somministrato ai pazienti. Il questionario è stato, inoltre, revisionato da una psicologa, per rendere le domande più sensibili al target di persone al quale ci si rivolge, e da un medico specializzato (Dott. Dal Conte), per un riscontro sulle informazioni richieste.
Le MST su cui il team Bridge si è maggiormente concentrato sono:

- HIV
- Condiloma ano-genitale
- Sifilide
- Herpes genitale
- Clamidia
- Gonorrea

in quanto le prime cinque sono le MST più diffuse in Italia [2], mentre l'ultima, l'HIV (Human Immunodeficiency Virus), presenta una forte correlazione con le altre patologie, infatti un soggetto affetto da MST ha una maggiore probabilità di contrarre o trasmettere l'HIV.
Lo scopo finale di tale ricerca è ricavare informazioni aggiuntive sull'importanza a livello teorico delle singole domande proposte, sia per la distinzione dei pazienti sani dai pazienti affetti da MST, sia per la distinzione delle singole patologie. Inoltre, può essere un utile strumento di confronto per la fase di ottimizzazione e selezione

delle informazioni e domande sufficienti e necessarie per permettere all'algoritmo di definire il grado di rischio e fornire le indicazioni necessarie all'utente.

2.1 HIV

L'infezione da virus dell'immunodeficienza umana (HIV) è causata da due retrovirus simili (HIV-1 e HIV-2) che distruggono i linfociti T CD4+ e inibiscono la risposta immunitaria cellulo-mediata, aumentando il rischio di alcune infezioni e di alcuni tumori.
L'infezione primaria può produrre una sindrome febbrile aspecifica, mentre le successive manifestazioni (correlate allo stato di immunodeficienza), è proporzionale al livello di deplezione del linfocita CD4+ circolante. L'HIV può causare un deterioramento cognitivo, ipogonadismo, insufficienza renale e cardiomiopatia.
Il paziente può manifestare da uno stato di portatore asintomatico fino alla sindrome da immunodeficienza acquisita, definita da una conta di CD4 circolanti < 200/mcL, oppure dall'insorgenza di gravi infezioni opportunistiche e/o tumori.
La terapia è in grado di ripristinare la funzione immunitaria nella maggior parte dei pazienti se la soppressione della replicazione è sostenuta [3].

2.1.1 Questionario

Si riportano le informazioni presenti nelle domande del questionario, rivolte ai soggetti che si presentano in consultorio, in grado di distinguere i pazienti affetti da HIV dai pazienti sani o da pazienti affetti dalle altre MST. Vengono riportate anche le informazioni presenti nel lavoro di tesi precedentemente seguito dal team Bridge [1].

- *Data dell'ultimo rapporto sessuale*: prime settimane dall'infezione si possono presentare sintomi simil-influenzali come febbre, dolori articolari e muscolari, debolezza, sudorazioni notturne.
- *Precedente controllo generale sulle MST (sì/no)*: un no aumenta il rischio per tutte le MST.
- *Ultimo test MST*: la data dell'ultimo test è influente solo dal punto di vista dello stadio.
- *MST passate*: avere avuto una MST in passato aumenta la probabilità di contrarre l'HIV.
- *Il paziente ha subito una trasfusione di sangue prima del 1992 (sì/no)*: a rischio in quanto il sangue non veniva controllato per l'HIV.

- *Rapporti con più di un partner negli ultimi sei mesi*: aumento del rischio se non viene usato il preservativo.
- Masturbazione e utilizzo sex toys: la masturbazione reciproca non trasmette l'HIV. se i sex toys non vengono utilizzati nella maniera corretta, disinfettati ed entrano a contatto con più parti genitali durante lo stesso rapporto [4].
- *Uso del preservativo*: fondamentale in questo caso.
- *Paziente costretto ad avere rapporti sessuali*: le possibili lacerazioni possono causare il contagio mediante il sangue, inoltre, non si conoscono le condizioni di salute del partner sessuale [3].
- *Rapporti sessuali dietro compenso (sì/no)*: aumenta il rischio in quanto implica un maggior numero di partner sessuali diversi e con condizioni di salute non conosciute.
- *Uso di droghe non a scopo medico*: l'eventuale uso di aghi aumenta il rischio di contagio [3].
- *Partner che ha fatto uso di droghe non a scopo medico (sì/no)*: l'uso di aghi aumenta il rischio di contagio, e l'uso di droghe comporta rapporti sessuali non prevedibili [3].
- *Partner con MST*: rischio se senza preservativo.
- *Sintomi possibili*:
  - *Linfonodi ingrossati*: sintomo correlato [1];
  - *Febbre*: uno dei possibili sintomi influenzali delle prime settimane [1].

2.2 Condiloma ano-genitale

I condilomi ano-genitali sono causati da alcuni tipi di virus del papilloma umano (HPV del tipo 6, 11 o altri). Si presentano come piccole escrescenze benigne presenti prevalentemente a livello degli organi genitali esterni, intorno all'ano, nella bocca e nella gola (oro-faringe).
Generalmente non provocano sintomi ma tendono ad aumentare di numero e ad estendersi nelle aree circostanti e, in alcuni casi, possono recare disturbo a funzione sessuale, minzione e defecazione. Inoltre, tendono a recidivare e il HPV che li causa può anche provocare tumori maligni a carico del collo dell'utero (cervice uterina), del pene, dell'ano e dell'oro-faringe [5].

2.2.1 Questionario

Si riportano le informazioni presenti nelle domande del questionario, rivolte ai soggetti che si presentano in consultorio, in grado di distinguere i pazienti affetti da condilo ano-genitale dai pazienti sani o da pazienti affetti dalle altre MST. Vengono riportate anche le informazioni presenti nel lavoro di tesi precedentemente seguito dal team Bridge [1].

- *Sesso del paziente*: vi è un'incidenza maggiore nei pazienti di sesso biologico maschile.
- *Età del paziente*: gran parte dei casi diagnosticati interessa giovani e adulti (15-40 anni).
- *Data dell'ultimo rapporto sessuale*: i sintomi, se presenti, possono comparire a settimane o mesi dal contagio.
- *Paziente in gravidanza*: la presenza dei condilomi ano-genitali in gravidanza può comportare un peggioramento dei sintomi.
- *Uso di contraccettivi a rilascio di farmaco*: l'utilizzo di contraccettivi orali rappresenta un fattore di rischio per la comparsa dei condilomi [6].
- *Data dell'ultimo Pap-Test*: si considera un aumento del rischio se il test non viene svolto da almeno sei mesi.
- *Risultato del Pap-test*: da considerare in relazione alla data, presenta un rischio aumentato se è positivo o non è mai stato mai svolto.
- *Precedente controllo generale sulle MST (sì/no)*: un no aumenta il rischio per tutte le MST.
- *Vaccinazione Hpv a tre dosi*: una vaccinazione parziale o completa riduce il rischio.
- *Rapporti con più di un partner negli ultimi sei mesi*: aumenta il rischio, in particolare se non viene utilizzato il preservativo.
- *Contatto diretto con i genitali (sì/no)*: essendo la clamidia un'infezione del tratto urinario, il contatto è indispensabile, ma non è necessario che sia diretto, per cui un sì a questa domanda aumenta il rischio, ma un no non lo annulla.
- *Masturbazione e utilizzo sex toys*: entrambi aumentano il rischio, soprattutto se questi ultimi non vengono utilizzati nella maniera corretta, disinfettati ed entrano a contatto con più parti genitali durante lo stesso rapporto.

- *Uso del preservativo*: l'assenza del preservativo o un preservativo rotto aumenta il rischio.
- *Rapporti sessuali dietro compenso (sì/no)*: aumenta il rischio in quanto implica un maggior numero di partner sessuali diversi e con condizioni di salute non conosciute.
- *Sintomi possibili*: l'unico è l'infiammazione e gonfiore in zona genitale [7], infatti prurito e gonfiore in zona anale e vaginale possono presentarsi unitamente alle verruche genitali.
- *Presenza di eruzioni cutanee*: da considerarsi unitamente ai tempi compatibili.
- *Identificazione delle verruche genitali dalle relative immagini*: da considerarsi unitamente ai tempi compatibili.
- *Identificazione delle verruche genitali dal riassunto immagini*: da considerarsi unitamente ai tempi compatibili.

2.3 Sifilide

La sifilide è una malattia a trasmissione sessuale causata dal batterio *Treponema pallidum*, la cui sintomatologia si può manifestare in tre fasi, intervallate da periodi di assenza di sintomi [8].

2.3.1 Sintomi

Nel primo stadio si presenta un'ulcera indolore (detta sifiloma) nella sede di infezione (genitali), oppure in altra sede (ano, retto, labbra, lingua, gola, cervice, dita o altre parti del corpo). Il sifiloma si presenta come una piccola zona arrossata e rilevata, che evolve rapidamente in una piaga aperta di consistenza dura, rilevata e relativamente indolore. Esso di solito guarisce in 3-12 settimane.
Nel secondo stadio i batteri entrano in circolo causando eritema cutaneo diffuso (in genere indolore), gonfiore ai linfonodi e, meno comunemente, altra sintomatologia ad altri organi. Questo stadio può essere di breve durata o persistere per mesi; anche senza trattamento, alla fine si risolve, ma potrebbe ripresentarsi. Nelle aree cutanee umide, possono svilupparsi neoformazioni piane lisce, definite condilomi piani. Questi sono indolori ma molto infettivi, in quanto contengono molti batteri della sifilide e possono aprirsi e trasudare.
Segue uno stadio latente, in cui la sifilide può rimanere per sempre e, generalmente, in questo stadio non è contagiosa. Raramente, all'inizio dello stadio latente possono comparire ulcere infettive sulla pelle o sulle mucose.

La sifilide terziaria si sviluppa in circa un terzo delle persone non trattate, anni o decenni dopo l'infezione iniziale.
Vi sono tre forme cliniche principali di sifilide terziaria:

1. *Sifilide terziaria benigna*: rara, sulla cute compaiono neoformazioni molli di consistenza gommosa, chiamate gomme, a volte anche sugli organi interni. Queste possono ulcerare;

2. *Sifilide cardiovascolare*: i batteri infettano i vasi sanguigni collegati al cuore, inclusa l'aorta;

3. *Neurosifilide*: può essere asintomatica, ma anche colpire meningi, midollo spinale e altre parti del cervello [8].

2.3.2 Questionario

Si riportano le informazioni presenti nelle domande del questionario, rivolte ai soggetti che si presentano in consultorio, in grado di distinguere i pazienti affetti da sifilide dai pazienti sani o da pazienti affetti dalle altre MST. Vengono riportate anche le informazioni presenti nel lavoro di tesi precedentemente seguito dal team Bridge [1].

- *Sesso del paziente*: per le persone di sesso biologico maschile, il rapporto a rischio è in genere di tipo eterosessuale e risalente dai tre mesi ai due anni precedenti. Per i pazienti di sesso biologico femminile, i rapporti sono in genere di tipo omosessuale e risalente dai tre mesi ai due anni precedenti.

- *Età*: i pazienti affetti da sifilide sono in maggioranza tra i 20 e i 44 anni (la fascia di età maggiore è tra 25 e 34 anni) [9].

- *Data dell'ultimo rapporto sessuale*: l'insorgenza dei primi sintomi avviene dopo 10-90 giorni dal contagio, successivamente la durata dei vari stadi e del tempo trascorso tra l'insorgenza dei sintomi diventa molto variabile, se il paziente non viene trattato aumenta il rischio di stadi tardivi.

- *Precedente controllo generale sulle MST (sì/no)*: un no aumenta il rischio per tutte le MST.

- *Ultimo test MST*: la data dell'ultimo test è influente solo dal punto di vista dello stadio.

- *Tipo di rapporto*: particolarmente a rischio i partner sessuali di donne affette da sifilide primaria, secondaria o latente, il rischio quindi aumenta per rapporti eterosessuali o omosessuali femminili, a seconda del sesso biologico del paziente.

- *Almeno uno dei partner conduce una vita promiscua*: la ripetuta esposizione sessuale con partner differenti risulta essere un grande fattore di rischio per l'acquisizione dell'herpes.
- *Contatto diretto con i genitali (sì/no)*: anche il contatto con l'eritema tipico del secondo stadio della sifilide può causare un contagio, perciò può essere sufficiente [8].
- *Masturbazione e utilizzo sex toys*: entrambi aumentano il rischio, se si viene a contatto con un sifiloma o con un eritema e se questi ultimi non vengono utilizzati nella maniera corretta, disinfettati ed entrano a contatto con più parti genitali durante lo stesso rapporto [8].
- *Uso del preservativo*: riduce il rischio ma non lo annulla [8].
- *Paziente costretto ad avere rapporti sessuali*: le possibili lacerazioni possono causare il contagio mediante il sangue, inoltre, non si conoscono le condizioni di salute del partner sessuale [10].
- *Rapporti sessuali dietro compenso (sì/no)*: aumenta il rischio in quanto implica un maggior numero di partner sessuali diversi e con condizioni di salute non conosciute.
- *Partner con MST*: particolarmente a rischio i partner sessuali di donne affette da sifilide primaria, secondaria o latente.
- *Sintomi possibili*:
  - *Eruzione cutanea in zona genitale* [1];
  - *Linfonodi ingrossati*: presenza possibile nello stadio primario [8] e secondario [1];
  - *Sensibilità alla luce* [1];
  - *Colore e consistenza dell'urina*: tipico del primo stadio [8];
  - *Febbre*: uno dei possibili sintomi influenzali del secondo stadio della sifilide [1].
- *Presenza di eruzioni cutanee, rossori (sì/no)*: sifiloma nello stadio primario, eritema a macchie rosse nello stato secondario.
- *Identificazione del sifiloma*: da considerarsi unitamente ai tempi compatibili.
- *Identificazione dell'eritema*: da considerarsi unitamente ai tempi compatibili.
- *Identificazione dell'eritema o del sifiloma dal riassunto delle immagini*: da considerarsi unitamente ai tempi compatibili.

2.4 Herpes genitale

L'herpes genitale è la patologia ulcerativa a trasmissione sessuale Sono causati da alcuni tipi di virus del papilloma umano (HPV del tipo 6, 11 o altri) dei Paesi sviluppati, causata da herpes virus umani 1 (*virus herpes simplex-1*) o 2 (*virus herpes simplex-2*).
Dopo l'infezione iniziale, il virus rimane quiescente nei gangli nervosi a partire dai quali può periodicamente riattivarsi, causando lesioni genitali. La trasmissione può avvenire attraverso il contatto con le lesioni o, più spesso, mediante spargimento asintomatico, cioè attraverso il contatto pelle a pelle con partner sessuali quando le lesioni non sono evidenti. Le vescicole solitamente si erodono formando ulcere che possono confluire, possono lasciare cicatrici e recidivano nel 50% dei casi per il virus dell'*herpes simplex di tipo 1* e nell' 80% dei casi per il virus dell'*herpes simplex di tipo 2* [11].

2.4.1 Questionario

Si riportano le informazioni presenti nelle domande del questionario, rivolte ai soggetti che si presentano in consultorio, in grado di distinguere i pazienti affetti da herpes genitale dai pazienti sani o da pazienti affetti dalle altre MST. Vengono riportate anche le informazioni presenti nel lavoro di tesi precedentemente seguito dal team Bridge [1].

- *Data dell'ultimo rapporto sessuale*: dopo un periodo di incubazione pari a 4-7 giorni, i sintomi possono comparire dai 4 ai 10 gg dopo il rapporto.

- *Gravidanza*: c'è il rischio di un taglio cesareo se la madre è stata contagiata entro i 3 mesi di gravidanza [12].

- *Precedente controllo generale sulle MST (sì/no)*: un no aumenta il rischio per tutte le MST.

- *MST passate*: con una diagnosi precedente di herpes, è possibile che ci sia una recidiva, infatti la frequenza di ricomparsa dei sintomi di solito si riduce con il passare del tempo ma i sintomi potrebbero ripresentarsi sporadicamente per diversi anni.

- *Rapporti con più di un partner negli ultimi sei mesi*: la ripetuta esposizione sessuale con partner differenti risulta essere un grande fattore di rischio per l'herpes genitale, in quanto può esserci contagio anche in assenza dei sintomi [11].

- *Almeno uno dei partner conduce una vita promiscua*: la ripetuta esposizione sessuale con partner differenti risulta essere un grande fattore di rischio per l'acquisizione dell'herpes.
- *Contatto diretto con i genitali (sì/no)*: è sufficiente il semplice contatto con i genitali per la trasmissione dell'infezione.
- *Masturbazione e utilizzo sex toys*: entrambi aumentano il rischio, soprattutto se questi ultimi non vengono utilizzati nella maniera corretta, disinfettati ed entrano a contatto con più parti genitali durante lo stesso rapporto.
- *Uso del preservativo*: non è sufficiente, in quanto le vescicole possono essere presenti anche intorno ai genitali.
- *Ruolo nel rapporto*: può variare la zona coinvolta [11].
- *Rapporti sessuali dietro compenso (sì/no)*: aumenta il rischio in quanto implica un maggior numero di partner sessuali diversi.
- *Sintomi possibili*:
  - *Eruzione cutanea in zona genitale* [1];
  - *Infiammazione e gonfiore in zona genitale* [1];
  - *Abbondanti secrezioni/perdite intime* [1];
  - *Febbre*, per possibile immunosoppressione momentanea [1].
- *Presenza di eruzioni cutanee, rossori (sì/no)*: presente in quanto le vescicole sono il sintomo principale.
- *Identificazione delle vescicole da herpes genitale*: da considerarsi unitamente ai tempi compatibili.
- *Identificazione delle vescicole da herpes genitale*: da considerarsi unitamente ai tempi compatibili.

2.5 Clamidia

Le infezioni da clamidia sono causate dal batterio *Chlamydia trachomatis*, il quale colpisce l'uretra, la cervice uterina e il retto. L'infezione dell'uretra può essere causata anche da altri batteri come quelli dei generi *Ureaplasma* e *Mycoplasma* [13].

2.5.1 Sintomi

I sintomi nell'uomo iniziano 7-28 giorni dopo la contrazione dell'infezione viene contratta durante il rapporto sessuale e includono una leggera sensazione di bruciore nell'uretra e/o secrezione trasparente (meno densa rispetto alla gonorrea) o torbida dal pene, al mattino presto l'orifizio del pene è spesso arrossato e chiuso da secrezioni secche. In alcuni casi i sintomi si manifestano in maniera più grave.
I sintomi nella donna sono spesso assenti, ma se presenti possono includere frequente stimolo alla minzione, minzione dolorosa e secrezione di muco giallo e di pus dalla vagina, con possibile dolore durante i rapporti sessuali.
In entrambi i sessi, in caso di coinvolgimento del retto, si possono manifestare dolore e secrezioni. Le infezioni alla gola sono asintomatiche. I sintomi, se presenti, si riducono in circa 4 settimane [13].

2.5.2 Complicanze

Le complicanze negli uomini comprendono infezione dell'epididimo e stenosi dell'uretra.
Le complicanze nelle donne comprendono: fibrosi delle tube di Falloppio (con rischio associato di sterilità e gravidanza ectopica), infezione tube di Falloppio (salpingite, che causa dolori nella parte inferiore dell'addome), peritonite e infezione dell'area circostante il fegato (Fitz-Hugh-Curtis che causa dolore, febbre e vomito).
In entrambi i sessi potrebbe manifestarsi artrite reattiva, in genere da una a tre settimane dopo l'infezione iniziale, che interessa una o poche articolazioni [13].

2.5.3 Questionario

Si riportano le informazioni presenti nelle domande del questionario, rivolte ai soggetti che si presentano in consultorio, in grado di distinguere i pazienti affetti da clamidia dai pazienti sani o da pazienti affetti dalle altre MST. Vengono riportate anche le informazioni presenti nel lavoro di tesi precedentemente seguito dal team Bridge [1].

- *Sesso del paziente*: i sintomi sono molto diversi sulla base del genere biologico, inoltre le donne hanno una maggiore probabilità di essere asintomatiche, queste informazioni sono cui estensibili ai pazienti transessuali, ma non ai pazienti transgender.
- *Età del paziente*: gran parte dei casi diagnosticati interessa i giovani (15-24 anni).
- *Data dell'ultimo rapporto sessuale*: i sintomi, se presenti, possono comparire da una a quattro settimane dopo il contagio [13].

- *Paziente in gravidanza*: un contagio da clamidia aumenta il rischio di gravidanza extrauterina e altre problematiche relative alla fertilità a causa dell'insorgenza della Malattia Infiammatoria Pelvica (MIP).
- *Uso di contraccettivi a rilascio di farmaco*: il loro utilizzo aumenta il rischio di contrarre un'infezione da clamidia a livello delle vie urinali inferiori, nonostante riesca a proteggere le pazienti dalla MIP [14].
- *Precedente controllo generale sulle MST (sì/no)*: un no aumenta il rischio per tutte le MST.
- *Rapporti con più di un partner negli ultimi sei mesi*: aumenta il rischio, in particolare se non viene utilizzato il preservativo [15].
- *Almeno uno dei partner conduce una vita promiscua*: aumento del rischio che possa aver contratto la clamidia, rimanendo magari anche asintomatico. Il rischio aumenta ulteriormente in caso il rapporto sia avvenuto meno di tre mesi dal contagio.
- *Contatto diretto con i genitali (sì/no)*: essendo la clamidia un'infezione del tratto urinario, il contatto è indispensabile, ma non è necessario che sia diretto, per cui un sì a questa domanda aumenta il rischio, ma un no non lo annulla.
- *Masturbazione e utilizzo sex toys*: entrambi aumentano il rischio, soprattutto se questi ultimi non vengono utilizzati nella maniera corretta, disinfettati ed entrano a contatto con più parti genitali durante lo stesso rapporto.
- *Uso del preservativo*: comportano sintomi diversi, ma l'assenza del preservativo o un preservativo rotto aumenta il rischio.
- *Rapporti sessuali dietro compenso (sì/no)*: aumenta il rischio in quanto implica un maggior numero di partner sessuali diversi e con condizioni di salute non conosciute.
- *Partner con MST*: aumenta il rischio, maggiormente se senza preservativo.
- *Sintomi possibili*:
  - *Congiuntivite* [13];
  - *Adesione all'apertura del pene* [13];
  - *Infiammazione e gonfiore in zona genitale* [1];
  - *Abbondanti secrezioni/perdite intime* [1];
  - *Forte sensazione di bruciore durante l'orinazione* [13];
  - *Secchezza della gola* [1];
  - *Febbre*, in caso di comparsa di infezioni come complicanze [1] [13].

2.6 Gonorrea

La gonorrea è una malattia causata dal batterio *Neisseria gonorrhoeae*, che infetta le mucose dell'uretra, della cervice uterina, del retto e della gola, o le membrane che rivestono la parte anteriore dell'occhio (congiuntiva e cornea). Gli antibiotici possono curare l'infezione, ma è sempre più comune la resistenza agli antibiotici utilizzati [16].

2.6.1 Sintomi

Solitamente, i sintomi si manifestano soltanto nei punti dell'organismo in cui è avvenuto il contagio, ad esempio secrezioni dal pene o dalla vagina o aumento della frequenza e dell'urgenza di urinare. In alcuni casi, l'infezione si diffonde attraverso il sangue e raggiunge altre parti del corpo, soprattutto la pelle e/o le articolazioni. Raramente la gonorrea colpisce articolazioni, pelle o cuore.
Il sesso anale con un partner infetto può avere come conseguenza una gonorrea rettale, che causa dolore durante le evacuazioni, oppure stipsi, prurito, sanguinamento e secrezioni dal retto. L'area perianale può arrossarsi e infiammarsi e le feci possono essere rivestite da muco e pus.
Il sesso orale con un partner infetto possono determinare la gonorrea della gola (faringite gonococcica). Solitamente, queste infezioni sono asintomatiche, ma può manifestarsi dolore alla gola.
Può svilupparsi la congiuntivite gonococcica, mediante il contatto tra liquidi infetti ed occhi, che causa gonfiore delle palpebre e secrezione oculare purulenta. Negli adulti, spesso solo un occhio viene colpito dall'infezione. I neonati di solito sviluppano l'infezione a entrambi gli occhi e, se non trattata precocemente, può causare cecità.
Circa il 25% degli uomini presentano una sintomatologia minima, un leggero fastidio nell'uretra, seguito in poche ore da un dolore di grado da lieve a grave durante la minzione, una secrezione di pus di colore giallo-verde dal pene e un frequente stimolo a urinare. L'orifizio del pene può arrossarsi e gonfiarsi. È possibile che i batteri si diffondano fino all'epididimo, causando gonfiore e ipersensibilità alla palpazione dello scroto.
Circa il 10-20% delle donne presentano una sintomatologia minima o sono asintomatiche. Alcuni pazienti provano soltanto un leggero fastidio nell'area genitale e presentano secrezione vaginale simile a pus. Tuttavia, in caso di coinvolgimento dell'uretra, sia manfiestano necessità di urinare frequentemente e dolore durante la minzione.
A volte l'infezione si diffonde alle tube di Falloppio causando la salpingite, che provoca forti dolori nella parte inferiore dell'addome, soprattutto durante i rapporti sessuali [16].

2.6.2 Complicanze

In alcune donne, l'infezione si diffonde al peritoneo, causando una peritonite o una malattia infiammatoria pelvica (MIP), che può provocare forti dolori nella parte inferiore dell'addome e aumenta il rischio di sterilità e di gravidanze ectopiche.
È possibile la comparsa della periepatite (o sindrome di Fitz-HughCurtis), principalmente nelle donne, a causa della diffusione dell'infezione al fegato [16].

2.6.3 Questionario

Si riportano le informazioni presenti nelle domande del questionario, rivolte ai soggetti che si presentano in consultorio, in grado di distinguere i pazienti affetti da gonorrea dai pazienti sani o da pazienti affetti dalle altre MST. Vengono riportate anche le informazioni presenti nel lavoro di tesi precedentemente seguito dal team Bridge [1].

- *Sesso del paziente*: le donne hanno una maggiore probabilità di essere asintomatiche rispetto agli uomini, queste informazioni sono cui estensibili ai pazienti transessuali, ma non ai pazienti transgender.
- *Età del paziente*: gran parte dei casi diagnosticati interessa soggetti di età inferiore a 30 anni [2].
- *Data dell'ultimo rapporto sessuale*: i sintomi si presentano in genere dopo 3-8 gg, ma ci possono essere casi di asintomatici che presentano successivamente sintomi più rari e tardivi, ad esempio l'artrite gonococcica.
- *Paziente in gravidanza*: Per una donna che contrae la gonorrea in gravidanza aumenta la probabilità di aborto o di parto prematuro.
- *Precedente controllo generale sulle MST (sì/no)*: un no aumenta il rischio per tutte le MST.
- *Ultimo test MST*: l'assenza di un test per la gonorrea comporta un maggiore rischio di stadi tardivi.
- *Rapporti con più di un partner negli ultimi sei mesi*: la ripetuta esposizione sessuale con partner differenti risulta essere un grande fattore di rischio per l'acquisizione della gonorrea.
- *Tipo di rapporto*: per rapporti eterosessuali, se senza preservativo, c'è il 20% di rischio di trasmissione da donna a uomo, mentre da uomo a donna. Ciò vale anche per i rapporti omosessuali maschili senza preservativo (rischio del 20%) [16].

- *Almeno uno dei partner conduce una vita promiscua*: la ripetuta esposizione sessuale con partner differenti risulta essere un grande fattore di rischio per l'acquisizione della gonorrea.
- *Contatto diretto con i genitali (sì/no)*: è sufficiente il semplice contatto con i genitali per la trasmissione dell'infezione [17].
- *Masturbazione e utilizzo sex toys*: entrambi aumentano il rischio, soprattutto se questi ultimi non vengono utilizzati nella maniera corretta, disinfettati ed entrano a contatto con più parti genitali durante lo stesso rapporto [17].
- *Uso del preservativo*: sesso orale, vaginale e anale comportano sintomi diversi e di diversa entità, ma l'assenza del preservativo o un preservativo rotto aumenta il rischio.
- *Rapporti sessuali dietro compenso (sì/no)*: aumenta il rischio in quanto implica un maggior numero di partner sessuali diversi e con condizioni di salute non conosciute.
- *Partner con MST*: aumenta il rischio, maggiormente se senza preservativo [16].
- *Sintomi possibili*:
  - *Congiuntivite* [1];
  - *Infiammazione e gonfiore in zona genitale* [1];
  - *Abbondanti secrezioni/perdite intime* [1];
  - *Forte sensazione di bruciore durante l'orinazione* [16];
  - *Secchezza della gola* [1];
  - *Febbre*, in caso di comparsa di infezioni come complicanze [1].

Capitolo 3

Classificatore

In questo capitolo vengono discussi i metodi e i risultati delle prove eseguite per creare l'algoritmo di classificazione, implementato inizialmente per conoscere l'entità della distinzione tra pazienti sani e malati e per valutare la riduzione del numero delle domande del questionario al fine di migliorarne l'efficienza:

- i primi tentativi di clusterizzazione mediate reti neurali SOM (Self-Organizing Map), utilizzati principalmente per la definizione della codifica dei dati;
- le prove di clusterizzazione sul massimo dei dati ricevuti;
- l'utilizzo della tecnica di feature selection al fine di selezionare le domande del questionario utili, quindi non ridondanti e prive di informazione, e ridurne il numero;
- la clusterizzazione effettuata con le sole features selezionate in precedenza;
- le prove di clusterizzazione ripetute altre quattro volte, in modo da raggiungere un totale di cinque ripetizioni, al fine di valutare l'entità del miglioramento ottenuto a seguito della feature selection.

3.1 Raccolta ed elaborazione dei dati

In questa sezione verrà trattato il processo per la scelta della codifica dei dati a nostra disposizione e l'ottimizzazione del questionario, migliorato sia dal punto di vista dei contenuti sia nella forma, fino all'utilizzo della clusterizzazione con reti neurali SOM sui primi dati disponibili per testare diverse soluzioni.

3.1.1 Raccolta dati

Il team Bridge sta raccogliendo dati somministrando direttamente ai pazienti un questionario al quale si può accedere anonimamente scannerizzando il QR-code presente sulle locandine di Bridge distribuite ai consultori e centri MST delle 13 regioni italiane che hanno accettato di collaborare al progetto.
A causa della pandemia da Covid-19 attualmente in corso, è stato possibile raccogliere solo 101 dati dai consultori nel periodo trascorso tra gennaio 2021 e agosto 2021. Considerata la scarsa affluenza nei consultori, in questa fase del progetto e per questo lavoro di tesi, si è deciso di utilizzare dati simulati, raccolti parallelamente somministrando lo stesso questionario a medici e specializzandi.

3.1.2 Modifica del questionario

Prima di procedere con il processo di selezione delle domande si è deciso di revisionare, con il supporto di una psicologa neo-laureata, il questionario per migliorarne la comprensibilità e l'inclusività delle opzioni fornite, per fornire spiegazioni aggiuntive relative all'argomento della domanda o del suo utilizzo e, non meno importante, per approcciare domande sensibili e molto personali in maniera più leggera. Si riportano alcuni esempi dei diversi tipi di modifiche apportate:

- *Aggiunta di nuove opzioni per aumentare la comprensibilità e l'inclusività*: nella domanda sul sesso del paziente (Figura 3.1), ad esempio, sono state aggiunte più opzioni di quelle necessarie dal punto di vista informativo in modo da rappresentare ogni possibile situazione del paziente; infatti, sarebbero necessarie solo informazioni sul sesso biologico del paziente. L'aggiunta delle opzioni *"Transessuale M/F"* e *"intersessuale"* e la modifica dell'opzione *"Transgender"* in *"Transgender MtF/FtM"* danno al paziente la possibilità di rispondere con maggiore rapidità sentendosi rispettato. Resta comunque possibile decidere di estrapolare l'informazione necessaria a posteriori, se ritenuto necessario.

Figura 3.1: Domanda "Indica il tuo sesso".

- *Aggiunta della spiegazione dell'utilizzo di alcune domande, che possono sembrare non inerenti e troppo personali durante la compilazione*: per la domanda relativa all'etnia del paziente (Figura 3.2), si è deciso di uniformare i termini in modo da rendere più comprensibili le varie opzioni, inoltre si è aggiunta un'informazione relativa all'utilizzo della risposta, perché si ritiene possibile che l'utente la ritenga non inerente al problema delle MST, a questo scopo si mantiene anche l'opzione *"Non rispondo"*. Le modifiche in questo caso abbattono le barriere che potrebbero crearsi tra la domande e il paziente.

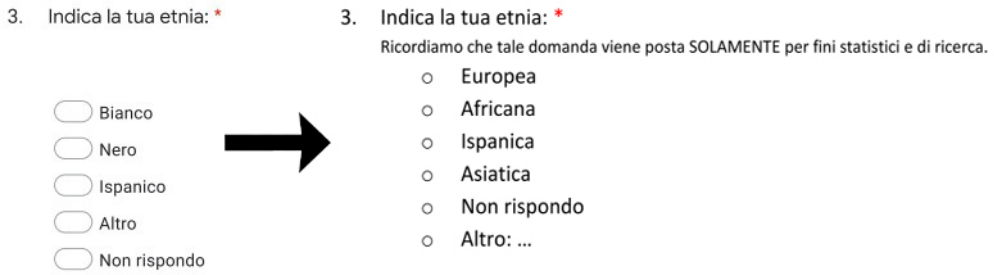


Figura 3.2: Domanda "Indica la tua etnia".

- *Aggiunta di una spiegazione e di informazioni relative all'oggetto della domanda*: in Figura 3.3 si mostra un'aggiunta di una breve spiegazione informativa sul Pap test, in particolare per le giovani ragazze che non lo hanno ancora svolto e non sono ancora state informate (il Pap test viene infatti consigliato per le donne di età superiore ai 25 anni), e per fornire informazioni utili a coloro che accettano di compilare il questionario al fine di renderlo uno strumento utile non solo per la raccolta dati.

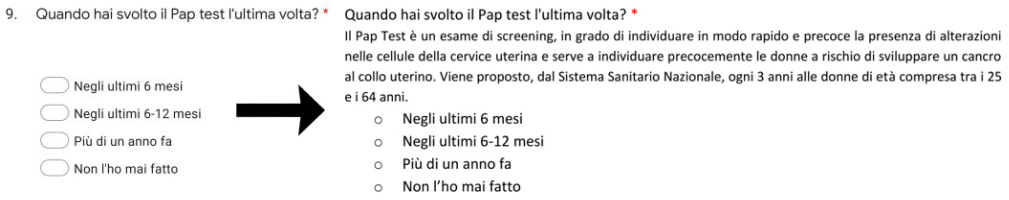


Figura 3.3: Domanda "Quando hai svolto il Pap test l'ultima volta?".

- *Aggiunta di opzioni anche su suggerimento dei pazienti e/o di medici e specializzandi*: è stata segnalata dagli utenti la mancanza dei condilomi ano-genitali tra le opzioni della domanda relativa alle MST avute in passato (Figura 3.4). Dopo un ulteriore controllo è stata aggiunta anche l'Epatite A. Il questionario,

infatti, rimane aperto a possibili modifiche, con il fine di poter apportare miglioramenti a seguito di suggerimenti di medici e pazienti.

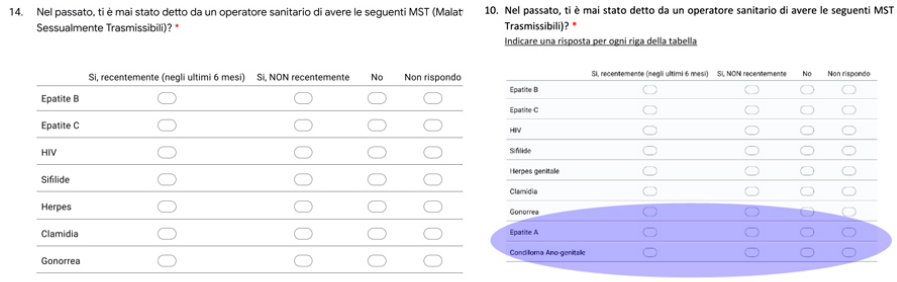


Figura 3.4: Domanda "Nel passato, ti è mai stato detto da un operatore sanitario di avere le seguenti MST?".

- *Si è deciso anche di modificare interamente la domanda, nella forma e nelle informazioni richieste*: nella domanda presentata in Figura 3.5, essendo l'aspetto importante il tipo di rapporti, le opzioni presentate racchiudono tutte le opzioni possibili. Inoltre, tali modifiche rendono la domanda più chiara per l'utente e non sarà quindi necessario agire successivamente sui dati raccolti, combinando la risposta di questa domanda con quella relativa al sesso del paziente, per definirne il tipo di rapporto.

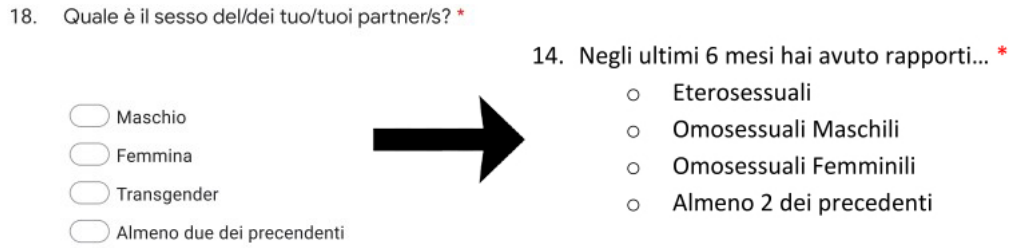


Figura 3.5: Domanda "Negli ultimi 6 mesi hai avuto rapporti...".

- *Inserimento di una nuova porzione della domanda*: all'interno della domanda relativa alla masturbazione (Figura 3.6), costituita da quattro possibili risposte, si è deciso di inserire altre tre opzioni per l'utilizzo di sex toys. Tale informazione è stata aggiunta a seguito di una assemblea scolastica, nella quale il team Bridge ha presentato il progetto, durante la quale sono emersi con frequenza i dubbi e le domande relative ai sex-toys. In questo caso i dati vengono successivamente separati in fase di codifica e, per i dati precedenti alla modifica, si considera che i pazienti abbiano scelto l'opzione *"Non rispondo"*.

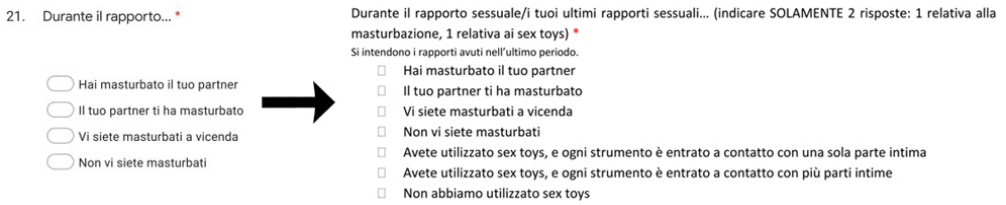


Figura 3.6: Domanda relativa alla masturbazione e all'uso di sex toys.

- *Modifica relativa alle immagini*: nelle quattro domande dedicate all'identificazione delle eruzioni cutanee correlate a condiloma ano-genitale, sifilide (sifiloma e eritema a macchie rosse) ed herpes genitale, sono state cambiate alcune immagini al fine di rappresentare casi più lievi ed è stata modificata la spiegazione medica della casistica con frasi più semplici e comprensibili (Figura 3.7), al fine di agevolare il target di pazienti che solitamente si reca presso i consultori.

Figura 3.7: Domanda relativa all'eritema a macchie rosa tipico del secondo stadio della Siflide.

3.1.3 Trattamento dei dati

Relativamente al trattamento dei dati inviati dagli utenti durante l'utilizzo dell'app e quindi la compilazione del questionario, nel precedente lavoro di tesi [1] è stato descritto sia lo *User Agreement* per la mobile-app, sia il trattamento dei dati provenienti dalla locandina affissa nei consultori.
Considerando l'evoluzione del progetto Bridge, da mobile app a web app, in questa nuova versione lo *User Agreement* verrà proposto all'utente ogni volta in cui accederà al sito e deciderà di usufruire della compilazione del questionario per ottenere informazioni e suggerimenti relativi alle MST e alla proprio salute sessuale.

Essendo un contratto che assegna all'utente la licenza d'uso dei servizi offerti dal servizio Bridge, nei termini stabiliti dal contratto stesso, l'utente deve visionare e accettare i termini e le condizioni dichiarate nell'accordo per poter usufruire del servizio. Rimangono inalterate le condizioni proposte dal team Bridge all'utente del servizio [1] in termini di trattamento dati e avvertenze sull'utilizzo improprio dei servizi e delle informazioni suggerite.
Per quanto riguarda i dati prelevati presso i consultori, non si pongono problemi per il trattamento dati e per la privacy, in quanto esiste un rapporto diretto tra il paziente che decide volontariamente di scannerizzare il QR-code e il team Bridge, che riceve anonimamente le risposte e non ha alcun modo per risalire all'identità del soggetto. Essendo la compilazione volontaria da parte dell'utente, il personale presente nei consultori e centri MST non ha alcuna informazione circa i pazienti che decidono di prendere parte alla raccolta dati. Si è deciso di evitare la ricezione dei dati in forma indiretta, quindi assegnando il compito ai consultori di prelevare i dati da inviare in seguito al team Bridge, per evitare l'intercettazione dei dati da parte di terzi e la possibilità di un incrocio di dati che possa consentire l'identificazione del paziente. Inoltre, solamente il team Bridge ha accesso ai dati.
Infine, nel questionario, come descritto nel Paragrafo 3.1.2, sono state aggiunte maggiori spiegazioni relative all'utilizzo delle riposte. All'inizio della compilazione, si richiede all'utente di leggere una breve informativa in cui viene sottolineato come i dati vengano tramessi in modo assolutamente anonimo, come descritto sopra, e si avverte l'utente che alcune domande possono risultare personali e intime, ma che vengono poste solamente per fini statistici, e che vengono mostrate immagini relative alle IST. Viene riportato lo screen che compare ai soggetti che decidono di scannerizzare il QR-code della locandina per rispondere al questionario:

"Nel questionario che stai per compilare ti verranno poste alcune domande PERSONALI (età, Paese di provenienza, etnia, etc.) e INTIME (sesso del/i partner, informazioni sul tipo di rapporti sessuali, etc.) SOLAMENTE per fini statistici e di ricerca. Inoltre, ti verranno mostrate alcune immagini sensibili relative a possibili eruzioni cutanee. Ti ricordiamo che le risposte saranno assolutamente ANONIME e che in nessun modo potremo risalire all'utente."

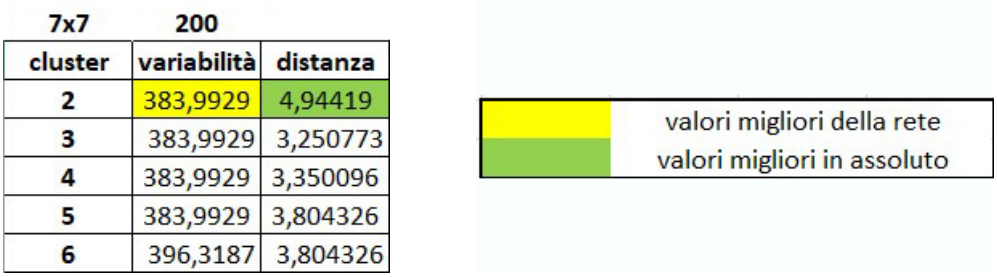
3.1.4 Elaborazione e codifica

Le prime prove sono state utilizzate per testare metodi di codifica delle riposte al questionario e contemporaneamente ottenere dei risultati che potessero creare una base di partenza per poi confrontare i risultati ottenuti da un numero maggiore di dati.
In questa fase sono stati usati due diversi dataset: il primo costituito da 116 pazienti di cui 16 sani, il secondo costituito da 139 pazienti di cui 25 sani. Si era deciso di

ripetere le prove più di una volta per aumentare la consistenza; perciò, si è deciso di includere anche gli ultimi dati ricevuti in quel periodo, considerato che la poca differenza crea varietà mantenendo le prove confrontabili.
Ogni dataset è stato suddiviso in tre sezioni, ovvero le sezioni che suddividono il questionario stesso: vita personale, vista sessuale e sintomi. Le risposte ottenute sono quindi state analizzate in tre diversi modi, utilizzando sezioni singole o accorpate: intero questionario (66 *features*), vita sessuale (17 *features*), vita sessuale e sintomi (35 *features*).

Metodo di clusterizzazione

Si è scelto di testare i tre possibili metodi di codifica utilizzando le reti neurali SOM di dimensione da 4x4 a 8x8 per il *training* dei dati e il dendrogramma per l'effettiva clusterizzazione, con tagli da 2 a 6. Per ognuno dei tagli del dendrogramma, al fine di suddividere i dati in cluster, sono stati valutati due parametri: la variabilità intra-cluster massima e la distanza inter-cluster minima. Successivamente sono state scelte le clusterizzazioni migliori per ogni rete testata, ovvero quelle che presentavano la variabilità intra-cluster minima e/o la distanza inter-cluster massima (Figura 3.8).



7x7	200	
cluster	variabilità	distanza
2	383,9929	4,94419
3	383,9929	3,250773
4	383,9929	3,350096
5	383,9929	3,804326
6	396,3187	3,804326

Figura 3.8: Esempio di valori per i vari tagli e legenda

	cluster1	cluster2
sani	9	16
malati	23	91
tot	32	107

Figura 3.9: Confusion Matrix corrispondente al taglio.

Al fine di valutare le prestazioni della clusterizzazione, sono stati confrontati i *clusters* ottenuti con le classi dei pazienti, in particolare costruendo le *confusion matrix* (di cui è presente un esempio in Figura 3.9) e valutando due indicatori:

- Differenza tra la percentuale dei pazienti sani nel *cluster* e la percentuale del totale dei pazienti nel *cluster*, in modo da tener conto non solo della divisione dei pazienti sani e pazienti affetti da MST, ma anche delle dimensioni dei *cluster*;

- Percentuale di pazienti sani rispetto al totale dei pazienti del *cluster*.

Codifica dei dati

Le risposte ottenute del questionario vengono esportate in un file .CSV da cui i dati vengono letti come una tabella e successivamente convertiti in dati di tipo categorical mediante la funzione Matlab *"categorical.m"*. La serie di prove, eseguite con il metodo precedentemente indicato, è stata utilizzata anche per il confronto di due metodi differenti di codifica dei dati.

- *Prove "no encoder"*: indica le prove effettuate con la codifica ottenuta dalla funzione di Matlab *"categorical.m"*, la quale però presentava un problema: questa funzione assegna i valori alle risposte uguali incrementandone il valore sulla base della successione temporale con la quale si presentano per la prima volta all'algoritmo (ad esempio la risposta "Non rispondo" si ripete in varie domande, ha lo stesso valore assegnato, e il valore indica in quale momento la risposta è comparsa nel dataset fornito). Tale metodo porta ad avere valori non omogenei per le risposte di una singola domanda, e pesi random alle singole risposte delle domande, come mostrato in Figura 3.10.

- *Prove "encoder"*: indica le prove effettuate con un nuovo encoder creato appositamente per questi dati. Si è quindi risolto il problema in oggetto creando inizialmente una matrice (in parte mostrata in Figura 3.11) con tutte le risposte possibili e riordinate (ove possibile) a seconda della natura della domanda: da *"Non rispondo"* a *"Sì"*, secondo una frequenza crescente, o una probabilità crescente.

- *Prove "norm"*: indica le prove effettuate con dati normalizzati. Le risposte codificate sono state successivamente normalizzate sulla base del valore assegnato, al fine di ottenere per ogni risposta lo stesso peso minimo e massimo possibile, ovvero tra 0 e 1.

Risultati dei test preliminari

Una volta terminata la serie di prove di clusterizzazione per ognuno dei metodi sopracitati, si è proceduto a confrontare i valori dei parametri ottenuti e la

	1	2	3	4	5
1	4	28	2	3	8
2	4	28	2	3	8
3	6	1	4	5	2
4	6	2	4	5	1
5	8	1	3	5	4
6	8	2	4	5	1
7	8	1	4	5	7
8	8	2	4	5	1
9	8	1	4	5	2
10	6	1	3	4	12
11	8	1	3	4	7
12	5	2	3	4	1
13	9	1	3	4	8
14	7	1	3	6	4
15	6	1	2	5	3
16	8	1	3	6	4

Figura 3.10: Esempio dei valori ottenuti con la funzione “categorical”.

qualità della clusterizzazione mediante i due indicatori precedentemente descritti, al fine di scegliere il metodo più adatto ai dati a nostra disposizione.
Il confronto è stato ripetuto sia sul dataset contenente 116 pazienti, sia sul dataset contenente 139 pazienti, al fine di ridurre l’influenza della componente casuale insita nel metodo delle reti SOM.
Si riportano quindi quattro diverse tabelle per ognuno dei dataset: le prime due riportano i valori migliori di variabilità intra-cluster e distanza inter-cluster per ognuna delle reti (Figura 3.12 - 3.15), le ultime due riportano il valore degli indicatori relativi al cluster migliore di ogni rete analizzata per il solo dataset costituito da 139 pazienti (Figura 3.16 - 3.17).
In ogni tabella è possibile confrontare i risultati ottenuti con i tre metodi analizzati per l’intero questionario e le due sezioni.

2 Var2	3 Var3	4 Var4	5 Var5	6 Var6	7 Var7	8 Var8
'meno di 20 anni'	'Non rispondo'	'Europa occidentale'	'Nelle ultime 2 ore'	''	''	''
'21-30 anni'	'Altro'	'Europa orientale'	'Oggi più di 2 ore fa'	'No'	'No'	'Non l''ho mai fatto'
'31-40 anni'	'Europea'	'Nord America'	'Ieri - 2 giorni fa'	'Forse'	'Si'	'Negli ultimi 6 mesi'
'41-50 anni'	'Africana'	'America centrale'	'3-5 giorni fa'	'Si'	'0'	'Negli ultimi 6-12 mesi'
'51-60 anni'	'Ispanica'	'Sud America'	'6-10 giorni fa'	'0'	'0'	'Più di un anno fa'
'superiore ai 61 anni'	'Asiatica'	'Australia'	'11-15 giorni fa'	'0'	'0'	'0'
'0'	'0'	'Africa'	'16-20 giorni fa'	'0'	'0'	'0'
'0'	'0'	'Asia'	'21-30 giorni fa'	'0'	'0'	'0'
'0'	'0'	'0'	'Da 1 a 12 mesi fa'	'0'	'0'	'0'
'0'	'0'	'0'	'Più di 12 mesi fa'	'0'	'0'	'0'
'0'	'0'	'0'	'0'	'0'	'0'	'0'
'0'	'0'	'0'	'0'	'0'	'0'	'0'

Figura 3.11: Parte della matrice "Answer".

- *Variabilità intra-cluster*
  I valori evidenziati in tabella (Figura 3.12 - 3.13) indicano i valori migliori tra le reti analizzate.
  Come si può riscontrare, per entrambi i dataset, la variabilità intra-cluster diminuisce progressivamente al miglioramento della codifica e normalizzazione dei dati: è un risultato atteso, in quanto con l'introduzione del nuovo metodo di codifica si è proceduto a ridurre i valori dei pesi massimi possibili per le risposte di ogni domanda, che sono stati successivamente normalizzati per far sì che ogni domanda avesse uguale peso. Si riscontra quindi un vantaggio nell'avere dati molto simili all'interno dello stesso cluster.

		no encoder	encoder	norm
intero quest.	5x5	3049,8000	785,7173	229,6865
	6x6	4498,6000	910,0923	231,8465
	7x7	4498,6000	778,2991	232,3003
	8x8	4498,6000	789,2580	239,0007
vita sessuale	4x4	221,1029	183,9391	37,9417
	5x5	233,8927	169,7508	36,9217
vita sess. + sintomi	4x4	790,9119	351,6949	94,4030
	8x8	1132,6000	349,6746	102,3641

Figura 3.12: Variabilità intra-cluster del dataset con 116 pazienti.

		no encoder	encoder	norm
intero quest.	4x4	3322,1126	792,8367	238,4866
	5x5	3380,9073	782,8679	230,5386
	6x6	3380,9073	779,5649	232,3820
	8x8	3380,9073	786,5509	230,6905
vita sessuale	6x6	217,8732	175,9627	37,1966
	7x7	236,3340		36,9160
	8x8	241,9729	163,3439	36,7032
vita sess. + sintomi	4x4	795,4578	367,6490	94,4548
	7x7	1280,3042	383,9929	93,9211

Figura 3.13: Variabilità intra-cluster del dataset con 139 pazienti.

- *Distanza inter-cluster*
  I valori evidenziati in tabella (Figura 3.14 - 3.15) indicano i valori migliori tra le reti analizzate.
  Come precedentemente osservato per la variabilità intra-cluster, anche per il parametro di distanza inter-cluster si ha una sensibile diminuzione dei valori per entrambi i dataset, ma in questo caso i valori ottenuti con la normalizzazione risultano molto più simili tra le diverse reti testate. In questo caso si osserva un effetto negativo sui dati, in quanto si vuole ottenere un insieme di cluster molto separati tra loro per garantire buone prestazioni delle reti.

		no encoder	encoder	norm
intero quest.	5x5	26,274000	5,822642	1,311556
	6x6	50,731100	5,955427	1,616938
	7x7	50,731100	5,898929	1,626090
	8x8	50,731100	6,250006	1,403852
vita sessuale	4x4	5,847400	5,191390	1,191056
	5x5	6,141500	4,589081	1,212105
vita sess. + sintomi	4x4	11,863900	4,139009	1,353839
	8x8	23,659400	4,087373	1,496231

Figura 3.14: Distanza inter-cluster del dataset con 116 pazienti.

		no encoder	encoder	norm
intero quest.	4x4	29,584767	6,222722	1,439416
	5x5	30,914768	6,033983	1,679078
	6x6	30,914768	5,776445	1,363425
	8x8	30,914768	6,141772	1,623048
vita sessuale	6x6	6,025968	4,859711	1,210994
	7x7	6,091805		1,191742
	8x8	6,138144	4,248635	1,325418
vita sess. + sintomi	4x4	11,650143	4,677381	1,338390
	7x7	30,606759	4,944190	1,323838

Figura 3.15: Distanza inter-cluster del dataset con 139 pazienti.

- *Clusterizzazione*
  Per entrambi i dataset (in Figura 3.16 - 3.17) vengono mostrati i risultati relativi al dataset con 139 pazienti) il metodo di codifica migliore risulta essere quello che prevede una normalizzazione: esso, infatti, presenta la clusterizzazione migliore sia per il primo indicatore (con il 50% per il primo dataset e il 55,56% per il secondo), sia per il secondo indicatore (con il 75% per il primo dataset e il 55,56% per il secondo). Pertanto, si è deciso di proseguire le prove utilizzando la codifica creata appositamente e normalizzando i dati.

3.2 Algoritmo di clusterizzazione

Verrà ora discusso nel dettaglio il metodo e i risultati dell'algoritmo di clusterizzazione ottenuta dal terzo e ultimo dataset costituito da 314 dati simulati, distribuiti come descritto in Figura 3.18.

		no encoder	encoder	norm
intero quest.	4x4	10,25%	1,77%	20,83%
	5x5	✓ 12,26%	15,25%	17,73%
	6x6	✓ 12,26%	✓ 23,48%	✓ 28,66%
	8x8	✓ 12,26%	11,65%	11,97%
vita sessuale	6x6	9,71%	7,54%	3,97%
	7x7	✓ 21,46%		5,09%
	8x8	21,16%	✓ 10,19%	✓ 10,71%
vita sess. + sintomi	4x4	✓ 13,85%	6,50%	✓ 21,55%
	7x7	1,44%	✓ 12,98%	17,32%

Figura 3.16: Confronto clusterizzazione per il dataset con 139 pazienti, che riporta la differenza tra percentuale di pazienti sani calcolata rispetto al totale (25) e la percentuale di pazienti presenti all'interno del cluster calcolata rispetto al totale (139).

		no encoder	encoder	norm
intero quest.	4x4	20,00%	23,46%	26,67%
	5x5	✓ 20,43%	22,22%	23,46%
	6x6	✓ 20,43%	✓ 26,03%	✓ 36,84%
	8x8	✓ 20,43%	21,05%	21,35%
vita sessuale	6x6	21,42%	✓ 23,53%	19,10%
	7x7	✓ 23,26%		19,35%
	8x8	22,45%	23,40%	✓ 20,35%
vita sess. + sintomi	4x4	✓ 21,18%	21,95%	✓ 27,11%
	7x7	17,51%	✓ 28,12%	23,68%

Figura 3.17: Confronto clusterizzazione con percentuale di pazienti sani rispetto al totale nel cluster (139 pazienti).

Di questi, i due pazienti affetti da HPV (Papilloma Virus Umano), sono tenuti separati dalla classe "Altro", in quanto il condiloma ano-genitale deriva dalla contrazione di alcune tipologie del virus HPV, ma, a causa dell'assenza dei condilomi in questi casi, è stata creata una classe separata. Inoltre, per il paziente affetto da piattole, essendo un parassita, si è stabilito di non includerlo nella classe "Altro", ma si è deciso comunque di non eliminarlo in quanto non ci sono limiti nel questionario relativamente alle patologie (in seguito si potrebbero ricevere altri dati con questa classe) e ogni informazione potrebbe risultare utile per la prima divisione di pazienti sani e pazienti affetti da MST. I pazienti che non hanno consultato il

medico sono stati utilizzati nell'allenamento della rete neurale SOM, ma non è stata verificata la loro clusterizzazione, in quanto non possono contribuire né alla separazione tra pazienti sani e pazienti malati né a una futura distinzione tra le patologie.
Il dataset qui descritto verrà mantenuto in questa forma fino alla conclusione del lavoro e, pertanto, i risultati del procedimento di clusterizzazione descritto in questa sezione costituirà l'elemento di confronto per le prove successive. L'obiettivo della clusterizzazione è di valutare, in una forma preliminare, se i dati prelevati possono essere significativi per il riconoscimento iniziale dei pazienti affetti da una MST, ovvero se è possibile una divisione in cluster di soggetti simili, quindi sani e affetti da MST. Per questo motivo dall'algoritmo ci si aspetta una divisione preferenziale in due clusters. Nonostante ciò, nelle prime prove, qualora fossero preferenziali tagli con maggiore numerosità di clusters per la divisione degli individui, è stata svolta una maggiore analisi circa la capacità di divisione della rete, per valutare se anche una divisione per patologie sarà possibile successivamente con un dataset più ampio. Per completezza, questa analisi è stata poi estesa anche ai tagli in due clusters, per verificare se la divisione delle patologie poteva essere soddisfacente anche in questo caso e, se ci fossero eventuali relazioni tra le patologie stesse, considerato che alcune presentano sintomatologie simili, come ad esempio la Clamidia e la Gonorrea.

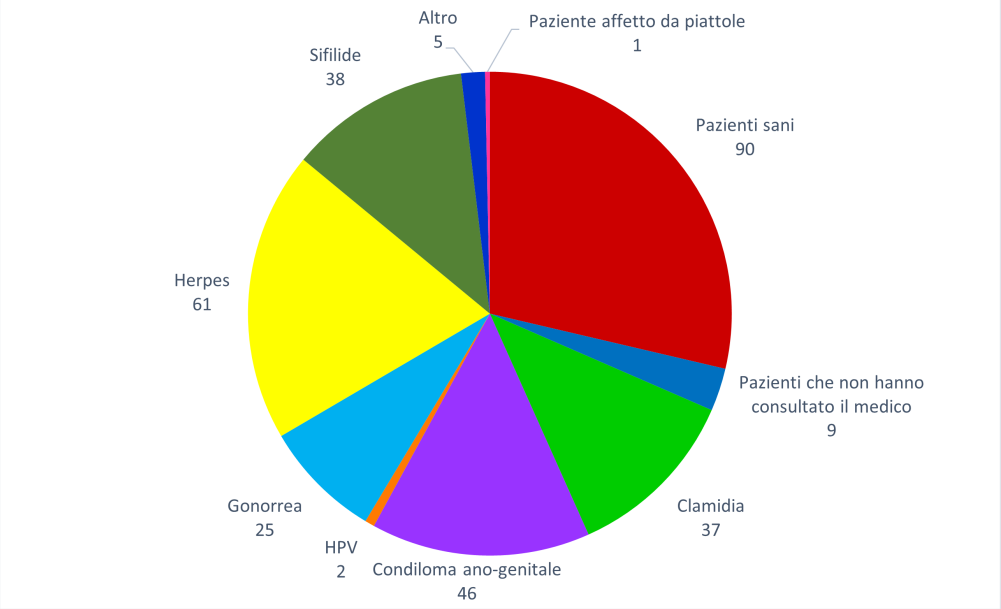


Figura 3.18: Grafico che mostra la distribuzione delle classi dei 314 pazienti simulati.

3.2.1 Metodo

L'analisi dei 314 dati simulati ricevuti è stata effettuata nuovamente secondo le tre diverse divisioni in sezioni delle domande (*features*): intero questionario (66 *features*), vita sessuale (17 *features*), vita sessuale e sintomi (35 *features*).
Come detto in precedenza, per la clusterizzazione sono state utilizzate le reti neurali SOM, allenate impostando i seguenti parametri, scelti arbitrariamente in base alla dimensione e alla tipologia del dataset utilizzato e al numero di *clusters* che ci si aspetta:

- dimensione delle reti da 4x4 a 8x8;
- vicinato uguale a 1;
- 200 iterazioni.

Sono stati provati diversi tagli del dendrogramma, in questo caso da 2 a un massimo di 7, fermandosi quando i parametri diminuivano sensibilmente per due volte consecutive, oppure quando i valori diminuivano significativamente da un taglio all'altro, in modo che fosse impossibile che aumentassero abbastanza da migliorare i valori precedenti.
Questi criteri sono stati elaborati durante le prove precedenti e, in questa fase, si è notato che la distanza inter-cluster è più determinante della variabilità intra-cluster, la quale quasi sempre presenta il suo valore migliore (cioè il valore minore) nel taglio minimo (2 *clusters*), per poi rimanere uguale e/o aumentare progressivamente.
Durante l'analisi dei diversi tagli, sono stati considerati i valori peggiori dei due parametri, quindi la variabilità intra-cluster massima e la distanza inter-cluster minima, in modo da valutare i diversi tagli con maggiore consistenza. Una volta terminata l'analisi dei tagli di un'intera rete, sono stati selezionati i tagli con i valori migliori, cioè la variabilità intra-cluster minima e la distanza inter-cluster massima.
Successivamente, sono state verificate le clusterizzazioni dei soli tagli selezionati verificando inizialmente la distribuzione dei pazienti sani (informazione utile per la divisione ottimale, quindi tra pazienti sani e pazienti affetti da MST), successivamente per completezza la distribuzione delle diverse patologie (informazione utile per definire una futura possibilità di distinzione delle patologie stesse).
Sono state ottenute le tabelle, dette *Confusion Matrix*, indicanti le distribuzioni dei pazienti sani e pazienti affetti da MST, dalle quali sono stati estrapolati due diversi indicatori:

- la differenza tra la percentuale di pazienti sani presenti all'interno del *cluster* (n° sani/90), calcolata sul totale di pazienti sani (90), e la percentuale della numerosità del *cluster* (n° pazienti/314), calcolata sul totale dei pazienti (314).

- la percentuale costituita da pazienti sani rispetto alla numerosità del *cluster* (n° sani/n° pazienti).

Il primo indicatore è utile a valutare la distribuzione dei pazienti sani in relazione alla numerosità del *cluster* in oggetto, il secondo a sottolineare l'eventuale prevalenza di pazienti sani indipendentemente dalla numerosità del *cluster*.
A causa delle diverse numerosità dei pazienti affetti dalle diverse patologie, si è deciso di valutare singolarmente le distribuzioni nei soli *clusters* selezionati, da queste è stato ricavato un grafico che verrà descritto nel relativo paragrafo.

3.2.2 Risultati

Verranno ora riportati i risultati per ognuna delle sezioni e ognuna delle reti SOM utilizzate e precedentemente descritte.
Per ognuna di esse saranno riportati gli elementi più rilevanti: innanzitutto i tagli che presentano le migliori variabilità intra-cluster e/o distanza inter-cluster della rete, per ogni taglio vengono riportati entrambi indipendentemente che siano i valori migliori, ma viene indicato ogni qual volta sia presente solo un parametro migliore; saranno inoltre mostrati i valori dei due indicatori per il *cluster/s* migliore/i della rete. Infine, viene indicato quando un parametro e/o un indicatore risultano essere i migliori dell'intera sezione, in questo modo è possibile valutare le performances ottenute con ognuna di esse.

Prove con l'intero questionario

Rete SOM 4x4
È stato selezionato solo il taglio a 2 (Figura 3.19 - 3.20), con i seguenti valori:

- variabilità intra-cluster = 221,850
- distanza inter-cluster = 1,597

La clusterizzazione indica che il *cluster* 1 contiene:

- 203 individui su 314 individui totali = 64,65%
- 75 pazienti sani su 90 pazienti sani totali = 83,33%

La differenza è quindi 18,68% una delle due migliori della sezione "intero questionario" (inferiore dello 0,007%).
La percentuale di pazienti sani nel *cluster* è 36,94% (75 pazienti sani su 203 individui).

	cluster1	cluster2
sani	75	15
malati	128	96
tot	203	111

Figura 3.19: Confusion matrix relativa alla sezione "intero questionario", rete 4x4, taglio a 2.

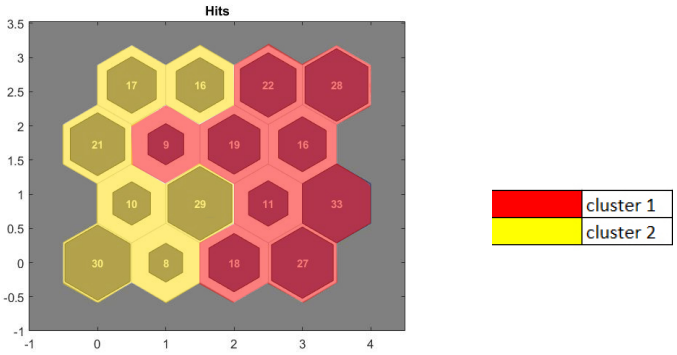


Figura 3.20: Mappa della rete SOM relativa alla sezione "intero questionario", rete 4x4, taglio a 2, con evidenziati i neuroni relativi ai clusters.

Rete SOM 5x5

È stato selezionato solo il taglio a 2 (Figura 3.21 - 3.22), con i seguenti valori:

- variabilità intra-cluster = 221,185
- distanza inter-cluster = 1,526

La variabilità intra-cluster risulta essere la migliore della sezione "intero questionario".

La clusterizzazione indica che il *cluster* 1 contiene:

- 196 individui su 314 individui totali = 62,42%
- 73 pazienti sani su 90 pazienti sani totali = 81,11%

La differenza è quindi 18,69% una delle due migliori della sezione "intero questionario" (superiore dello 0,007%).
La percentuale di pazienti sani nel *cluster* è 36,95% (73 pazienti sani su 196 individui).

	cluster1	cluster2
sani	73	17
malati	123	101
tot	196	118

Figura 3.21: Confusion matrix relativa alla sezione "intero questionario", rete 5x5, taglio a 2.

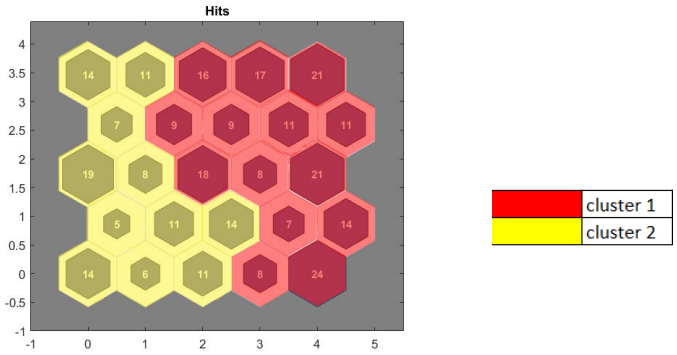


Figura 3.22: Mappa della rete SOM relativa alla sezione "intero questionario", rete 5x5, taglio a 2, con evidenziati i neuroni relativi ai clusters.

Rete SOM 6x6

È stato selezionato il taglio a 2 (Figura 3.23 - 3.24), con la migliore variabilità intra-cluster della rete:

- variabilità intra-cluster = 221,648
- distanza inter-cluster = 1,433

e il taglio a 3 (Figura 3.25 - 3.26), con la migliore distanza inter-cluster della rete:

- variabilità intra-cluster = 233,597
- distanza inter-cluster = 1,569

La clusterizzazione del taglio a 2 indica che il *cluster* 1 contiene:

- 176 individui su 314 individui totali = 56,05%
- 67 pazienti sani su 90 pazienti sani totali = 74,44%

La differenza è quindi 18,39%.
La percentuale di pazienti sani nel *cluster* è 38,07% (67 pazienti sani su 176 individui), il migliore della sezione "intero questionario".

Per quanto riguarda la clusterizzazione del taglio a 3, il *cluster* migliore è lo stesso del taglio a 2, perciò i valori risultano gli stessi.

	cluster1	cluster2
sani	67	23
malati	109	115
tot	176	138

Figura 3.23: Confusion matrix relativa alla sezione "intero questionario", rete 6x6, taglio a 2.

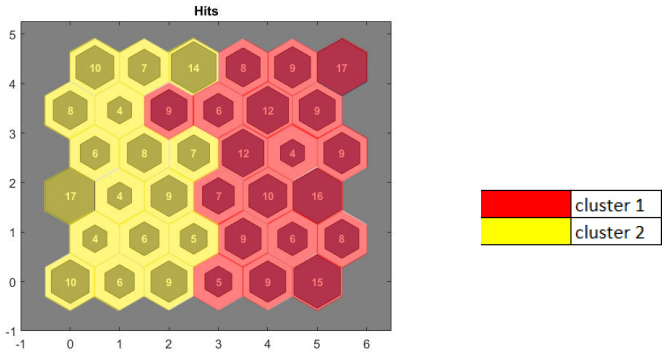


Figura 3.24: Mappa della rete SOM relativa alla sezione "intero questionario", rete 6x6, taglio a 2, con evidenziati i neuroni relativi ai clusters.

	cluster1	cluster2	cluster3
sani	5	18	67
malati	30	85	109
tot	35	103	176

Figura 3.25: Confusion matrix relativa alla sezione "intero questionario", rete 6x6, taglio a 3.

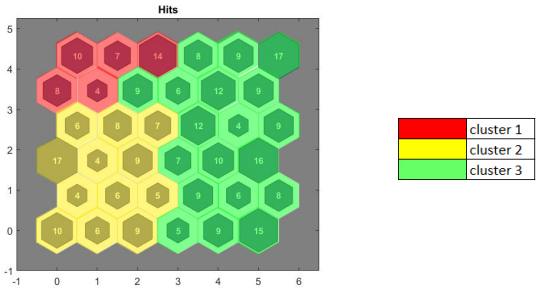


Figura 3.26: Mappa della rete SOM relativa alla sezione "intero questionario", rete 6x6, taglio a 3, con evidenziati i neuroni relativi ai clusters.

Rete SOM 7x7

È stato selezionato solo il taglio a 2 (Figura 3.27 - 3.28), con i seguenti valori:

- variabilità intra-cluster = 221,367
- distanza inter-cluster = 1,622

	cluster1	cluster2
sani	17	73
malati	90	134
tot	107	207

Figura 3.27: Confusion matrix relativa alla sezione “intero questionario”, rete 7x7, taglio a 2.

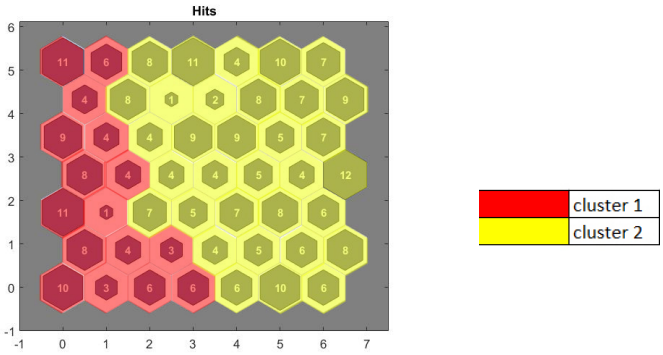


Figura 3.28: Mappa della rete SOM relativa alla sezione “intero questionario”, rete 7x7, taglio a 2, con evidenziati i neuroni relativi ai clusters.

La clusterizzazione indica che il *cluster* 2 contiene:

- 207 individui su 314 individui totali = 65,92%
- 73 pazienti sani su 90 pazienti sani totali = 81,11%

La differenza è quindi 15,19%.
La percentuale di pazienti sani nel *cluster* è 35,26% (73 pazienti sani su 207 individui).

Rete SOM 8x8
È stato selezionato solo il taglio a 2 (Figura 3.29 - 3.30, con i seguenti valori:

- variabilità intra-cluster = 222,078
- distanza inter-cluster = 1,647

La distanza inter-cluster risulta essere la migliore della sezione “intero questionario”.

La clusterizzazione indica che il *cluster* 2 contiene:

- 215 individui su 314 individui totali = 68,47%
- 75 pazienti sani su 90 pazienti sani totali = 83,33%

	cluster1	cluster2
sani	15	75
malati	84	140
tot	99	215

Figura 3.29: Confusion matrix relativa alla sezione “intero questionario”, rete 8x8, taglio a 2.

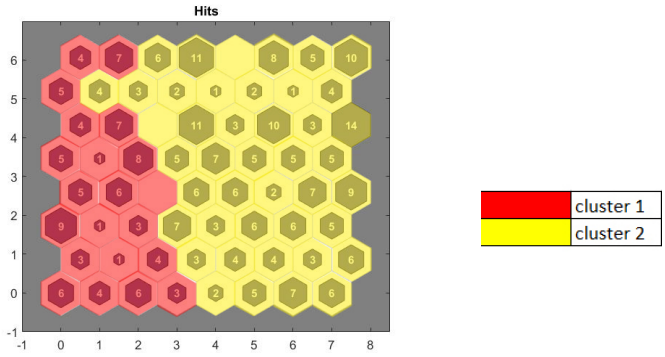


Figura 3.30: Mappa della rete SOM relativa alla sezione “intero questionario”, rete 8x8, taglio a 2, con evidenziati i neuroni relativi ai clusters.

La differenza è quindi 14,86%.
La percentuale di pazienti sani nel *cluster* è 34,88% (75 pazienti sani su 215 individui).

Le prove di questa sezione hanno quindi indicato quattro reti diverse per i migliori parametri e i migliori indicatori globali:

- 5x5 per la variabilità intra-cluster;
- 8x8 per la distanza inter-cluster;
- 4x4 e 5x5 per il primo indicatore, con una differenza minima, dello 0,007%;
- 6x6 per la percentuale di sani nel *cluster*;

Nessuno dei due indicatori della clusterizzazione risultano i migliori tra le sezioni.

Prove con le domande relative alla vita sessuale

Rete SOM 4x4

È stato selezionato il taglio a 2 (Figura 3.31 - 3.32), con la migliore variabilità intra-cluster della rete:

- variabilità intra-cluster = 34,258

- distanza inter-cluster = 1,114

	cluster1	cluster2
sani	59	31
malati	107	117
tot	166	148

Figura 3.31: Confusion matrix relativa alla sezione "vita sessuale", rete 4x4, taglio a 2.

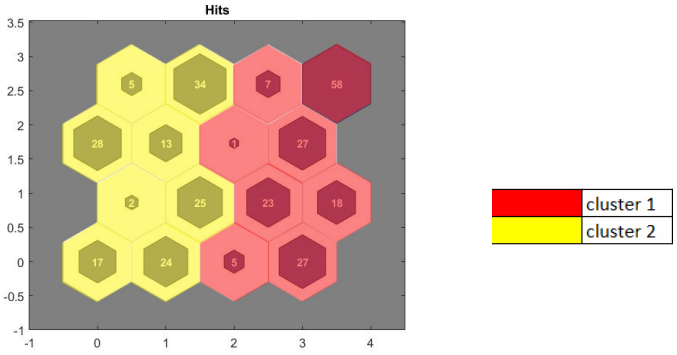


Figura 3.32: Mappa della rete SOM relativa alla sezione "vita sessuale", rete 4x4, taglio a 2, con evidenziati i neuroni relativi ai clusters.

e il taglio a 3 (Figura 3.33 - 3.34), con la migliore distanza inter-cluster della rete:

- variabilità intra-cluster = 34,258
- distanza inter-cluster = 1,150

	cluster1	cluster2	cluster3
sani	15	16	59
malati	53	64	107
tot	68	80	166

Figura 3.33: Confusion matrix relativa alla sezione "vita sessuale", rete 4x4, taglio a 3.

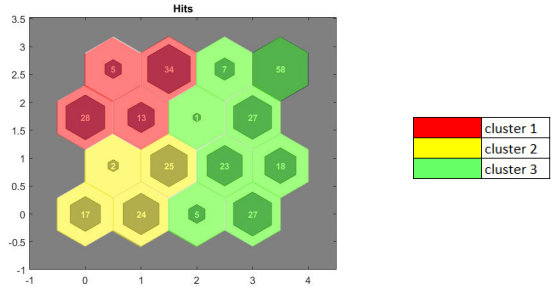


Figura 3.34: Mappa della rete SOM relativa alla sezione "vita sessuale", rete 4x4, taglio a 3, con evidenziati i neuroni relativi ai clusters.

La clusterizzazione del taglio a 2 indica che il *cluster* 1 contiene:

- 166 individui su 314 individui totali = 52,87%
- 59 pazienti sani su 90 pazienti sani totali = 65,56%

La differenza è quindi 12,69%.
La percentuale di pazienti sani nel *cluster* è 35,54% (59 pazienti sani su 166 individui).

Per quanto riguarda la clusterizzazione del taglio a 3, il *cluster* migliore è lo stesso del taglio a 2, perciò i valori risultano gli stessi.

Rete SOM 5x5
È stato selezionato solo il taglio a 2 (Figura 3.35 - 3.36), con i seguenti valori:

- variabilità intra-cluster = 33.286
- distanza inter-cluster = 1,385

Entrambi risultano essere i migliori della sezione "vita sessuale".

La clusterizzazione indica che il *cluster* 2 contiene:

- 256 individui su 314 individui totali = 81,53%
- 85 pazienti sani su 90 pazienti sani totali = 94,44%

La differenza è quindi 12,92%.
La percentuale di pazienti sani nel *cluster* è 33,20% (85 pazienti sani su 256 individui).

Rete SOM 6x6
È stato selezionato il taglio a 2 (Figura 3.37 - 3.38), con la migliore variabilità intra-cluster della rete:

- variabilità intra-cluster = 33,886
- distanza inter-cluster = 1,102

e il taglio a 3 (Figura 3.39 - 3.40), con la migliore distanza inter-cluster della rete:

- variabilità intra-cluster = 34,625
- distanza inter-cluster = 1,222

La clusterizzazione del taglio a 2 indica che il *cluster* 1 contiene:

	cluster1	cluster2
sani	5	85
malati	53	171
tot	58	256

Figura 3.35: Confusion matrix relativa alla sezione "vita sessuale", rete 5x5, taglio a 2.

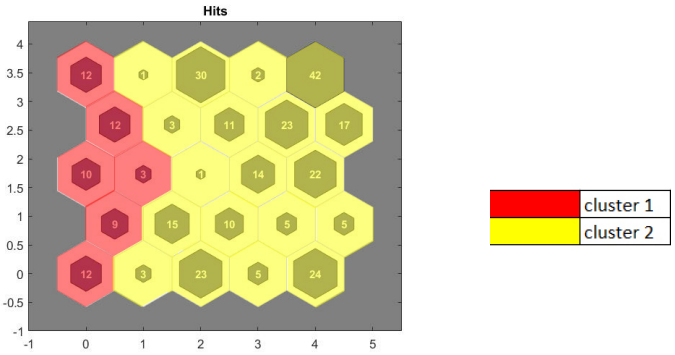


Figura 3.36: Mappa della rete SOM relativa alla sezione "vita sessuale", rete 5x5, taglio a 2, con evidenziati i neuroni relativi ai clusters.

	cluster1	cluster2
sani	63	27
malati	112	112
tot	175	139

Figura 3.37: Confusion matrix relativa alla sezione "vita sessuale", rete 6x6, taglio a 2.

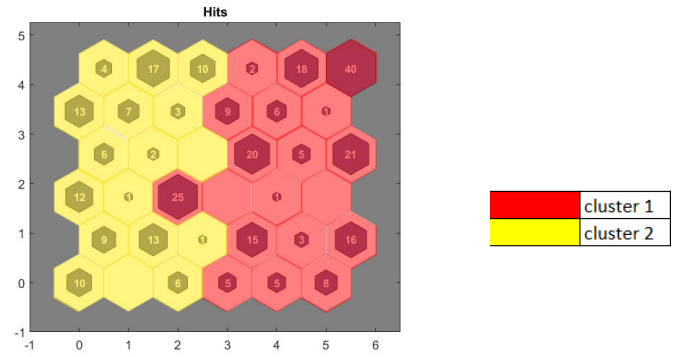


Figura 3.38: Mappa della rete SOM relativa alla sezione "vita sessuale", rete 6x6, taglio a 2, con evidenziati i neuroni relativi ai clusters.

- 175 individui su 314 individui totali = 55,73%
- 63 pazienti sani su 90 pazienti sani totali = 70,00%

La differenza è quindi 14,27%, la migliore della sezione.
La percentuale di pazienti sani nel *cluster* è 36,00% (63 su 175), il migliore della sezione "vita sessuale".

	cluster1	cluster2	cluster3
sani	13	14	63
malati	28	84	112
tot	41	98	175

Figura 3.39: Confusion matrix relativa alla sezione "vita sessuale", rete 6x6, taglio a 3.

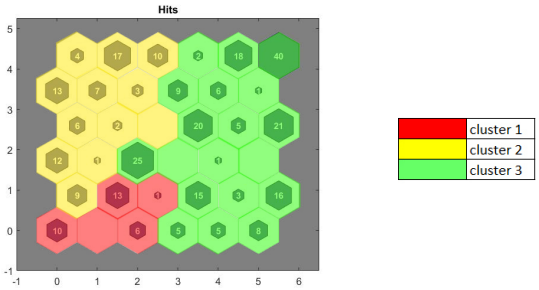


Figura 3.40: Mappa della rete SOM relativa alla sezione "vita sessuale", rete 6x6, taglio a 3, con evidenziati i neuroni relativi ai clusters.

Per quanto riguarda la clusterizzazione del taglio a 3, il *cluster* migliore è lo stesso del taglio a 2, perciò i valori risultano gli stessi.

Rete SOM 7x7

È stato selezionato solo il taglio a 2 (Figura 3.41 - 3.42), con i seguenti valori:

- variabilità intra-cluster = 33.327
- distanza inter-cluster = 1,342

La clusterizzazione indica che il *cluster* 2 contiene:

- 250 individui su 314 individui totali = 79,62%
- 84 pazienti sani su 90 pazienti sani totali = 93,33%

La differenza è quindi 13,72%.
La percentuale di pazienti sani nel *cluster* è 33,60% (84 pazienti sani su 250 individui).

Rete SOM 8x8

È stato selezionato solo il taglio a 2 (Figura 3.43 - 3.44), con i seguenti valori:

- variabilità intra-cluster = 33.681
- distanza inter-cluster = 1,208

	cluster1	cluster2
sani	6	84
malati	58	166
tot	64	250

Figura 3.41: Confusion matrix relativa alla sezione "vita sessuale", rete 7x7, taglio a 2.

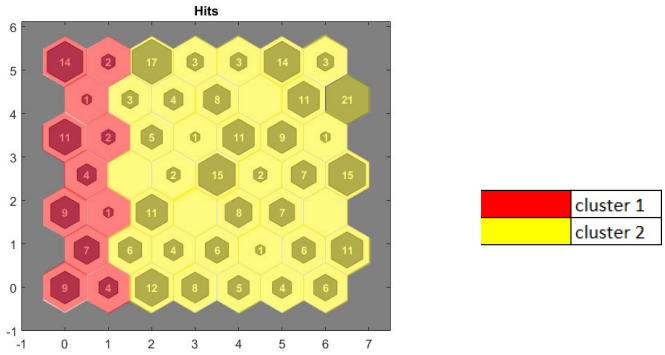


Figura 3.42: Mappa della rete SOM relativa alla sezione "vita sessuale", rete 7x7, taglio a 2, con evidenziati i neuroni relativi ai clusters.

	cluster1	cluster2
sani	19	71
malati	82	142
tot	101	213

Figura 3.43: Confusion matrix relativa alla sezione "vita sessuale", rete 8x8, taglio a 2.

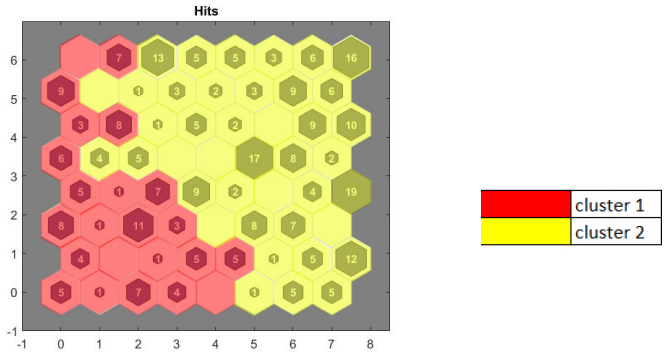


Figura 3.44: Mappa della rete SOM relativa alla sezione "vita sessuale", rete 8x8, taglio a 2, con evidenziati i neuroni relativi ai clusters.

La clusterizzazione indica che il *cluster* 2 contiene:

- 213 individui su 314 individui totali = 67,83%
- 71 pazienti sani su 90 pazienti sani totali = 78,89%

La differenza è quindi 11,05%.

La percentuale di pazienti sani nel *cluster* è 33,33% (71 pazienti sani su 213 individui).

Le prove di questa sezione hanno quindi indicato due reti diverse, per i migliori parametri e i migliori indicatori globali:

- 5x5 per la variabilità intra-cluster e la distanza inter-cluster;
- 6x6 per entrambi gli indicatori della clusterizzazione;

Entrambi gli indicatori della clusterizzazione hanno prodotto i risultati peggiori tra le sezioni.

Prove con le domande relative alla vita sessuale e ai sintomi

Rete SOM 4x4

È stato selezionato solo il taglio a 2 (Figura 3.45 - 3.46), con i seguenti valori:

- variabilità intra-cluster = 94,522
- distanza inter-cluster = 1,299

	cluster1	cluster2
sani	11	79
malati	91	133
tot	102	212

Figura 3.45: Confusion matrix relativa alla sezione "vita sessuale e sintomi", rete 4x4, taglio a 2.

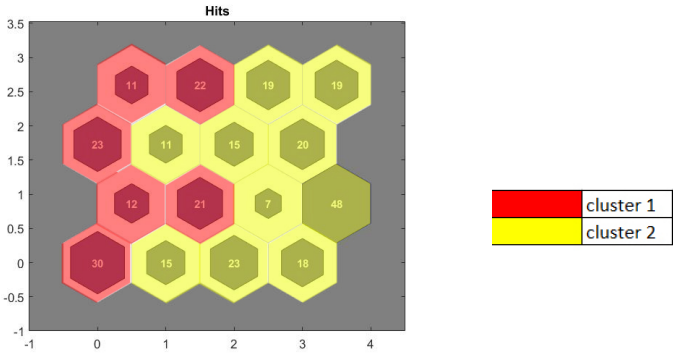


Figura 3.46: Mappa della rete SOM relativa alla sezione "vita sessuale e sintomi", rete 4x4, taglio a 2, con evidenziati i neuroni relativi ai clusters.

La clusterizzazione indica che il *cluster* 1 contiene:

- 212 individui su 314 individui totali = 67,52%
- 79 pazienti sani su 90 pazienti sani totali = 87,78%

La differenza è quindi 20,26%.
La percentuale di pazienti sani nel *cluster* è 37,26% (79 pazienti sani su 212 individui).

Rete SOM 5x5
È stato selezionato solo il taglio a 2 (Figura 3.47 - 3.48), con i seguenti valori:

- variabilità intra-cluster = 93,423
- distanza inter-cluster = 1,274

	cluster1	cluster2
sani	13	77
malati	84	140
tot	97	217

Figura 3.47: Confusion matrix relativa alla sezione "vita sessuale e sintomi", rete 5x5, taglio a 2.

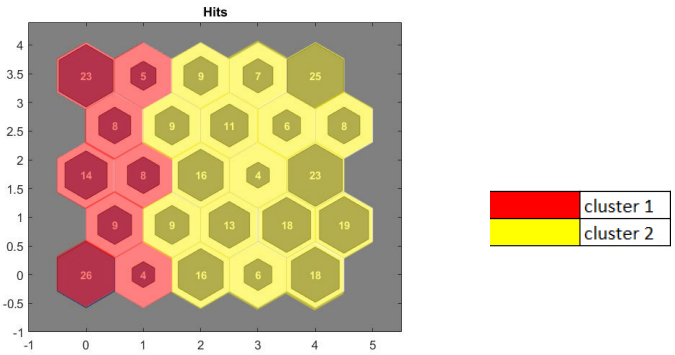


Figura 3.48: Mappa della rete SOM relativa alla sezione "vita sessuale e sintomi", rete 5x5, taglio a 2, con evidenziati i neuroni relativi ai clusters.

La clusterizzazione indica che il *cluster* 1 contiene:

- 217 individui su 314 individui totali = 69,11%
- 77 pazienti sani su 90 pazienti sani totali = 85,56%

La differenza è quindi 16,45%.
La percentuale di pazienti sani nel *cluster* è 35,48% (77 pazienti sani su 217 individui).

Rete SOM 6x6
È stato selezionato il taglio a 2 (Figura 3.49 - 3.50), con la migliore variabilità intra-cluster della rete:

- variabilità intra-cluster = 94,144
- distanza inter-cluster = 1,178

	cluster1	cluster2
sani	75	15
malati	108	116
tot	183	131

Figura 3.49: Confusion matrix relativa alla sezione "vita sessuale e sintomi", rete 6x6, taglio a 2.

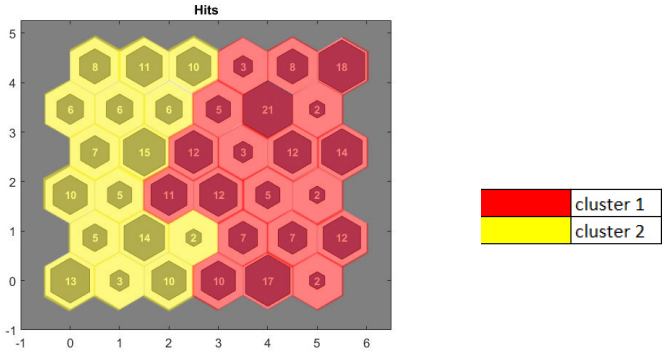


Figura 3.50: Mappa della rete SOM relativa alla sezione "vita sessuale e sintomi", rete 6x6, taglio a 2, con evidenziati i neuroni relativi ai clusters.

e il taglio a 4 (Figura 3.51 - 3.52), con la migliore distanza inter-cluster della rete:

- variabilità intra-cluster = 99,008
- distanza inter-cluster = 1,301

La clusterizzazione del taglio a 2 indica che il *cluster* 1 contiene:

- 175 individui su 314 individui totali = 58,28%
- 63 pazienti sani su 90 pazienti sani totali = 83,33%

La differenza è quindi 25,05%.
La percentuale di pazienti sani nel *cluster* è 36,00% (75 su 183).

	cluster1	cluster2	cluster3	cluster4
sani	1	9	5	75
malati	36	38	42	108
tot	37	47	47	183

Figura 3.51: Confusion matrix relativa alla sezione "vita sessuale e sintomi", rete 6x6, taglio a 4.

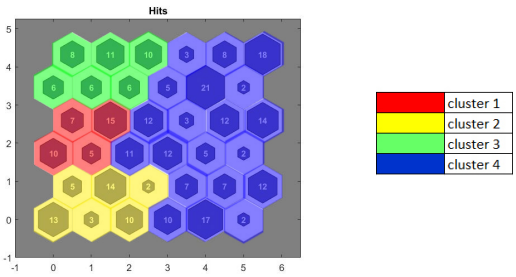


Figura 3.52: Mappa della rete SOM relativa alla sezione "vita sessuale e sintomi", rete 6x6, taglio a 4, con evidenziati i neuroni relativi ai clusters.

Per quanto riguarda la clusterizzazione del taglio a 4, il *cluster* migliore è lo stesso del taglio a 2, perciò i valori risultano gli stessi.

Rete SOM 7x7
È stato selezionato il taglio a 2 (Figura 3.53 - 3.54), con la migliore variabilità intra-cluster della rete:

- variabilità intra-cluster = 94,432
- distanza inter-cluster = 1,214

e il taglio a 3 (Figura 3.56 - 3.56), con la migliore distanza inter-cluster della rete:

- variabilità intra-cluster = 94,432
- distanza inter-cluster = 1,214

e il taglio a 6 (Figura 3.58 - 3.58), con la migliore distanza inter-cluster della rete:

- variabilità intra-cluster = 99,860
- distanza inter-cluster = 1,224

La clusterizzazione del taglio a 2 indica che il *cluster* 2 contiene:

- 225 individui su 314 individui totali = 71,66%
- 79 pazienti sani su 90 pazienti sani totali = 87,78%

	cluster1	cluster2
sani	11	79
malati	78	146
tot	89	225

Figura 3.53: Confusion matrix relativa alla sezione "vita sessuale e sintomi", rete 7x7, taglio a 2.

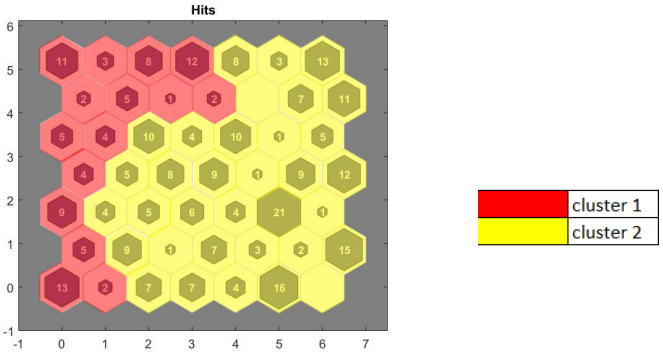


Figura 3.54: Mappa della rete SOM relativa alla sezione "vita sessuale e sintomi", rete 7x7, taglio a 2, con evidenziati i neuroni relativi ai clusters.

	cluster1	cluster2	cluster3
sani	53	26	11
malati	41	105	78
tot	94	131	89

Figura 3.55: Confusion matrix relativa alla sezione "vita sessuale e sintomi", rete 7x7, taglio a 3.

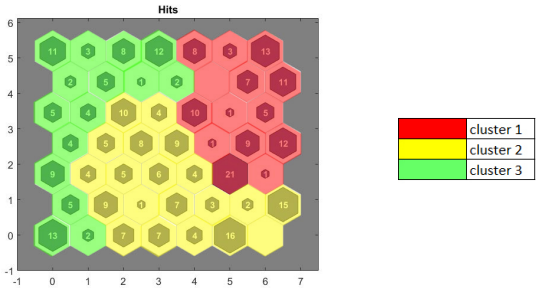


Figura 3.56: Mappa della rete SOM relativa alla sezione "vita sessuale e sintomi", rete 7x7, taglio a 3, con evidenziati i neuroni relativi ai clusters.

La differenza è quindi 16,12%.
La percentuale di pazienti sani nel *cluster* è 35,11% (79 su 225).

La clusterizzazione del taglio a 3 indica che il *cluster* 1 contiene:

- 94 individui su 314 individui totali = 29,94%
- 53 pazienti sani su 90 pazienti sani totali = 58,89%

	cluster1	cluster2	cluster3	cluster4	cluster5	cluster6
sani	10	43	24	2	4	7
malati	13	28	46	59	17	21
tot	23	71	70	61	21	28

Figura 3.57: Confusion matrix relativa alla sezione "vita sessuale e sintomi", rete 7x7, taglio a 6.

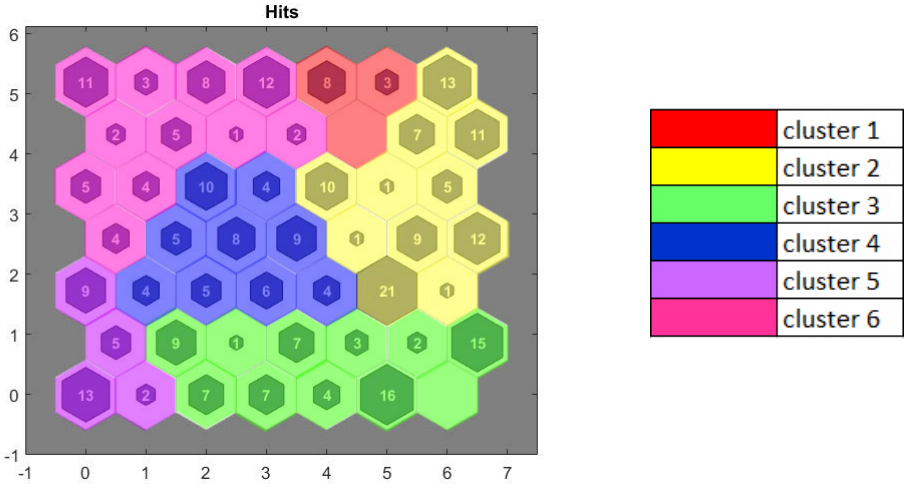


Figura 3.58: Mappa della rete SOM relativa alla sezione "vita sessuale e sintomi", rete 7x7, taglio a 6, con evidenziati i neuroni relativi ai clusters.

La differenza è quindi 28,95%, la migliore della rete e della sezione "vita sessuale e sintomi".
La percentuale di pazienti sani nel *cluster* è 56,38% (53 su 94).

La clusterizzazione del taglio a 6 indica che il *cluster* 2 contiene:

- 71 individui su 314 individui totali = 22,61%
- 43 pazienti sani su 90 pazienti sani totali = 47,78%

La differenza è quindi 25,17%.
La percentuale di pazienti sani nel *cluster* è 60,56% (43 su 71), la migliore della sezione "vita sessuale e sintomi".

Questo è un caso particolare rispetto agli altri, si è deciso di privilegiare il taglio a 6, che presenta una maggiore prevalenza di sani nel *cluster* 2, rispetto al taglio a 3 che presenta il 3,72% in più nella differenza. Con questa scelta entrambi gli indicatori sono comunque i migliori della sezione "vita sessuale e sintomi".

Rete SOM 8x8
È stato selezionato il taglio a 2 (Figura 3.49 - 3.50), con la migliore variabilità intra-cluster della rete:

- variabilità intra-cluster = 93,108
- distanza inter-cluster = 1,167

La variabilità intra-cluster risulta essere la migliore della sezione "vita sessuale e sintomi".

	cluster1	cluster2
sani	73	17
malati	130	94
tot	203	111

Figura 3.59: Confusion matrix relativa alla sezione "vita sessuale e sintomi", rete 8x8, taglio a 2.

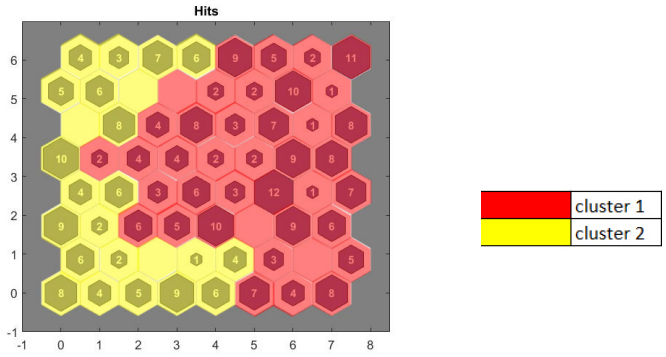


Figura 3.60: Mappa della rete SOM relativa alla sezione "vita sessuale e sintomi", rete 8x8, taglio a 2, con evidenziati i neuroni relativi ai clusters.

e il taglio a 4 (Figura 3.61 - 3.62), con la migliore distanza inter-cluster della rete:

- variabilità intra-cluster = 98,060
- distanza inter-cluster = 1,375

La distanza inter-cluster risulta essere la migliore della sezione “vita sessuale e sintomi”.

La clusterizzazione del taglio a 2 indica che il *cluster* 2 contiene:

- 203 individui su 314 individui totali = 64,65%
- 73 pazienti sani su 90 pazienti sani totali = 81,11%

La differenza è quindi 16,46%.
La percentuale di pazienti sani nel *cluster* è 35,96% (73 su 203).

Per quanto riguarda la clusterizzazione del taglio a 4, il *cluster* migliore è lo stesso del taglio a 2, perciò i valori risultano gli stessi.

Le prove di questa sezione hanno quindi indicato due reti diverse, per i migliori parametri e i migliori indicatori globali:

- 8x8 per la variabilità intra-cluster e la distanza inter-cluster;
- 7x7 per entrambi gli indicatori della clusterizzazione;

Entrambi gli indicatori della clusterizzazione hanno prodotto i risultati migliori tra le sezioni.

	cluster1	cluster2	cluster3	cluster4
sani	7	6	4	73
malati	17	32	45	130
tot	24	38	49	203

Figura 3.61: Confusion matrix relativa alla sezione “vita sessuale e sintomi”, rete 8x8, taglio a 4.

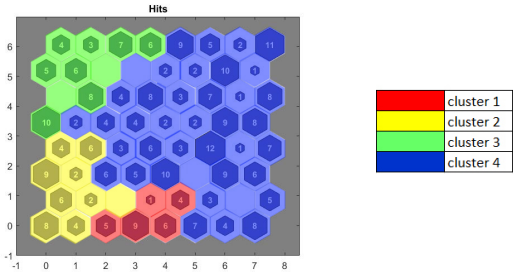


Figura 3.62: Mappa della rete SOM relativa alla sezione “vita sessuale e sintomi”, rete 8x8, taglio a 4, con evidenziati i neuroni relativi ai clusters.

3.2.3 Distribuzione delle patologie nei clusters

È stata analizzata la distribuzione nei *cluster* dei pazienti affetti dalle patologie considerate, per valutare una possibile successiva divisione rispetto ai soli pazienti sani e pazienti affetti da MST. I dati a disposizione delle singole MST sono:

- Clamidia (37 dati simulati);
- Condiloma ano-genitale (46 dati simulati);
- Gonorrea (25 dati simulati);
- Herpes genitale (61 dati simulati);
- Sifilide (38 dati simulati);
- patologie diverse denominate "Altro" (5 dati simulati).

Alcune categorie sono state escluse, in quanto sono state aggiunte dai pazienti e non risultano sufficientemente significative in numero per l'analisi in corso:

- HPV, i casi presenti in questo dataset sono soltanto due, non sufficienti a valutare la distribuzione;
- il paziente simulato affetto da piattole che, come per il procedimento di clusterizzazione, non è stato eliminato ma non è stato considerato;
- i pazienti che non hanno consultato il medico, anche questi mantenuti ma non valutabili.

Metodo e risultati

Per ognuno dei tagli delle reti riportati nel paragrafo precedente (3.2.2), sono state ricavate le tabelle che riportano il numero di pazienti affetti da ogni patologia in ognuno dei *cluster*.
Da queste è stata successivamente calcolata la percentuale dei pazienti per ogni *cluster* e confrontata con la percentuale del totale dei pazienti nel *cluster*, con lo scopo di valutare la distribuzione in relazione alla numerosità.
Le caselle delle patologie sono state quindi evidenziate come in Figura 3.63. In questo caso sono state selezionate la Clamidia e la Gonorrea.

- in giallo se la differenza tra patologia e totale del *cluster* migliore supera il 10%;
- in verde se la differenza tra patologia e totale del *cluster* migliore supera il 25%.

4x4							
	Clamidia	Sifilide	Condiloma	Herpes	Gonorrea	Altro	Totale
1	29	20	24	40	9	2	203
2	8	18	22	21	16	3	111

Figura 3.63: Esempio di tabella per le patologie (rete 4x4 della sezione "Intero questionario")

.

La Clamidia presenta la maggioranza dei pazienti all'interno del *cluster* 1, il quale contiene:

- 203 individui su 314 individui totali = 64,65%
- 29 pazienti affetti da Clamidia su 37 casi di Clamidia totali = 78,38%

La differenza è quindi 13,73%, per cui la Clamidia è stata evidenziata in giallo.

La Gonorrea presenta la maggioranza dei pazienti all'interno del *cluster* 2, il quale contiene:

- 111 individui su 314 individui totali = 35,35%
- 16 pazienti affetti da Gonorrea su 25 casi di Gonorrea totali = 64,00%

La differenza è quindi 28,65%, per cui la Gonorrea è stata evidenziata in verde.

L'Herpes genitale presenta la maggioranza dei pazienti all'interno del *cluster* 1, il quale contiene:

- 203 individui su 314 individui totali = 64,65%
- 40 pazienti affetti da Herpes genitale su 61 casi di Herpes genitale totali = 65,67%

La differenza è quindi 1,02%, per cui l'Herpes genitale non viene selezionato, in quanto la sua distribuzione è troppo simile a quella del totale degli individui.

Successivamente è stato ricavato un grafico che riportasse i *clusters* selezionati, in entrambi i casi in percentuale al totale disponibile per ogni sezione (questa scelta è dovuta alla necessità di uniformare i valori a causa del diverso numero di tagli selezionati nelle tre serie di prove analizzate). In questo modo è stato possibile confrontare la resa sia delle diverse patologie, sia delle diverse divisioni del questionario.

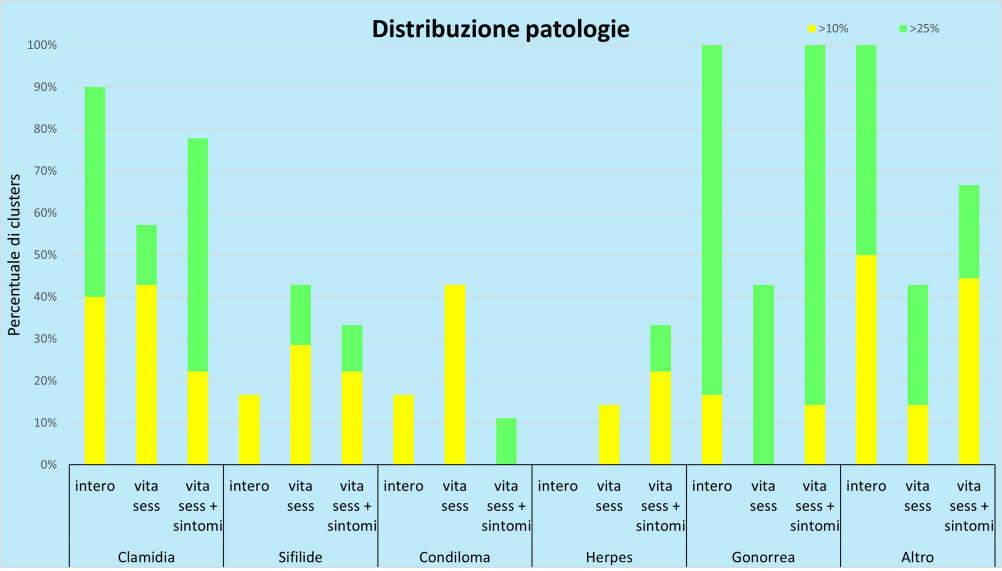


Figura 3.64: Grafico relativo alla qualità delle distribuzioni delle patologie relativamente alle tre diverse sezioni analizzate.

.

Il grafico (in Figura 3.64) mostra come la maggiore percentuale di *clusters* selezionati è data dalla Gonorrea, seguita rispettivamente da Clamidia, "Altro", Sifilide, Condiloma ano-genitale ed Herpes genitale. La Gonorrea risulta la patologia con la maggior quantità di clusters con differenze di percentuale maggiori del 25%, seguita dalle altre nello stesso ordine ad eccezione del condiloma, che possiede una grande prevalenza di differenze di percentuale maggiori del 10%.
Relativamente alle tre divisioni del questionario, è la parte con le domande relative alla vita sessuale e ai sintomi del paziente a dare risultati decisamente migliori, sia come selezione totale sia come qualità delle distribuzioni, seguita poi dall'intero questionario.

3.2.4 Conclusioni

Rispetto alle prove effettuate in precedenza (Paragrafo 3.1.4), l'aumento dei dati simulati da 139 a 314 ha prodotto un miglioramento nella maggior parte degli aspetti di ogni sezione del questionario analizzata, con una clusterizzazione più efficace per le prove relative alle domande sulla vita sessuale e i sintomi del paziente. Tale miglioramento probabilmente è dovuto alla buona coerenza tra le risposte simulate dei medici e le risposte reali dei pazienti, in particolare alle domande

relativa ai sintomi.
Si riportano i dettagli degli aspetti migliorati:

- la variabilità intra-cluster si è ridotta in media del 5,97%, risultato positivo in quanto all'interno di un *cluster* gli elementi devono essere il più simili possibile;
- la distanza inter-cluster è aumentata del 1,04% (la percentuale minore è data da valori complessivamente più simili rispetto alla variabilità intra-cluster), indicando che gli elementi che costituiscono diversi *clusters* sono più separati/distanti tra loro, ovvero elementi più dissimili nei diversi *clusters*;
- entrambi gli indicatori della clusterizzazione sono migliorati rispettivamente del 1,64% e del 13,57%, suggerendo che la separazione dei pazienti sani rispetto alla separazione in *clusters* del totale dei pazienti è sostanzialmente simile, ma vengono creati dei singoli *clusters* che possiedono un maggior numero di sani rispetto ai malati di quanto avveniva in precedenza.

In merito alla distinzione delle singole patologie, i risultati ottenuti sono migliori di quelli attesi, data l'esiguità dei dati: probabilmente la decisione di mantenere le patologie separate ha permesso di avere un dataset più bilanciato (sono presenti 90 pazienti sani e la patologia più numerosa è l'Herpes genitale con 61 pazienti). La Gonorrea (con 25 pazienti) possiede le clusterizzazioni migliori, risultato atteso date le risposte particolarmente coerenti con le informazioni trovate in fase di ricerca bibliografica, in particolare si segnalano la varietà e la particolarità dei sintomi come descritto al Paragrafo 3.2.3 (Domanda 30).
Dalla serie di prove riportate e discusse, sono stati ottenuti risultati di partenza soddisfacenti, in relazione al numero di dati disponibili, e la conferma che la ricezione di un maggior numero di dati possa favorire sensibilmente i risultati ottenuti con questa analisi.

3.3 Feature Selection

In questa sezione viene discussa la scelta dei metodi e i risultati ottenuti con il processo di *Feature Selection*, per la quale si è deciso di utilizzare un Algoritmo Genetico (AG). Tale passaggio risulta fondamentale per la selezione delle domande più importanti al fine della classificazione dei pazienti, scopo ultimo del progetto, e quindi per consentire l'eliminazione delle domande che non portano informazione e/o risultano ridondanti per l'algoritmo. La scelta dell'AG deriva da una serie di osservazioni:

- la molteplicità dei parametri disponibili rende l'algoritmo adatto a una maggiore varietà di prove e a una maggiore adattabilità, considerata l'esiguità dei dati disponibili al momento attuale;

- il metodo conserva le features originali, quindi domande e risposte, aspetto fondamentale se si lavora su un questionario.

3.3.1 Metodi

Si è quindi deciso di procedere con la ricerca bibliografica, volta in particolare all'identificazione di una funzione di fitness adeguata ai dati disponibili, dalla quale vengono considerati diversi articoli. Un primo esempio è l'AG proposto da Zhang et al. (*Variable selection in Logistic regression model with Genetic Algorithm* [18]), nel quale vengono proposti vari metodi per il calcolo della funzione di *fitness*, i quali però presentano l'utilizzo della media (la quale, in quanto caratteristica globale, può non essere sufficientemente informativa), oppure l'aggregazione di due indicatori e/o la creazione di uno "score" che risulta particolarmente utile solo se paragonato a quello di un diverso dataset. Si è quindi considerato un secondo metodo che potrebbe rivelarsi utile per la fase successiva al presente lavoro di tesi, descritto da Benteng Ma e Wong Xia (*A Tribe Competition-Based Genetic Algorithm for Feature Selection in Pattern Classification* [19]). Esso lavora su due diversi livelli: la popolazione viene divisa in tribù, le quali si evolveranno secondo i meccanismi di un classico AG, per poi successivamente scontrarsi tra loro, risultando nell'aumento della dimensione della tribù vincente e nella riduzione di tutte le altre. Questo aspetto potrebbe infatti essere utile nella fase di lavoro relativa all'analisi del rischio dei pazienti di aver contratto una MST. Infine, si è scelto il metodo proposto da Babatunde Oluleye et al., (*A Genetic Algorithm-Based Feature Selection* [20]), in quanto è stato creato per dati di tipo *categorical*, lo stesso formato in cui vengono salvati i dati provenienti dalle risposte al questionario dopo l'esportazione e, si ritiene, una funzione di *fitness* più adatta perché basata interamente sulla performance del classificatore KNN implementato e sul numero di *features* selezionate. Si riportano quindi gli aspetti fondamentali di tale metodo:

- *Generazione della popolazione iniziale*: la popolazione iniziale è stata generata in maniera randomica, inoltre, si è deciso di seguire l'indicazione che la popolazione debba essere almeno di lunghezza pari alla lunghezza del cromosoma, cioè il totale delle *features* (66), perciò si è deciso di iniziare con un numero di individui pari ai dati utilizzabili per questo metodo, ovvero 305 individui iniziali.

- *Funzione di fitness*: è basata su classificatore KNN, il quale classifica ogni elemento cercando gli elementi con una distanza minore dall'elemento in oggetto. La distanza viene calcolata secondo l'Eq. 3.1: in questo caso il numero di elementi "vicini" selezionati è 3, perciò questi verranno utilizzati per stabilire la classe dell'elemento in oggetto.

$$D\left(x_{test}, x_i\right) = \sqrt{\sum_{m=1}^{M} \left(x_{test} - x_i\right)^2} \tag{3.1}$$

Il valore di interesse in questo caso è l'errore di classificazione: esso viene calcolato per ogni feature di ogni individuo e viene selezionato il valore minore per ogni individuo. Successivamente viene scelto l'individuo con l'errore di classificazione minore per il calcolo del valore "finale" della *fitness* (Eq. 3.2), dove:

- α = errore di classificazione del KNN;
- N_f = numero di features selezionate.

$$fit = \frac{\alpha}{N_f} + e^{\left(\frac{-1}{N_f}\right)} \tag{3.2}$$

Pertanto, il valore migliore della funzione di *fitness* sarà il valore minore, in quanto basato sull'errore di classificazione.

- *Selezione della nuova popolazione*, basata sugli individui ottenuti secondo tre diversi procedimenti:
  - *Selezione degli "Élite Children"*: i due individui con il minore (e quindi migliore) valore della funzione di fitness passano alla generazione successiva senza subire modifiche;
  - *Crossover*: tecnica che riprende l'equivalente meccanismo genetico della riproduzione umana, ovvero unendo nel figlio il patrimonio genetico derivante dai due genitori all'interno di un singolo cromosoma; in questo caso, esso è ottenuto mediante l'utilizzo di una probabilità di crossover che si è deciso, a differenza dell'articolo, di non mantenere fissa, ma variare tra 0,8 e 0,95 per la definizione della migliore scelta del parametro;
  - *Mutazione*: anch'essa rappresenta una tecnica che riprende l'equivalente meccanismo genetico della riproduzione umana, ovvero consiste nella perturbazione di un singolo cromosoma indipendentemente dai genitori; è ottenuta mediante l'utilizzo di una probabilità di mutazione (in genere minore della probabilità di crossover) che è vincolata, in quanto è ottenuta dal numero di individui rimanenti dopo la selezione degli *"Élite Children"* (da cui restano 312 individui disponibili) e i nuovi individui ottenuti mediante crossover (ad esempio nc=305*0,8=244).
- *Condizioni di stop*: si definiscono le condizioni che comportano lo stop dell'iterazione dell'AG, ovvero il raggiungimento del numero massimo di iterazioni

scelto (5000) o una variazione della funzione dei fitness minore o uguale a 0,000001.

Implementazione e prove

Si è deciso di implementare l'Algoritmo Genetico mediante:

- l'utilizzo sia della distanza descritta in precedenza (Eq. 3.1), sia della classica distanza Euclidea;
- la codifica delle classi sia con la codifica originale (mantenendo le patologie separate), sia come paziente sano o paziente affetto da MST (con due sole classi). In questa fase si è deciso di eliminare i 9 pazienti con classe *"Non ho consultato il medico"*, in quanto nel secondo caso precedentemente descritto non potevano essere inserite né nella classe dei pazienti sani né nella classe dei pazienti affetti da MST.

Si ottengono così quattro diverse forme, per ognuna delle quali sono state svolte una serie di prove, variando la probabilità di crossover da 0,8 a 0,95, per un totale di 16 prove.

3.3.2 Analisi dei risultati dell'Algoritmo Genetico

Per ognuno dei quattro metodi utilizzati (ottenuti variando la formula relativa alla valutazione della distanza del KNN e la codifica delle classi relative ai pazienti affetti da MST come descritto nel Paragrafo 3.3.1) sono state effettuate 16 prove variando la probabilità di crossover da 0,8 a 0,95 a passi di 0,01. Per ognuna di queste prove sono state effettuate 10 ripetizioni, ottenendo quindi 10 soluzioni con relativo valore della funzione di fitness per ogni valore della probabilità di crossover analizzato.
Sono stati quindi valutati cinque diversi criteri:

- *numero di soluzioni uguali*: se una soluzione si ripete più di una volta, partendo da una generazione casuale della popolazione, questa viene considerata come preponderante;
- *numero di valori uguali della funzione di fitness*: in assenza di soluzioni uguali, si considerano i valori di funzione di *fitness* uguali come indicativi della bontà della prova;
- *numero di bit di differenza tra le soluzioni*: questo criterio è stato implementato al fine di verificare se ci fossero soluzioni simili in caso non si presentassero soluzioni uguali e/o valori della funzione di *fitness* uguali, a causa del grande

numero di *features* presenti (66). Tuttavia, probabilmente proprio a causa della presenza di 66 *features* diverse, in tutte le prove si sono trovate soluzioni con un minimo di 12 bit diversi fino a un massimo di 38: per questo motivo, questo criterio non è stato più considerato e non verrà riportato nei risultati delle prove (Paragrafo 3.3.3 e 3.3.4);

- *errore di classificazione*: espresso in percentuale, per valutare anche la performance del classificatore KNN implementato al fine della valutazione della funzione di *fitness*; viene riportato l'errore di classificazione tra pazienti sani e pazienti affetti da MST: esso viene calcolato dalla *confusion matrix* relativa alla soluzione salvata per ogni ripetizione, perciò, nel caso delle prove in cui i pazienti affetti dalle varie patologie sono stati mantenuti separati, questo valore è stato valutato sulla *confusion matrix* aggregata;
- *percentuale di corretta classificazione delle patologie*: nelle due prove effettuate con patologie separate, si è deciso di valutare anche quanto il classificatore KNN abbia classificato correttamente le singole patologie, per eventualmente identificarne una o alcune che peggiorino il risultato globale dato dall'errore di classificazione.

Sono stati quindi analizzati tre diversi criteri nel caso delle prove con le patologie unite, quattro nel caso delle prove a patologie separate.
È stato anche verificato il numero di *features* selezionate: le soluzioni considerate adatte selezionano da 20 a 40 *features* (per non ridurre eccessivamente il questionario pur ottenendo una selezione significativa): la maggior parte delle soluzioni presenta 20 *features* selezionate, solo alcune ne presentano 21 o 22.

Spazio delle soluzioni

Lo spazio delle soluzioni indica l'estensione di tutte le possibili soluzioni sulla base del problema in oggetto, ovvero il numero massimo di soluzioni ottenibili sulla base del numero di features disponibili per la selezione. La percentuale di spazio delle soluzioni esplorata è perciò un buon indicatore dell'efficacia e dell'esaustività dell'analisi condotta dall'Algoritmo Genetico.
Dato il numero esiguo di dati, si è verificato quanto spazio delle soluzioni venisse effettivamente esplorato, considerando come soddisfacente una percentuale del 10%.
Lo spazio delle soluzioni è definito secondo l'Eq. 3.3, mentre il numero di soluzioni generate dall'algoritmo genetico è dato dall'Eq. 3.4.

$$\text{spazio delle soluzioni} = 2^n - 1 \tag{3.3}$$

con n = numero di features

$$\text{numero individui} * \text{numero iterazioni} \geq 10\% \tag{3.4}$$

La dimensione dello spazio delle soluzioni permette di scegliere entrambi i parametri, quindi il numero di individui e il numero di iterazioni. Normalmente viene aumentato il secondo, in quanto aumentare il primo comporta un importante incremento del tempo computazionale.
Il problema in oggetto è costituito da 66 *features*, pertanto lo spazio delle soluzioni risulta molto esteso:

$$\text{Spazio delle soluzioni: } 2^{66} - 1 = 7{,}3787 * 10^{19}$$

In genere, si considera soddisfacente un'esplorazione del 10% dello spazio delle soluzioni:

$$\text{Esplorazione al 10\%: } (7{,}3787 * 10^{19}) * 0{,}1 = 7{,}3787 * 10^{18}$$

Considerata l'estensione dello spazio delle soluzioni, sia la dimensione della popolazione iniziale e il numero di iterazioni devono essere molto alti per raggiungere un adeguato ordine di grandezza. Aumentare il numero di iterazioni risulta non sufficiente, in quanto la seconda condizione di stop verrà verificata sempre prima del raggiungimento di un numero di iterazioni elevato; aumentare la popolazione iniziale in modo da verificare la condizione sull'esplorazione, invece, aumenterebbe eccessivamente il tempo computazionale e, considerati i pochi dati disponibili, una popolazione numerosa sarebbe poco significativa. Sono stati quindi impostati i seguenti parametri:

$$\text{n° soluzioni generate: } 305 \text{ individui} * 5000 \text{ iterazioni} = 1{,}525 * 10^{6}$$

Come descritto nel Paragrafo 3.3.2, la maggior parte delle soluzioni hanno selezionato 20 *features*, cioè il valore minimo impostato, riducendo lo spazio delle soluzioni finali di 13 ordini di grandezza:

$$\text{Spazio delle soluzioni: } 2^{20} - 1 = 1{,}0486 * 10^{6}$$

3.3.3 Risultati delle prove con due classi

Vengono qui riportati i risultati dei due metodi che utilizzano entrambi la codifica dei dati con le patologie unite (0 per i pazienti sani e 1 per i pazienti affetti da MST).

- *Metodo 1*: con patologie unite e la distanza definita dell'articolo scelto [20] e descritta dall'Eq. 3.1;

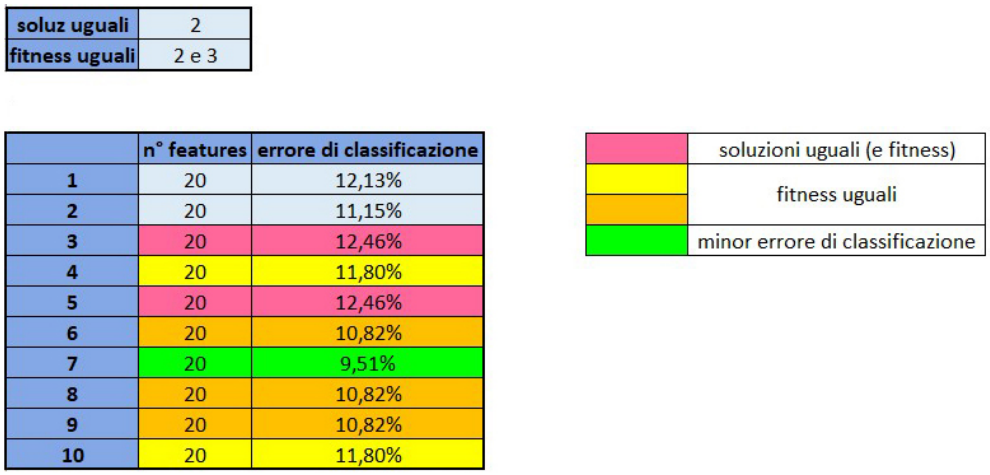
- *Metodo 2*: con patologie unite e l'utilizzo della distanza Euclidea.

Relativamente a questi due metodi, i criteri di confronto delle soluzioni per ogni probabilità di crossover sono stati riassunti come mostrato in Figura 3.65 : nella tabella in alto, viene riportato il numero di soluzioni uguali (se presenti) e il numero/i relativo alle soluzioni con valori uguali della sola funzione di *fitness* (se presenti); nella tabella in basso, vengono indicate, per ognuna delle 10 singole soluzioni (relative alle 10 ripetizioni dell'esecuzione dell'Algoritmo Genetico per ogni probabilità di crossover), il numero di *features* e l'errore di classificazione del KNN, espresso come percentuale. Ognuna delle soluzioni è stata evidenziata secondo la legenda riportata in figura, nella quale vengono indicati i tre criteri presenti per questi due metodi:

- *soluzioni uguali*: soluzioni che selezionano le stesse *features* (in rosa in figura);

- *fitness uguali*: le soluzioni possiedono lo stesso valore della funzione di *fitness*, ma selezionano *features* diverse (in giallo in figura);

- *minor errore di classificazione*: la soluzione evidenziata presenta il minor errore di classificazione relativo al classificatore KNN utilizzato per il calcolo della funzione di *fitness* (in verde in figura).



soluz uguali	2
fitness uguali	2 e 3

	n° features	**errore di classificazione**
1	20	12,13%
2	20	11,15%
3	20	12,46%
4	20	11,80%
5	20	12,46%
6	20	10,82%
7	20	9,51%
8	20	10,82%
9	20	10,82%
10	20	11,80%

Figura 3.65: Esempio di rappresentazione dei criteri di confronto per le soluzioni dell'Algoritmo Genetico relativo al metodo con distanza euclidea e patologie unite, con pc=0,89 (pc=probabilità di crossover).

.

A partire da queste informazioni, si è eseguita una valutazione globale del metodo, relativa sia alla performance generale, sia ai risultati migliori ottenuti.

Risultati del Metodo 1

Questo metodo, che utilizza l'Eq. 3.1 per l'implementazione della distanza nel classificatore KNN (come descritto nel Paragrafo 3.3.1), non ha prodotto risultati significativi: si ottiene sempre la medesima soluzione per ognuna delle probabilità di crossover testate. Verificando la *performance* del classificatore KNN, si è notato che tutti i pazienti vengono classificati come pazienti affetti da MST: questo è probabilmente dovuto in parte allo sbilanciamento del dataset (90 pazienti sani con 215 pazienti affetti da MST), il quale, insieme alla distanza utilizzata, porta sempre alla prevalenza di pazienti affetti da MST nei "vicini" selezionati dal classificatore KNN. Questa valutazione è stata poi confermata dai risultati ottenuti con gli altri metodi testati.

Risultati del Metodo 2

Questo metodo utilizza la distanza euclidea e la codifica con patologie unite. La Figura 3.66 riporta le singole soluzioni con almeno due criteri soddisfatti (in questo caso solo valori della funzione di fitness uguali e l'errore di classificazione minore) oppure le soluzioni che si sono ripetute per tutte e 10 le ripetizioni dell'Algoritmo Genetico. Per ognuna delle soluzioni riportate viene indicato l'errore di classificazione e si evidenzia il valore minore.

pc	n° soluzione	soluzioni uguali	fitness uguali	errore minore
0,82	TUTTE	10		9,84%
0,85	2-4 e 5		3	**9,51%**
0,87	TUTTE	10		10,49%
0,91	3-6 e 10		3	9,84%
0,93	4 e 10		2	10,49%
0,95	TUTTE	10		**9,51%**

Figura 3.66: Riassunto delle probabilità di crossover (pc) che presentano soluzioni che soddisfano almeno due criteri (evidenziati in tabella) e le probabilità di crossover che presentano tutte le soluzioni uguali.

.

Da questa tabella (Figura 3.66) è facile dedurre che la probabilità di crossover che ha prodotto la migliore soluzione è quella uguale 0,95: infatti essa risulta "stabile",

in quanto è stata ottenuta per tutte le 10 ripetizioni, inoltre è una delle due che ha prodotto l'errore di classificazione minore.
In Figura 3.67 viene mostrata la composizione delle domande selezionate secondo le tre sezioni in cui è stato diviso il questionario: vita personale, vita sessuale, sintomi. È possibile notare una prevalenza delle domande di vita personale, ma la distribuzione in realtà rispecchia quasi perfettamente la divisione delle *features* nelle tre sezioni: 31 *features* per la vita personale (46,97%), 17 *features* per la vita sessuale (25,76%), 18 *features* per i sintomi (27,27%).

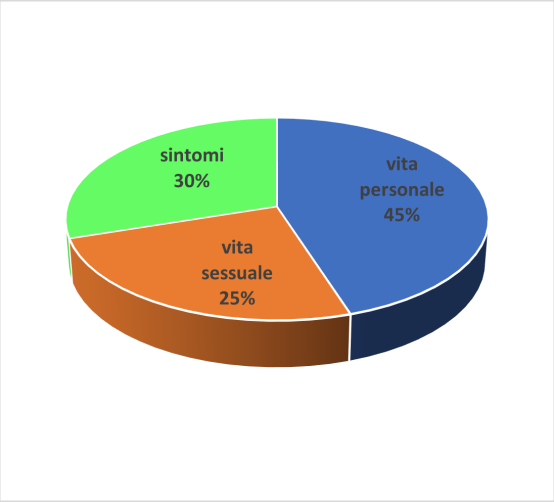


Figura 3.67: Grafico che mostra la composizione della soluzione con pc=0,95 per il metodo con due classi e utilizzo della distanza euclidea.

.

Sono state selezionate 20 *features*, cioè il minimo impostato, come nella maggior parte dei casi in ognuno dei metodi. Si segnala inoltre che, in questo caso, sono state selezionate molte *features* relative alla domanda sulle MST avute in passato dal paziente (ben 5 patologie sulle 9 disponibili) e anche 4 domande (su 6 disponibili) relative alle eruzioni cutanee, nella sezione sintomi.
Globalmente, questa serie di prove ha prodotto dei risultati accettabili: innanzitutto la grande presenza delle soluzioni uguali, in 14 casi su 16 (87,50%), 3 dei quali sono quelle riportate (18,75%). In ognuna delle probabilità di crossover analizzate sono presenti almeno una coppia di soluzioni con fitness uguali e l'errore di classificazione non è soddisfacente ma, considerato che i pazienti sani costituiscono il 29,51% dei pazienti totali, supera le aspettative. I valori dell'errore risultano inoltre piuttosto stabili, con una variazione dal 9,51% al 12,79%.

3.3.4 Risultati delle prove con patologie distinte

Vengono qui riportati i risultati dei due metodi che utilizzano, entrambi, la codifica dei dati con le patologie distinte, per un totale quindi di 8 classi (ovvero pazienti sani, Clamidia, Gonorrea, Condiloma ano-genitale, Herpes genitale, Sifilide, "Altro", HPV).

- *Metodo 3*: con patologie distinte e la distanza descritta dall'Eq. 3.1;
- *Metodo 4*: con patologie distinte e l'utilizzo della distanza Euclidea;

Relativamente a questi due metodi, i criteri di confronto delle soluzioni per ogni probabilità di crossover sono stati riassunti come mostrato in Figura 3.68: nella tabella in alto, viene riportato il numero di soluzioni uguali (se presenti) e il numero/i relativo alle soluzioni con valori uguali della sola funzione di *fitness* (se presenti); nella tabella in basso, vengono indicate, per ognuna delle 10 singole soluzioni (relative alle 10 ripetizioni dell'esecuzione dell'Algoritmo Genetico per ogni probabilità di crossover), il numero di *features* selezionate, l'errore di classificazione del KNN (espresso come percentuale), e la percentuale di corretta classificazione delle patologie. Per questi due metodi l'errore di classificazione viene valutato sulla *confusion matrix* a patologie aggregate, in modo da non perdere l'informazione relativa alla classificazione dei pazienti sani e dei pazienti affetti da MST. Ognuna delle soluzioni è stata evidenziata secondo la legenda riportata in figura, nella quale vengono indicati i quattro criteri presenti per questi due metodi:

- *soluzioni uguali (e fitness)*: che selezionano le stesse *features* e, di conseguenza, possiedono anche lo stesso valore della funzione di *fitness*;
- *fitness uguali*: le soluzioni possiedono lo stesso valore della funzione di *fitness*, ma selezionano *features* diverse;
- *minor errore di classificazione*: la soluzione evidenziata presenta il minor errore di classificazione relativo al classificatore KNN utilizzato per il calcolo della funzione di *fitness*;
- *maggiori patologie classificate*: maggiore percentuale di corretta classificazione delle patologie;

A partire da queste informazioni, si è eseguita una valutazione globale del metodo, relativa sia alla *performance* generale, sia ai risultati migliori ottenuti.

Risultati del Metodo 3

In Figura 3.69 vengono mostrate le soluzioni che soddisfano almeno due tra i criteri disponibili, riportati evidenziati in figura, poiché in nessun caso sono state

soluz uguali	6
fitness uguali	6 e 2

	n° features	errore di classificazione	patologie corrette
1	20	7,21%	84,19%
2	20	7,21%	84,19%
3	20	6,23%	87,44%
4	20	5,57%	83,72%
5	20	7,21%	84,19%
6	20	6,89%	83,26%
7	20	7,21%	84,19%
8	20	6,89%	81,40%
9	20	7,21%	84,19%
10	20	7,21%	84,19%

	soluzioni uguali (e fitness)
	fitness uguali
	minor errore di classificazione
	maggiori patologie classificate

Figura 3.68: Esempio di rappresentazione dei criteri di confronto per le soluzioni dell'Algoritmo Genetico relativo al metodo con distanza euclidea e patologie distinte, con pc=0,88 (pc=probabilità di crossover).

.

ottenute 10 ripetizioni con soluzioni uguali. Le soluzioni che possono essere ritenute migliori sono quelle con probabilità di crossover pari a 0,82 e 0,85, i quali presentano tre criteri soddisfatti e, rispettivamente, presentano anche la migliore classificazione delle patologie e il minor errore di classificazione tra pazienti sani e pazienti affetti da MST.

pc	n° soluzione	soluzioni uguali	fitness uguali	errore minore	patologie maggiore
0,81	1-2			23,94%	22,33%
0,82	4-5-7-8-10			29,18%	**26,51%**
0,83	5			25,25%	25,58%
0,85	1-2-3-4-5-6-7-9-10			**23,93%**	17,67%
0,86	1-2-4-5-7-10			27,21%	25,12%
0,92	10			22,30%	25,58%
0,93	2			29,18%	26,51%
0,94	1			26,56%	25,12%
0,95	8			25,90%	25,58%

Figura 3.69: Riassunto delle probabilità di crossover (pc) che presentano soluzioni che soddisfano almeno due criteri per il metodo 3 (criteri soddisfatti evidenziati in tabella).

.

Si riporta quindi la composizione delle domande selezionate dalle due soluzioni, secondo le tre sezioni in cui è stato diviso il questionario (vita personale, vita sessuale, sintomi):

- *pc=0,82*: in Figura 3.70 è possibile notare una prevalenza delle domande

di vita personale (55%), che risulta essere prevalente anche nel questionario. Rispetto alla soluzione mostrata in Figura 3.67, la differenza tra la sezione "vita sessuale" e la sezione "sintomi" rimane uguale, si riducono infatti allo stesso modo.

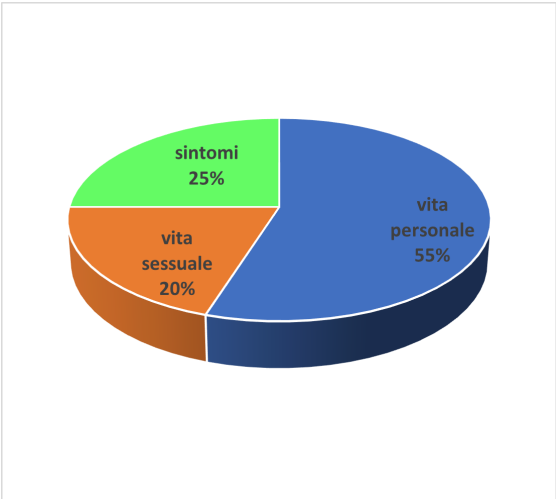


Figura 3.70: Grafico che mostra la composizione della soluzione con pc=0,82 con patologie separate.

.

Sono state selezionate 20 *features* finali, cioè il minimo numero impostato (come descritto nel Paragrafo 3.3.2), come nella maggior parte dei casi in ognuno dei metodi. Si segnala inoltre che, in questo caso, sono state selezionate molte *features* relative alla domanda relativa ai sintomi (nel questionario la domanda riporta una tabella con elencati i sintomi da selezionare): 4 *features* sulle 5 selezionate provengono proprio dalla sezione "sintomi".

- *pc=0,85*: questa soluzione (Figura 3.71) presenta una distribuzione molto più uniforme relativamente alle tre sezioni: le domande di vita personale sono molto ridotte rispetto ai casi precedenti. Sono presenti 21 *features*, uno dei pochi casi in cui sono state ottenute soluzioni con più di 20 *features* (il valore minimo impostato) e, non presenta una prevalenza di *features* relative a una particolare domanda e/o tipo di domande.

Globalmente, questa serie di prove ha prodotto dei risultati accettabili: innanzitutto per la grande presenza delle soluzioni uguali, in 12 casi su 16 (75%) al massimo 9 su 10, ma mai tutte uguali. In ognuna delle probabilità di crossover analizzate sono presenti almeno una coppia di soluzioni con *fitness* uguali. L'errore di classificazione peggiora rispetto al metodo con due classi (Paragrafo 3.3.3), probabilmente a causa

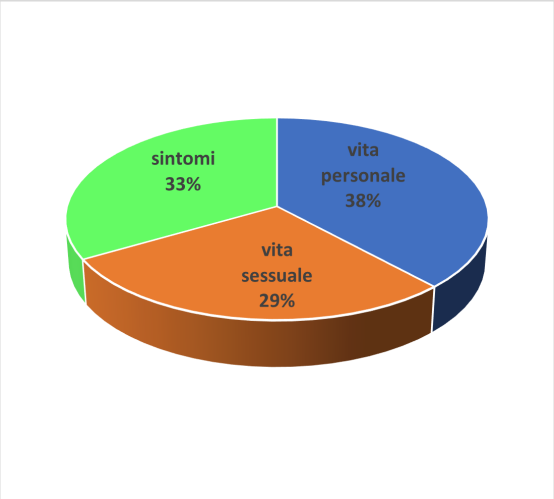


Figura 3.71: Grafico che mostra la composizione della soluzione con pc=0,85 con patologie separate.

.

dell'utilizzo di più di due classi, e la sua variazione è più consistente. Infatti, nel metodo a due classi si ha un errore di classificazione nel range 20,33% - 30,16%, che si riduce a 13,95% e 28,37% considerando la classificazione delle diverse patologie, variazione considerata comunque soddisfacente per il dataset a disposizione. Questa valutazione peggiora quando si verifica quali patologie siano state effettivamente classificate correttamente e con quale frequenza:

- il 72,19% del totale dei pazienti affetti da Herpes genitale, che compare nel 100% delle soluzioni disponibili;
- il 1,14% del totale dei pazienti affetti da Gonorrea, che compare nel 27,62% delle soluzioni disponibili;
- il 12,1% del totale dei pazienti affetti da Condiloma ano-genitale, che compare nel 22,86% delle soluzioni disponibili.

La corretta individuazione di solo 3 patologie principali su 5 (in totale 7 classi relative alle patologie), non può essere considerata utile, in quanto l'unica classificazione valida è quella dell'Herpes genitale.

Risultati del Metodo 4

In Figura 3.72 vengono mostrate le soluzioni ottenute mediante l'utilizzo della distanza euclidea, che soddisfano almeno due dei criteri disponibili, di cui alcune possiedono tutte le 10 soluzioni uguali, per una singola probabilità di crossover.

Le soluzioni che possono essere ritenute migliori sono quelle con probabilità di crossover pari a 0,84 e 0,93, infatti sono le uniche con 10 ripetizioni uguali e presentano, rispettivamente, il minor errore di classificazione tra pazienti sani e pazienti affetti da MST, e la migliore percentuale di corretta classificazione delle patologie. Si riporta quindi la composizione delle domande selezionate dalle due

pc	n° soluzione	soluzioni uguali	fitness uguali	errore minore	patologie maggiore
0,8	3-5-6-7-9			7,54%	87,91%
0,81	1			6,56%	86,05%
0,82	2			5,90%	81,86%
0,83	2 e 7			5,57%	86,05%
0,84	tutte	10		**5,25%**	84,19%
0,85	10			5,90%	85,12%
0,86	2			5,57%	86,05%
0,87	10			7,21%	86,51%
0,89	7			5,90%	85,58%
0,9	da 2 a 10			6,23%	85,12%
0,91	5 e 10			6,23%	85,58%
0,92	4 e 8			6,23%	84,65%
	3			6,89%	85,12%
0,93	tutte	10		5,57%	**95,40%**
0,94	10			5,90%	85,12%
0,95	10			5,90%	86,98%

Figura 3.72: Riassunto delle probabilità di crossover (pc) che presentano soluzioni che soddisfano almeno due criteri per il Metodo 4.

.

soluzioni, secondo le tre sezioni in cui è stato diviso il questionario (vita personale, vita sessuale, sintomi):

- *pc=0,84*: la distribuzione è nuovamente più equilibrata, come è possibile osservare in Figura 3.73: sono state selezionate 20 *features* e, tra le *features* selezionate, si nota la presenza di 2 delle 3 *features* dedicate alle vaccinazioni nella sezione "vita personale". Inoltre, la sezione "sintomi" è equamente costituita da sintomi e immagini relative alle eruzioni cutanee.

- *pc=0,93*:questa soluzione (Figura 3.74) presenta una distribuzione molto più uniforme relativamente alle tre sezioni: le domande di vita personale sono molto ridotte rispetto ai casi precedenti. Questa soluzione presenta 21 *features*, uno dei pochi casi in cui sono state ottenute soluzioni con più di 20 *features* (il valore minimo impostato) e, non presenta una prevalenza di *features* relative a una particolare domanda e/o tipo di domande.

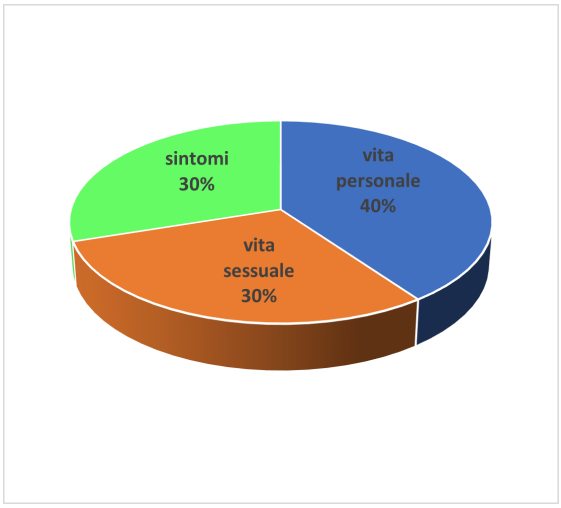


Figura 3.73: Grafico che mostra la composizione della soluzione con pc=0,84 con patologie separate e distanza euclidea.

.

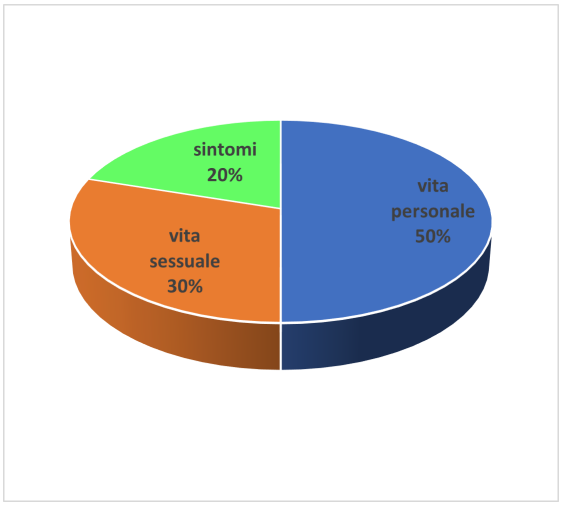


Figura 3.74: Grafico che mostra la composizione della soluzione con pc=0,93 con patologie separate e distanza euclidea.

.

Globalmente, questa serie di prove ha prodotto risultati molto buoni: innanzitutto la grande presenza delle soluzioni uguali, in 14 casi su 16 (87,5%), di cui 2 con tutte le soluzioni uguali (12,5%). In ognuna delle probabilità di crossover analizzate sono

presenti almeno una coppia di soluzioni con *fitness* uguali.
La *performance* del classificatore KNN in questo metodo è nettamente superiore ai metodi precedenti: sia dal punto di vista dell'errore di classificazione, il quale varia tra il 5,25% e il 9,18%, sia per la percentuale di patologie classificate correttamente, che varia tra il 76,28% e il 95,40%, con la maggior parte delle probabilità di crossover intorno al 80% - 85%.
Verificando la singola *performance* su ogni patologia, si è evidenziato come ogni patologia ad eccezione della classe "HPV" (la quale conta solo 2 pazienti), compaia tra le classificazioni corrette nel 100% delle soluzioni disponibili, con i seguenti risultati:

- 89,89% di pazienti correttamente classificati per il Condiloma ano-genitale;
- 86.48% di pazienti correttamente classificati per l'Herpes genitale;
- 82,80% di pazienti correttamente classificati per la Gonorrea;
- 81,22% di pazienti correttamente classificati per la Sifilide;
- 77,26% di pazienti correttamente classificati per la Clamidia;
- 70% di pazienti correttamente classificati per l'HPV;
- 69,58% di pazienti correttamente classificati per la classe "Altro".

Anche da questo punto di vista il metodo risulta più che soddisfacente, in particolare in relazione alla scarsità del dataset di partenza. Le percentuali più basse della classe "Altro" e "HPV" sono facilmente imputabili alla mancanza di un numero adeguato di pazienti che, nel primo caso, si unisce al raggruppamento di patologie diverse.

3.3.5 Scelta del metodo e analisi delle soluzioni

Considerati i risultati riportati al Paragrafo 3.3.4, l'ultimo metodo testato, che utilizza la distanza euclidea e le 8 classi diverse che comprendono i pazienti sani e i pazienti affetti dalle varie patologie, è il migliore sia dal punto di vista dell'errore di classificazione, sia dal punto di vista della classificazione delle patologie.
La selezione del metodo è necessaria in questa fase del progetto per individuare la soluzione dell'Algoritmo Genetico da utilizzare per ripetere la clusterizzazione mediante reti neurali SOM (descritta al Paragrafo 3.2.1), con un minore numero di *features*.
Si è deciso di analizzare e successivamente provare la clusterizzazione su entrambe le soluzioni descritte nel paragrafo precedente, al fine di non dare priorità né alla distinzione delle patologie (soluzione con probabilità di crossover pari a 0,93), né alla

separazione tra pazienti sani e pazienti affetti da MST (soluzione con probabilità di crossover pari a 0,84).
Si riporta, quindi, una descrizione delle *features* selezionate da ognuna delle soluzioni (Figura 3.75). Le *features* iniziali sono 66, ovvero il numero delle possibili riposte selezionabili nel questionario composto da 37 domande. Entrambe le soluzioni hanno selezionato 20 *features*, di cui 7 sono in comune, cioè il 35%, valore superiore alle aspettative dato l'elevato numero di *features* iniziali.
Dal punto di vista delle domande, invece, la soluzione con pc=0,84 contiene 13 domande, invece la soluzione con pc=0,93 ne contiene 14 (per questo fine si considera una domanda selezionata se compare almeno una delle sue *features*, indipendentemente dalla struttura della domanda stessa), di cui 7 domande sono in comune (circa il 50% delle domande selezionate), un dato meno significativo considerato che le domande sono 37 in totale.

Soluzione pc=0,84	Soluzione pc=0,93
	Sesso
Età	
	Etnia
	Gravidanza
Data dell'ultimo Pap Test	Data dell'ultimo Pap Test
	Ultimo test per la Epatite B
	Ultimo test per la Clamidia
Ultimo test per la Gonorrea	
Vaccinazione per Epatite A	
Vaccinazione per HPV	Vaccinazione per HPV
MST passate - Epatite B	MST passate - Epatite B
	MST passate - Epatite C
MST passate - Gonorrea	
MST passate - Epatite A	MST passate - Epatite A
Rapporti con più di una persona	
	Tipi di rapporti
	Vita promiscua del partner
Masturbazione	Masturbazione
	Sex toys
Uso del preservativo in rapporto orale	
	Ruolo in rapporto orale
Ruolo in rapporto vaginale	Ruolo in rapporto vaginale
Paziente costretto al rapporto	
Partner che fa uso di droghe	
	Sintomo - eruzione cutanea
	Sintomo - linfonodi
	Sintomo - congiuntivite
Sintomo - infiammazione in zona genitale	
Sintomo - sensibilità alla luce	
Sintomo - febbre	Sintomo - febbre
Immagini verruche genitali	
Immagini sifiloma	
Riassunto delle diverse immagini	

Figura 3.75: Tabelle che illustrano le *features* selezionate dalle due soluzioni dell'Algoritmo Genetico.

.

3.4 Clusterizzazione a seguito della Feature Selection

L'analisi dei 314 dati simulati ricevuti è stata ripetuta per valutare il risultato con le due soluzioni ottenute mediante l'Algoritmo Genetico, in modo da stabilire quale presenta la performance migliore. Per la clusterizzazione sono stati impostati gli stessi parametri descritti al Paragrafo 3.2.1:

- dimensione delle reti da 4x4 a 8x8;
- vicinato uguale a 1;
- 200 iterazioni;
- tagli del dendrogramma da 2 a 7, seguendo le stesse condizioni di stop.

Le *Confusion Matrix*, indicanti le distribuzioni dei pazienti sani e pazienti affetti da MST, sono state valutate mediante i medesimi due indicatori indicati in precedenza:

- la differenza tra la percentuale di pazienti sani presenti all'interno del cluster (n° sani/90), calcolata sul totale di pazienti sani (90), e la percentuale della numerosità del cluster (n° pazienti/314), calcolata sul totale dei pazienti (314). In questo modo si ha una valutazione sulla base del totale dei pazienti, in modo da non favorire i *clusters* ottenuti dai tagli maggiori.
- la percentuale costituita da pazienti sani rispetto alla numerosità del *cluster* (n° sani/n° pazienti).

3.4.1 Risultati

Si riportano i risultati per entrambe le soluzioni e ognuna delle reti SOM utilizzate e precedentemente descritte.
Per ognuna di esse saranno riportati gli elementi più rilevanti: innanzitutto i tagli che presentano le migliori variabilità intra-cluster e/o distanza inter-cluster della rete e per ogni taglio vengono riportati entrambi indipendentemente che siano i valori migliori, ma viene indicato ogni qual volta sia presente solo un parametro migliore; saranno inoltre mostrati i valori dei due indicatori per il/i *cluster/s* migliore/i della rete.
Infine, viene indicato quando un parametro e/o un indicatore risultano essere i migliori della soluzione, in questo modo è possibile valutare le *performances* ottenute con ognuna di esse.

Prove con soluzione 1 (pc=0,84)

Rete SOM 4x4

È stato selezionato solo il taglio a 2 (Figura 3.76 - 3.77), con i seguenti valori:

- variabilità intra-cluster = 35,979 (la migliore della soluzione 1)
- distanza inter-cluster = 1,074

	cluster1	cluster2
sani	20	70
malati	94	130
tot	114	200

Figura 3.76: Confusion matrix relativa alla soluzione 1, rete 4x4, taglio a 2.

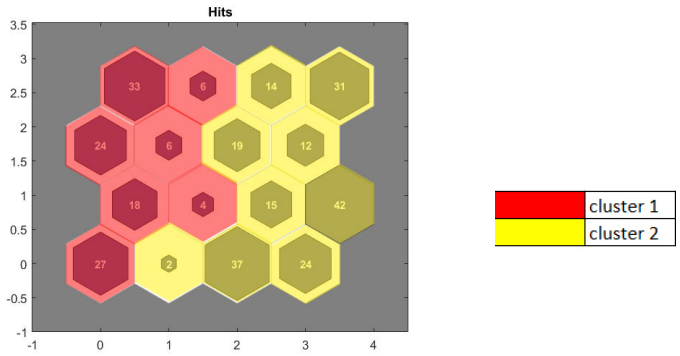


Figura 3.77: Mappa della rete SOM relativa alla soluzione 1, rete 4x4, taglio a 2, con evidenziati i neuroni relativi ai clusters.

La clusterizzazione indica che il *cluster* 2 contiene:

- 200 individui su 314 individui totali = 63,69%
- 70 pazienti sani su 90 pazienti sani totali = 77,78%

La differenza è quindi 14,08%, la migliore della soluzione 1.
La percentuale di pazienti sani nel *cluster* è 35,00% (70 pazienti sani su 200 individui), la migliore di tutte le reti della soluzione 1.

Rete SOM 5x5

È stato selezionato solo il taglio a 2 (Figura 3.78 - 3.79), con i seguenti valori:

- variabilità intra-cluster = 36,189
- distanza inter-cluster = 1,072

	cluster1	cluster2
sani	18	72
malati	85	139
tot	103	211

Figura 3.78: Confusion matrix relativa alla soluzione 1, rete 5x5, taglio a 2.

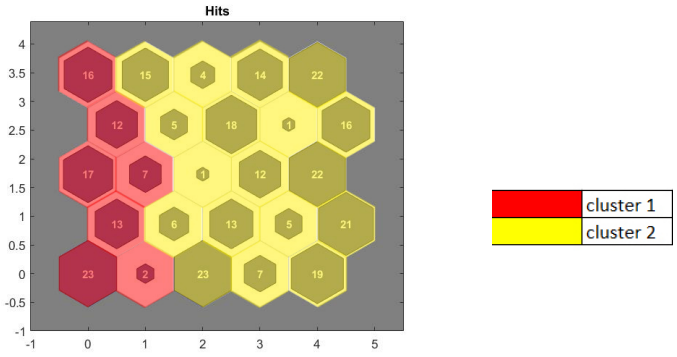


Figura 3.79: Mappa della rete SOM relativa alla soluzione 1, rete 5x5, taglio a 2, con evidenziati i neuroni relativi ai clusters.

La clusterizzazione indica che il *cluster* 2 contiene:

- 211 individui su 314 individui totali = 67,20%
- 72 pazienti sani su 90 pazienti sani totali = 80,00%

La differenza è quindi 12,80%.
La percentuale di pazienti sani nel *cluster* è 34,12% (72 pazienti sani su 211 individui).

Rete SOM 6x6
È stato selezionato solo il taglio a 2 (Figura 3.80 - 3.81), con i seguenti valori:

- variabilità intra-cluster = 36,189
- distanza inter-cluster = 1,072

La clusterizzazione indica che il *cluster* 1 contiene:

- 207 individui su 314 individui totali = 65,92%
- 72 pazienti sani su 90 pazienti sani totali = 80,00%

La differenza è quindi 14,08%.
La percentuale di pazienti sani nel *cluster* è 34,78% (72 pazienti sani su 207 individui).

	cluster1	cluster2
sani	72	18
malati	135	89
tot	207	107

Figura 3.80: Confusion matrix relativa alla soluzione 1, rete 6x6, taglio a 2.

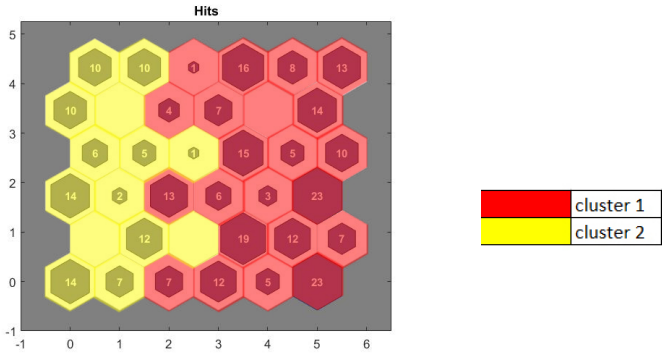


Figura 3.81: Mappa della rete SOM relativa alla soluzione 1, rete 6x6, taglio a 2, con evidenziati i neuroni relativi ai clusters.

Rete SOM 7x7

È stato selezionato solo il taglio a 2 (Figura 3.82 - 3.83), con i seguenti valori:

- variabilità intra-cluster = 36,272
- distanza inter-cluster = 1,118 (la migliore della soluzione 1)

	cluster1	cluster2
sani	70	20
malati	131	93
tot	201	113

Figura 3.82: Confusion matrix relativa alla soluzione 1, rete 7x7, taglio a 2.

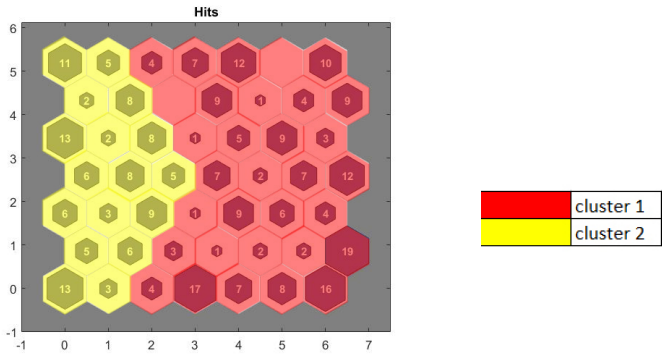


Figura 3.83: Mappa della rete SOM relativa alla soluzione 1, rete 7x7, taglio a 2, con evidenziati i neuroni relativi ai clusters.

La clusterizzazione indica che il *cluster* 1 contiene:

- 201 individui su 314 individui totali = 64,01%
- 70 pazienti sani su 90 pazienti sani totali = 77,78%

La differenza è quindi 13,77%.
La percentuale di pazienti sani nel *cluster* è 34,83% (70 pazienti sani su 201 individui).

Rete SOM 8x8
È stato selezionato solo il taglio a 2 (Figura 3.84 - 3.85), con i seguenti valori:

- variabilità intra-cluster = 36,426
- distanza inter-cluster = 1,077

	cluster1	cluster2
sani	71	19
malati	139	85
tot	210	104

Figura 3.84: Confusion matrix relativa alla soluzione 1, rete 8x8, taglio a 2.

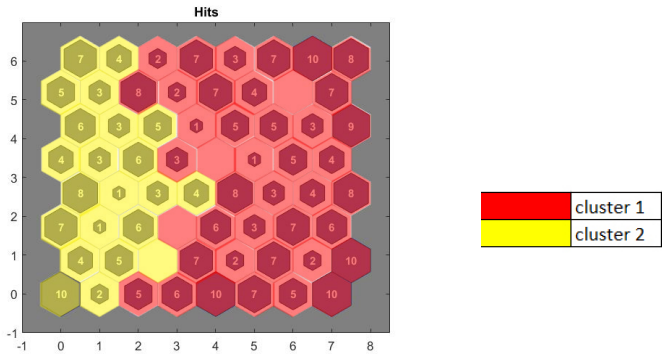


Figura 3.85: Mappa della rete SOM relativa alla soluzione 1, rete 8x8, taglio a 2, con evidenziati i neuroni relativi ai clusters.

La clusterizzazione indica che il *cluster* 1 contiene:

- 210 individui su 314 individui totali = 66,88%
- 71 pazienti sani su 90 pazienti sani totali = 78,89%

La differenza è quindi 12,01%.
La percentuale di pazienti sani nel cluster è 33,81% (71 pazienti sani su 210 individui).

Prove con soluzione 2 (pc=0,93)

Rete SOM 4x4

È stato selezionato il taglio a 2 (Figura 3.86 - 3.87), con la migliore variabilità intra-cluster della rete:

- variabilità intra-cluster = 36,940
- distanza inter-cluster = 0,882

	cluster1	cluster2
sani	7	83
malati	83	141
tot	90	224

Figura 3.86: Confusion matrix relativa alla soluzione 2, rete 4x4, taglio a 2.

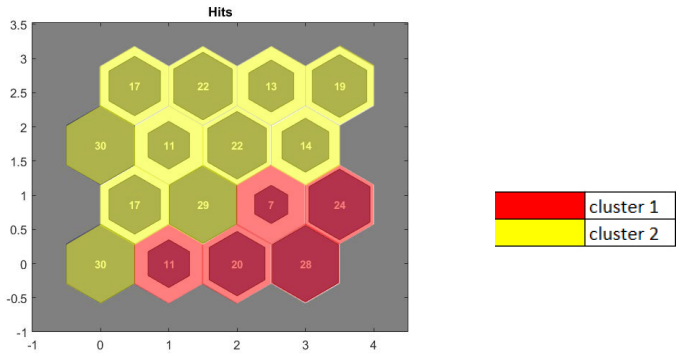


Figura 3.87: Mappa della rete SOM relativa alla soluzione 2, rete 4x4, taglio a 2, con evidenziati i neuroni relativi ai clusters.

e il taglio a 3 (Figura 3.88 - 3.89), con la migliore distanza inter-cluster della rete:

- variabilità intra-cluster = 37,827
- distanza inter-cluster = 0,990

La clusterizzazione del taglio a 2 indica che il *cluster* 1 contiene:

- 224 individui su 314 individui totali = 71,34%
- 83 pazienti sani su 90 pazienti sani totali = 92,22%

La differenza è quindi 20,88%. La percentuale di pazienti sani nel *cluster* è 37,05% (83 pazienti sani su 224 individui).

La clusterizzazione del taglio a 3 indica che il *cluster* 2 contiene:

- 145 individui su 314 individui totali = 46,18%
- 69 pazienti sani su 90 pazienti sani totali = 76,67%

La differenza è quindi 30,49%. La percentuale di pazienti sani nel *cluster* è 47,58% (69 pazienti sani su 145 individui).

	cluster1	cluster2	cluster3
sani	14	69	7
malati	65	76	83
tot	79	145	90

Figura 3.88: Confusion matrix relativa alla soluzione 2, rete 4x4, taglio a 3.

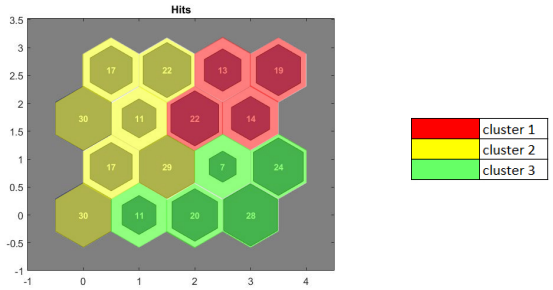


Figura 3.89: Mappa della rete SOM relativa alla soluzione 2, rete 4x4, taglio a 3, con evidenziati i neuroni relativi ai clusters.

Rete SOM 5x5
È stato selezionato solo il taglio a 2 (Figura 3.90 - 3.91), con i seguenti valori:

- variabilità intra-cluster = 36,889 (la migliore della soluzione 2)
- distanza inter-cluster = 1,014

La clusterizzazione indica che il *cluster* 2 contiene:

- 215 individui su 314 individui totali = 68,47%
- 75 pazienti sani su 90 pazienti sani totali = 83,33%

La differenza è quindi 14,86%.
La percentuale di pazienti sani nel *cluster* è 34,88% (75 pazienti sani su 215 individui).

Rete SOM 6x6
È stato selezionato solo il taglio a 2 (Figura 3.92 - 3.93), con i seguenti valori:

- variabilità intra-cluster = 37,015

	cluster1	cluster2
sani	75	15
malati	140	84
tot	215	99

Figura 3.90: Confusion matrix relativa alla soluzione 2, rete 5x5, taglio a 2.

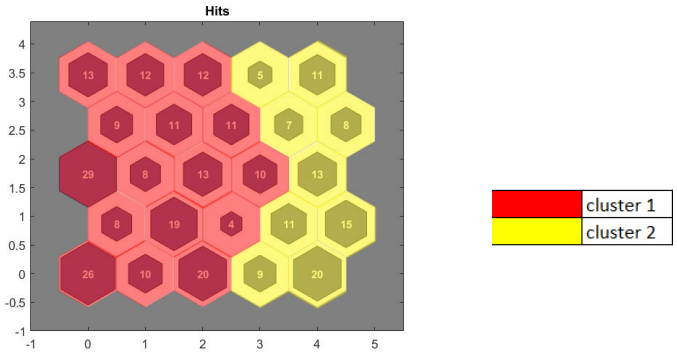


Figura 3.91: Mappa della rete SOM relativa alla soluzione 2, rete 5x5, taglio a 2, con evidenziati i neuroni relativi ai clusters.

	cluster1	cluster2
sani	79	11
malati	137	87
tot	216	98

Figura 3.92: Confusion matrix relativa alla soluzione 2, rete 6x6, taglio a 2.

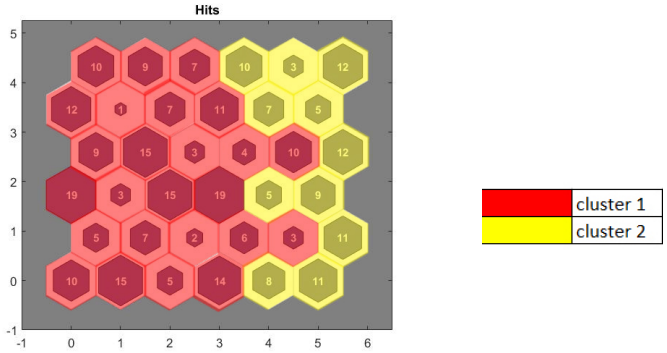


Figura 3.93: Mappa della rete SOM relativa alla soluzione 2, rete 6x6, taglio a 2, con evidenziati i neuroni relativi ai clusters.

- distanza inter-cluster = 1,039

La clusterizzazione indica che il *cluster* 2 contiene:

- 207 individui su 314 individui totali = 68,79%
- 72 pazienti sani su 90 pazienti sani totali = 87,78%

La differenza è quindi 18,99%.
La percentuale di pazienti sani nel *cluster* è 36,57% (79 pazienti sani su 216 individui).

Rete SOM 7x7
È stato selezionato solo il taglio a 2 (Figura 3.94 - 3.95), con i seguenti valori:

- variabilità intra-cluster = 37,005
- distanza inter-cluster = 1,064 (la migliore della soluzione 2)

	cluster1	cluster2
sani	16	74
malati	87	137
tot	103	211

Figura 3.94: Confusion matrix relativa alla soluzione 2, rete 7x7, taglio a 2.

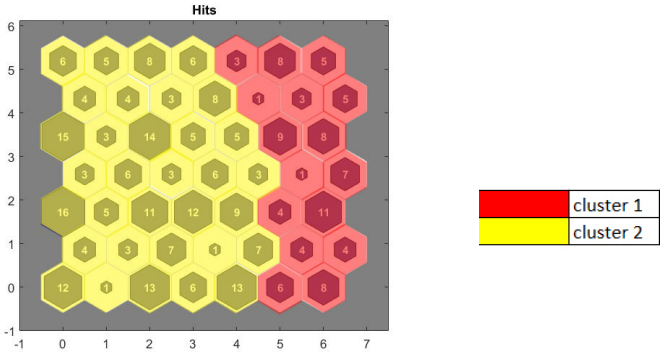


Figura 3.95: Mappa della rete SOM relativa alla soluzione 2, rete 7x7, taglio a 2, con evidenziati i neuroni relativi ai clusters.

La clusterizzazione indica che il *cluster* 2 contiene:

- 216 individui su 314 individui totali = 68,79%
- 79 pazienti sani su 90 pazienti sani totali = 87,78%

La differenza è quindi 18,99%.
La percentuale di pazienti sani nel cluster è 36,57% (79 pazienti sani su 216 individui).

Rete SOM 8x8

È stato selezionato il taglio a 2 (Figura 3.96 - 3.97), con la migliore variabilità intra-cluster della rete:

- variabilità intra-cluster = 37,073
- distanza inter-cluster = 0,843

	cluster1	cluster2
sani	10	80
malati	77	147
tot	87	227

Figura 3.96: Confusion matrix relativa alla soluzione 2, rete 8x8, taglio a 2.

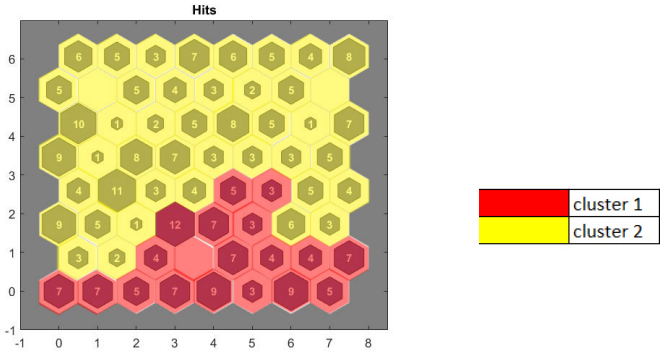


Figura 3.97: Mappa della rete SOM relativa alla soluzione 2, rete 8x8, taglio a 2, con evidenziati i neuroni relativi ai clusters.

e il taglio a 3 (Figura 3.98 - 3.99, con la migliore distanza inter-cluster della rete:

- variabilità intra-cluster = 37,073
- distanza inter-cluster = 0,920

La clusterizzazione del taglio a 2 indica che il *cluster* 2 contiene:

- 227 individui su 314 individui totali = 72,29%
- 80 pazienti sani su 90 pazienti sani totali = 88,89%

La differenza è quindi 16,60%. La percentuale di pazienti sani nel *cluster* è 35,24% (80 pazienti sani su 227 individui).

	cluster1	cluster2	cluster3
sani	13	67	10
malati	80	67	77
tot	93	134	87

Figura 3.98: Confusion matrix relativa alla soluzione 2, rete 8x8, taglio a 3.

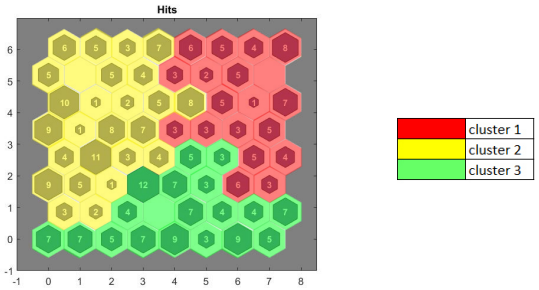


Figura 3.99: Mappa della rete SOM relativa alla soluzione 2, rete 8x8, taglio a 3, con evidenziati i neuroni relativi ai clusters.

La clusterizzazione del taglio a 3 indica che il *cluster* 2 contiene:

- 134 individui su 314 individui totali = 42,86%
- 67 pazienti sani su 90 pazienti sani totali = 74,44%

La differenza è quindi 31,77%.
La percentuale di pazienti sani nel *cluster* è 50,00% (67 pazienti sani su 134 individui).
Entrambi gli indicatori risultano i migliori della soluzione 2.

Confronto delle due soluzioni dell'Algoritmo Genetico

La clusterizzazione della seconda soluzione, corrispondente a una probabilità di crossover pari 0,93, introduce un miglioramento nella clusterizzazione rispetto al primo indicatore: dal 25,17%, il risultato migliore delle prove descritte al Paragrafo 3.2.2 (corrispondente alla sezione "vita sessuale e sintomi"), al 31,77% ottenuto a seguito di questa *Feature Selection*.
I risultati della clusterizzazione della prima soluzione (probabilità di crossover pari a 0,84), invece, risultano peggiori di tutte le prove effettuate in precedenza, con valori dei due indicatori (14,08% e 35,00%) leggermente inferiori ai valori migliori della sezione "vita sessuale", la quale aveva ottenuto i risultati peggiori delle prove precedenti (14,27% e 36,00%).
La variabilità intra-cluster, che si era già notato essere fortemente legata al numero di *features* in oggetto, presenta valori leggermente superiori alle prove con la sezione "vita sessuale", costituita da 17 *features*; queste prove, infatti, sono state effettuate entrambe con 20 *features*.

La distanza inter-cluster, complessivamente, peggiora in entrambi i casi, relativamente a tutte le prove precedenti.
Si procederà quindi considerando solo la *Feature Selection* ottenuta con la probabilità di crossover pari a 0,93, l'unica in grado di introdurre un miglioramento nelle *performances* delle reti neurali SOM.

3.4.2 Risultati ripetizioni

Al fine di aumentare la ripetibilità dei risultati ottenuti mediante le prove di clusterizzazione, si è deciso di ripetere per altre quattro volte le prove sulle tre sezioni del questionario e sulla selezione di domande ottenuta mediante l'Algoritmo Genetico, raggiungendo così un totale di cinque prove. Si ritiene infatti prematuro tentare l'utilizzo di un metodo di classificazione supervisionato con i pochi dati a disposizione, mentre un maggior numero di prove permette di estrapolare informazioni globali relative ai diversi subset, le quali permetteranno una comparazione più efficace.

Parametri

Si riportano quindi i valori dei parametri per:

- *Sezione "intero questionario"* (Figure 3.100 - 3.104): le ripetizioni hanno prodotto in entrambi i parametri un aumento dei valori, aspetto positivo per la distanza inter-cluster, ma negativo per la variabilità intra-cluster. Tuttavia, questo risultato era atteso, considerato l'utilizzo di tutte le 66 *features* e l'aumento dei valori testati.

		variabilità intra-cluster					media	dev. Std
		314	rip_1	rip_2	rip_3	rip_4		
intero quest	4x4	221,850	221,850	223,688	221,721	221,721	222,166	0,836
	5x5	221,185	221,185	222,982	225,125	224,859	223,067	1,604
	6x6	221,648	221,648	229,246	222,583	222,195	223,464	3,101
	7x7	221,367	221,367	222,329	221,884	222,709	221,931	0,505
	8x8	222,078	222,078	227,575	227,945	221,238	224,183	3,077

Figura 3.100: Valori della variabilità intra-cluster relativa alla sezione "intero questionario", per ogni dimensione delle reti neurali SOM.

.

		distanza inter-cluster					media	dev. Std
		314	rip_1	rip_2	rip_3	rip_4		
intero quest	4x4	1,597	1,597	1,640	1,589	1,581	1,601	0,023
	5x5	1,526	1,526	1,602	1,627	1,646	1,585	0,046
	6x6	1,569	1,569	1,713	1,589	1,577	1,603	0,059
	7x7	1,622	1,622	1,568	1,627	1,613	1,611	0,023
	8x8	1,647	1,647	1,700	1,665	1,627	1,657	0,027

Figura 3.101: Valori della distanza inter-cluster relativa alla sezione "intero questionario", per ogni dimensione delle reti neurali SOM.

.

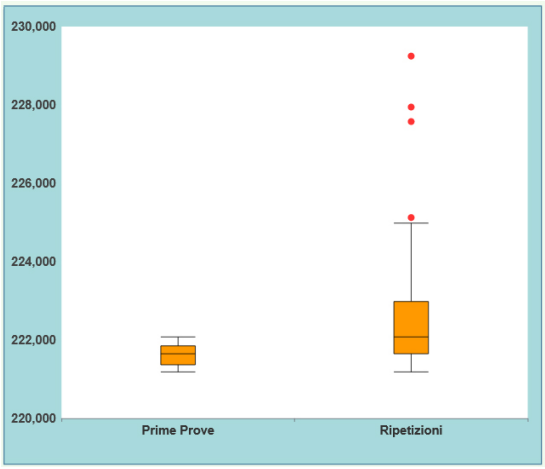


Figura 3.102: Boxplot della variabilità intra-cluster nelle prime prove (5 valori) e dopo le ripetizioni (25 valori) per la sezione "intero questionario".

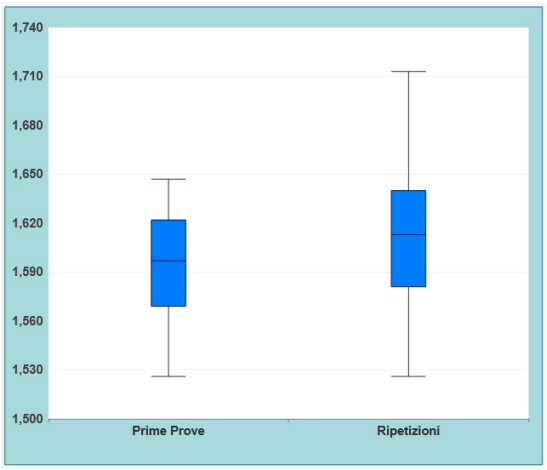


Figura 3.103: Boxplot della distanza inter-cluster nelle prime prove (5 valori) e dopo le ripetizioni (25 valori) per la sezione "intero questionario".

- *Sezione "vita sessuale"* (Figura 3.105 - 3.109): per questa sezione si registra un peggioramento di entrambi i parametri: infatti, la variabilità intra-cluster è aumentata (solo la metà dei due boxplots risultano sovrapposti), e la distanza inter-cluster è fortemente diminuita (i due boxplots non sono quasi sovrapposti).

	variabilità intra-cl		distanza inter-cl	
media primo tentativo	221,698	0,358	1,592	0,047
media tot	223,192	1,150	1,611	0,027

Figura 3.104: Tabella con media e deviazione standard per entrambi i parametri per la sezione "intero questionario".

.

		variabilità intra-cluster					media	dev. Std
		314	rip_1	rip_2	rip_3	rip_4		
vita sessuale	4x4	34,286	34,405	33,727	34,405	34,405	34,245	0,294
	5x5	33,886	33,286	34,063	33,286	33,286	33,561	0,338
	6x6	33,327	34,192	34,452	34,192	34,192	34,071	0,140
	7x7	33,327	34,337	34,549	34,337	34,337	34,178	0,132
	8x8	33,681	34,023	33,004	34,023	34,023	33,751	0,442

Figura 3.105: Valori della variabilità intra-cluster relativa alla sezione "vita sessuale", per ogni dimensione delle reti neurali SOM.

.

- *Sezione "vita sessuale e sintomi"* (Figura 3.110-3.3.114): la variabilità intra-cluster, in questo caso, peggiora maggiormente in relazione al valore medio ma, osservando i boxplots, è possibile notare che il peggioramento risulta meno sensibile. La distanza inter-cluster presenta un comportamento simile.

		distanza inter-cluster					media	dev. Std
		314	rip_1	rip_2	rip_3	rip_4		
vita sessuale	4x4	1,150	1,147	1,217	1,147	1,147	1,162	0,030
	5x5	1,385	1,385	1,140	1,385	1,385	1,336	0,107
	6x6	1,222	1,105	1,078	1,105	1,105	1,123	0,016
	7x7	1,342	1,103	1,120	1,103	1,103	1,154	0,022
	8x8	1,208	1,202	1,458	1,202	1,202	1,254	0,111

Figura 3.106: Valori della distanza inter-cluster relativa alla sezione "vita sessuale", per ogni dimensione delle reti neurali SOM.

.

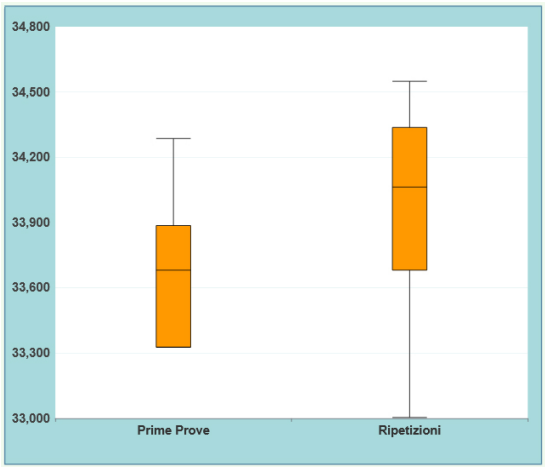


Figura 3.107: Boxplot della variabilità intra-cluster nelle prime prove (5 valori) e dopo le ripetizioni (25 valori) per la sezione "vita sessuale".

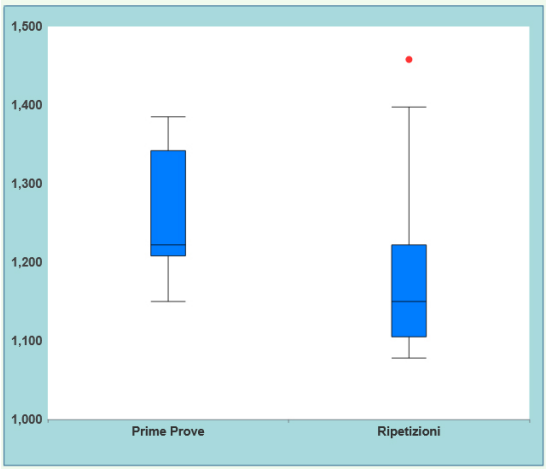


Figura 3.108: Boxplot della distanza inter-cluster nelle prime prove (5 valori) e dopo le ripetizioni (25 valori) per la sezione "vita sessuale".

- *Soluzione Algoritmo Genetico* (Figura 3.115-3.3.119): la soluzione dell'AG è l'unica che è rimasta stabile con l'aggiunta delle ripetizioni, registrando solo un leggero aumento di entrambi i parametri (maggiore nel caso dei boxplots della distanza inter-cluster), un dato positivo che indica che la *Features Selection* introdotta in queste prove produce risultati più ripetibili.

	variabilità intra-cl		distanza inter-cl	
media primo tentativo	33,701	0,405	1,262	0,098
media tot	33,961	0,293	1,206	0,088

Figura 3.109: Tabella con media e deviazione standard per entrambi i parametri per la sezione "vita sessuale".

.

		variabilità intra-cluster					media	dev. Std
		314	rip_1	rip_2	rip_3	rip_4		
vita sessuale + sintomi	4x4	94,522	94,522	94,424	94,603	93,044	94,223	0,641
	5x5	93,423	96,848	95,503	93,349	93,219	94,468	1,527
	6x6	93,144	95,254	95,963	92,733	92,645	93,948	1,484
	7x7	94,432	94,416	95,302	95,658	99,739	95,909	2,055
	8x8	93,108	94,736	93,600	93,373	93,143	93,592	0,615

Figura 3.110: Valori della variabilità intra-cluster relativa alla sezione "vita sessuale e sintomi", per ogni dimensione delle reti neurali SOM.

.

Indicatori della clusterizzazione

Relativamente alla clusterizzazione, sono stati analizzati i due indicatori (descritti al Paragrafo 3.2.1), in modo da verificare l'utilità della soluzione dell'Algoritmo Genetico per la distinzione tra pazienti sani e pazienti affetti da MST.
In Figura 3.120 vengono riportati i risultati del primo indicatore, in cui si è deciso di considerare solo i risultati migliori di ogni serie di prove per le valutazioni globali di metodi. La soluzione dell'Algoritmo Genetico ha prodotto un leggero miglioramento rispetto alla sezione "vita sessuale e sintomi", la migliore delle prime prove: osservando le prove della colonna "314", il miglioramento è più consistente, perciò le ripetizioni hanno, probabilmente, prodotto un risultato più realistico, considerato che in genere i sintomi risultano molto utili per la clusterizzazione. Il boxplot (Figura 3.121) però, mostra un netto miglioramento della percentuale media della differenza tra la percentuale di pazienti sani nel *cluster* e la percentuale dei pazienti totali nel *cluster*.
In Figura 3.122 vengono riportati i risultati del secondo indicatore, in cui si è deciso di considerare solo i risultati migliori di ogni serie di prove per le valutazioni globali di metodi. In questo caso la soluzione dell'Algoritmo Genetico non ha prodotto un miglioramento, probabilmente anche a causa del maggior numero di *clusters* prodotti mediamente dalla sezione "vita sessuale e sintomi", in cui il dendrogramma veniva tagliato più in basso. Il secondo indicatore è costituito dalle percentuali di pazienti sani contenuti in un singolo *cluster* e un numero maggiore di tagli

		distanza inter-cluster					media	dev. Std
		314	rip_1	rip_2	rip_3	rip_4		
vita sessuale + sintomi	4x4	1,299	1,299	1,231	1,298	1,272	1,280	0,028
	5x5	1,274	1,358	1,280	1,286	1,312	1,302	0,031
	6x6	1,301	1,193	1,231	1,338	1,248	1,262	0,054
	7x7	1,224	1,329	1,263	1,288	1,059	1,233	0,104
	8x8	1,375	1,252	1,297	1,248	1,244	1,283	0,023

Figura 3.111: Valori della distanza inter-cluster relativa alla sezione "vita sessuale e sintomi", per ogni dimensione delle reti neurali SOM.

.

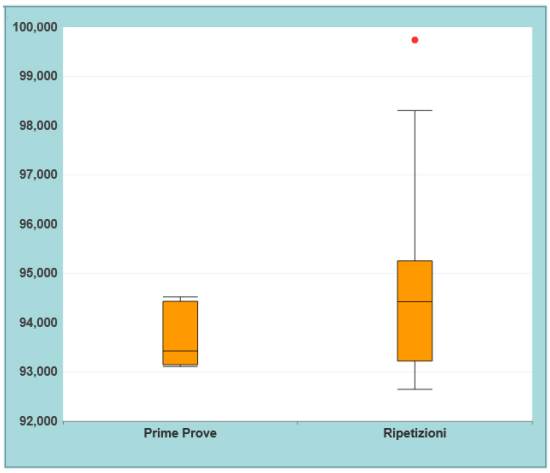


Figura 3.112: Boxplot della variabilità intra-cluster nelle prime prove (5 valori) e dopo le ripetizioni (25 valori) per la sezione "vita sessuale e sintomi".

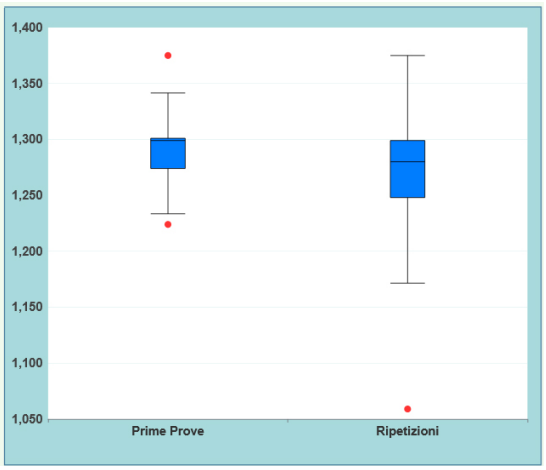


Figura 3.113: Boxplot della distanza inter-cluster nelle prime prove (5 valori) e dopo le ripetizioni (25 valori) per la sezione "vita sessuale e sintomi".

può produrre più facilmente percentuali più alte. Queste considerazioni vengono confermate dal boxplot mostrato in Figura 3.123.

	variabilità intra-cl		distanza inter-cl	
media primo tentativo	93,726	0,697	1,295	0,055
media tot	94,428	0,890	1,272	0,026

Figura 3.114: Tabella con media e deviazione standard per entrambi i parametri per la sezione "vita sessuale e sintomi".

.

		variabilità intra-cluster					media	dev. Std
		314	rip_1	rip_2	rip_3	rip_4		
GA	4x4	36,940	37,644	36,967	36,902	36,797	37,050	0,333
	5x5	36,889	36,771	37,026	37,102	36,974	36,953	0,123
	6x6	37,015	36,555	37,222	36,854	37,024	36,934	0,245
	7x7	37,005	36,640	37,173	36,498	37,211	36,905	0,316
	8x8	37,073	37,288	37,048	37,095	37,532	37,207	0,192

Figura 3.115: Valori della variabilità intra-cluster relativa alla soluzione dell'AG, per ogni dimensione delle reti neurali SOM.

.

Distribuzione delle patologie

Una volta verificate le performances sulla distinzione tra pazienti sani e pazienti affetti da MST, si è ripetuta la verifica della distribuzione delle patologie, secondo il metodo descritto al Paragrafo 3.2.3, in quanto la soluzione dell'Algoritmo Genetico che è stata selezionata è quella con una migliore distinzione delle singole patologie. Dai 5 grafici delle patologie, per ogni ripetizione si è creata una tabella (in Figura 3.124), che mostra quale sezione risulta più adatta per la distinzione di ogni singola patologia, al fine di individuare l'efficacia della soluzione dell'Algoritmo Genetico: essa migliora la distinzione delle patologie nel 50% dei casi, percentuale che sale al 60% se si esclude la classe "Altro", che al momento conta solo 5 pazienti relativi a patologie diverse.
Per poter ottenere un grafico (Figura 3.125) simile a quello precedentemente descritto, si è deciso di mediare i risultati delle 5 ripetizioni per ognuna delle categorie analizzate (differenza di percentuali maggiore del 10%, differenza di percentuali maggiore del 25%, classificazione "perfetta"), considerando il totale delle strutture disponibili. Il grafico ha permesso di confermare i risultati indicati in precedenza. È possibile, infatti, osservare come le patologie presentino degli andamenti paragonabili:

- la *Clamidia* presenta un netto miglioramento nell'utilizzo dell'Algoritmo

		distanza inter-cluster					media	dev. Std
		314	rip_1	rip_2	rip_3	rip_4		
GA	4x4	0,990	0,994	0,996	1,004	0,986	0,994	0,006
	5x5	1,014	1,043	1,051	1,054	1,057	1,044	0,006
	6x6	1,039	0,926	0,970	0,972	1,056	0,993	0,047
	7x7	1,061	1,021	1,053	0,996	1,049	1,036	0,023
	8x8	0,920	1,003	1,054	1,031	1,041	1,010	0,021

Figura 3.116: Valori della distanza inter-cluster relativa alla soluzione dell'AG, per ogni dimensione delle reti neurali SOM.

.

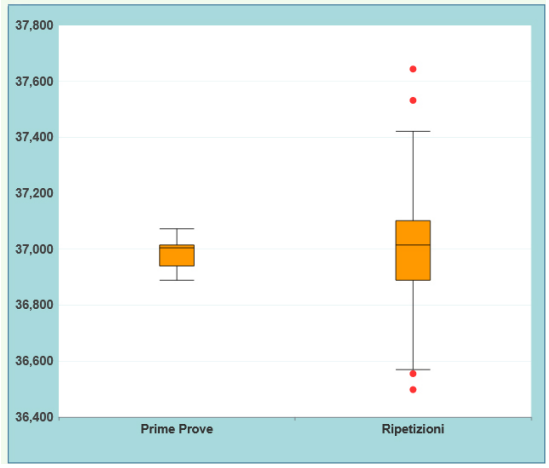


Figura 3.117: Boxplot della variabilità intra-cluster nelle prime prove (5 valori) e dopo le ripetizioni (25 valori) per la soluzione dell'AG.

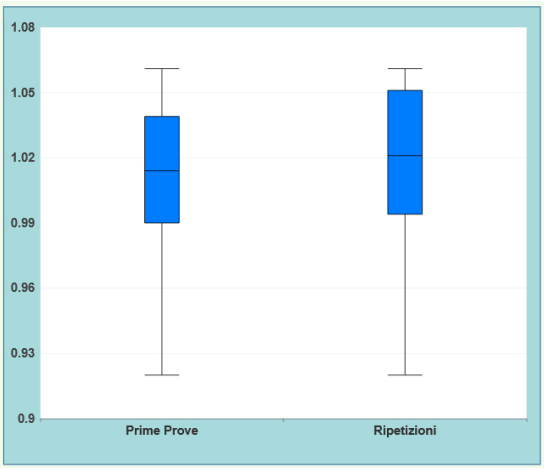


Figura 3.118: Boxplot della distanza inter-cluster nelle prime prove (5 valori) e dopo le ripetizioni (25 valori) per la soluzione dell'AG.

Genetico, come confermano i grafici ottenuti dalle singole ripetizioni;

- la *Sifilide*, che nella tabella risultava migliorata dall'Algoritmo Genetico in tutti i casi tranne uno (in cui la sezione "vita sessuale e sintomi" è risultata migliore) presenta in questo grafico un minor numero di strutture evidenziate, ma con un maggior numero di strutture con differenza di percentuali maggiore di 25%;

- il *Condiloma ano-genitale* presenta una riduzione delle patologie selezionate rispetto alle sezioni "vita sessuale" e "vita sessuale e sintomi", inoltre tutte le strutture dell'Algoritmo Genetico appartengono alla categoria con differenza di

	variabilità intra-cl		distanza inter-cl	
media primo tentativo	36,984	0,071	1,005	0,054
media tot	37,010	0,123	1,015	0,024

Figura 3.119: Tabella con media e deviazione standard per entrambi i parametri per la soluzione dell'AG.

.

		Differenza					media dei migliori	dev. std dei migliori
		314	rip1	rip2	rip3	rip4		
intero quest	4x4	18,68%	18,68%	20,72%	19,32%	19,32%	20,60%	3,36%
	5x5	18,69%	18,69%	14,54%	18,95%	19,77%		
	6x6	18,39%	18,39%	26,55%	14,39%	15,03%		
	7x7	15,19%	15,19%	15,51%	17,41%	12,95%		
	8x8	14,86%	14,86%	14,68%	17,22%	14,55%		
vita sessuale	4x4	12,69%	9,68%	10,27%	9,68%	9,68%	13,97%	0,37%
	5x5	12,92%	12,92%	13,32%	12,92%	12,92%		
	6x6	14,20%	12,85%	11,59%	12,85%	12,85%		
	7x7	13,72%	10,95%	11,90%	10,95%	10,95%		
	8x8	11,05%	14,11%	11,90%	14,11%	14,11%		
vita sessuale + sintomi	4x4	20,26%	20,26%	28,54%	20,58%	31,44%	28,57%	3,74%
	5x5	16,45%	15,94%	15,06%	14,59%	32,12%		
	6x6	25,05%	24,71%	32,67%	18,05%	17,72%		
	7x7	25,17%	15,80%	28,70%	28,16%	26,12%		
	8x8	16,46%	16,60%	18,02%	17,88%	17,40%		
GA	4x4	30,49%	21,24%	30,49%	30,81%	25,50%	29,27%	2,62%
	5x5	14,86%	21,68%	17,88%	16,30%	16,77%		
	6x6	18,99%	27,64%	20,28%	27,47%	15,66%		
	7x7	15,02%	18,36%	30,64%	27,15%	15,82%		
	8x8	31,77%	20,75%	17,09%	15,98%	15,78%		

Figura 3.120: Tabella relativa al primo indicatore della clusterizzazione.

.

percentuali maggiore di 10%. Nella tabella la sezione "vita sessuale e sintomi" risulta migliore in quasi tutte le singole ripetizioni, dato che viene confermato dal fatto che questa sezione, nel grafico, presenta più strutture con differenza di percentuali maggiore del 25%;

- l'*Herpes genitale* presenta, come la Clamidia, un miglioramento con l'Algoritmo Genetico sia per numero di strutture sia per la categoria, sia nella tabella, sia nel grafico;

- la *Gonorrea*, nelle singole patologie, presenta quattro ripetizioni con "intero questionario" come miglior sezione; nel grafico, la sezione "intero questionario"

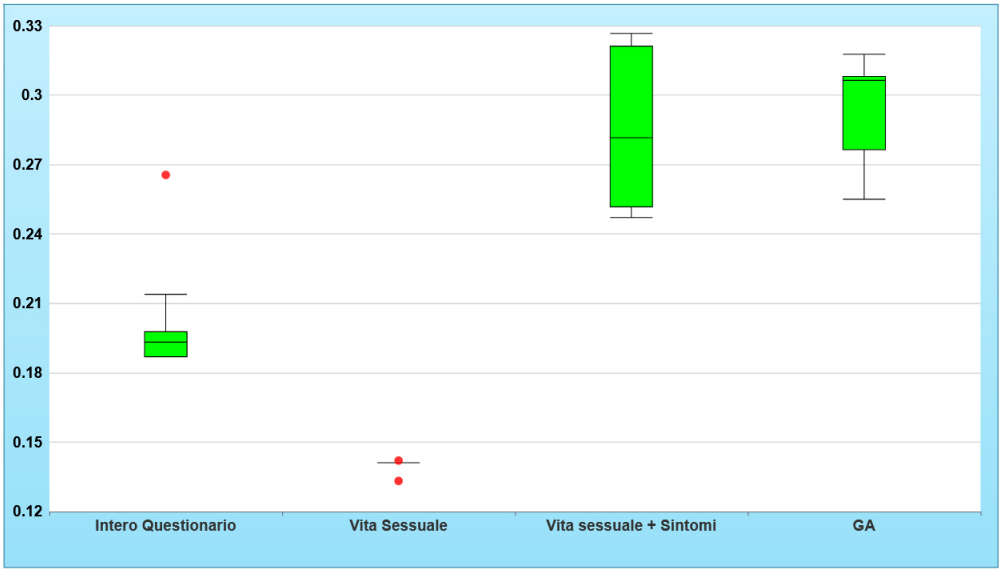


Figura 3.121: Boxplot relativo al primo indicatore della clusterizzazione.

		% sani nel cluster					media dei migliori	dev. std dei migliori
		314	rip1	rip2	rip3	rip4		
intero quest	4x4	36,94%	36,94%	36,84%	37,31%	37,31%	42,05%	5,92%
	5x5	36,95%	36,95%	34,72%	46,81%	36,36%		
	6x6	38,07%	38,07%	50,00%	34,95%	35,29%		
	7x7	35,26%	35,26%	35,43%	36,23%	33,93%		
	8x8	34,88%	34,88%	34,17%	35,15%	34,93%		
vita sessuale	4x4	35,54%	34,18%	33,17%	34,18%	34,18%	35,96%	0,17%
	5x5	33,20%	33,20%	35,67%	33,20%	33,20%		
	6x6	36,00%	35,50%	35,22%	35,50%	35,50%		
	7x7	33,60%	34,78%	35,44%	34,78%	34,78%		
	8x8	33,33%	36,05%	35,44%	36,05%	36,05%		
vita sessuale + sintomi	4x4	37,26%	37,26%	42,47%	39,71%	47,65%	52,63%	10,43%
	5x5	35,48%	35,88%	36,36%	36,20%	56,19%		
	6x6	40,98%	39,15%	44,38%	36,58%	36,15%		
	7x7	60,50%	34,95%	42,33%	55,91%	63,23%		
	8x8	35,96%	35,45%	35,62%	36,11%	35,98%		
GA	4x4	47,58%	38,83%	47,59%	47,91%	39,43%	46,46%	4,07%
	5x5	34,88%	37,61%	36,36%	35,75%	35,88%		
	6x6	36,57%	47,37%	38,22%	45,03%	35,41%		
	7x7	35,07%	36,49%	47,29%	44,74%	35,61%		
	8x8	50,00%	38,00%	36,06%	35,58%	34,41%		

Figura 3.122: Tabella relativa al secondo indicatore della clusterizzazione.

.

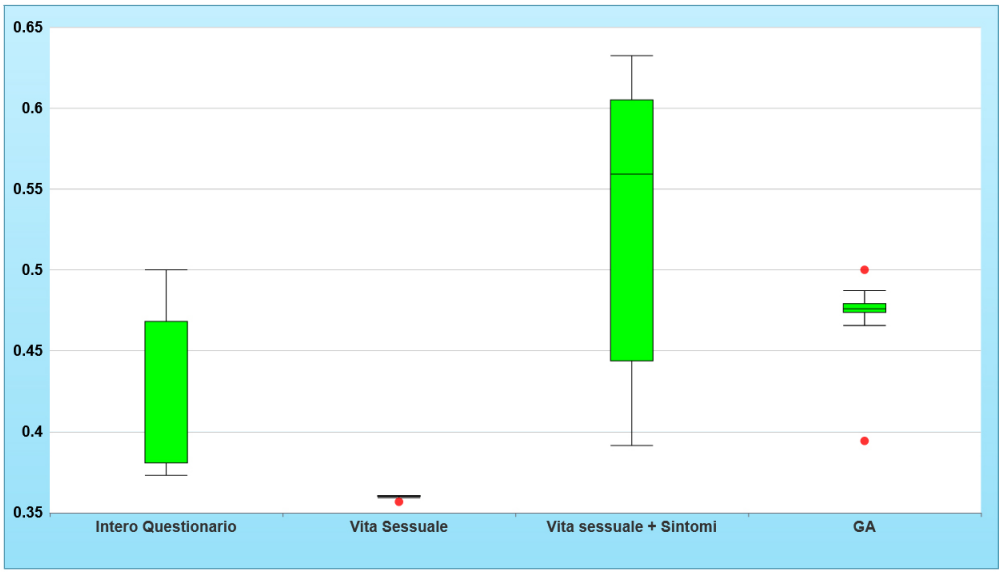


Figura 3.123: Boxplot relativo al secondo indicatore della clusterizzazione.

	314	n1	n2	n3	n4
Clamidia	GA	GA	2+3	GA	GA
Sifilide	GA	GA	2+3	GA	GA
Condiloma	GA	2+3	2+3	2+3	2+3
Herpes	GA	2+3/GA	GA	GA	GA
Gonorrea	2+3/GA	intero	intero	intero	intero
Altro	intero	intero	2+3	2+3	2+3

Figura 3.124: Tabella che indica quale sezione è maggiormente utile alla distinzione di ogni patologia per ogni serie di prove ("intero" indica la sezione "intero questionario", "2" indica la sezione "vita sessuale", "2+3" indica la sezione "vita sessuale e sintomi", "GA" indica la soluzione dell'Algoritmo Genetico).

.

ha un numero di strutture leggermente minore, ma con differenza di percentuali maggiori;

- per la classe *"Altro"*, nelle singole ripetizioni le sezioni migliori sono "intero questionario" e "vita sessuale e sintomi": infatti, la prima ha prevalenze migliori, mentre la seconda presenta un maggior numero di strutture selezionato e, in alcuni casi, i pazienti sono stati collocati tutti in un solo *cluster*.

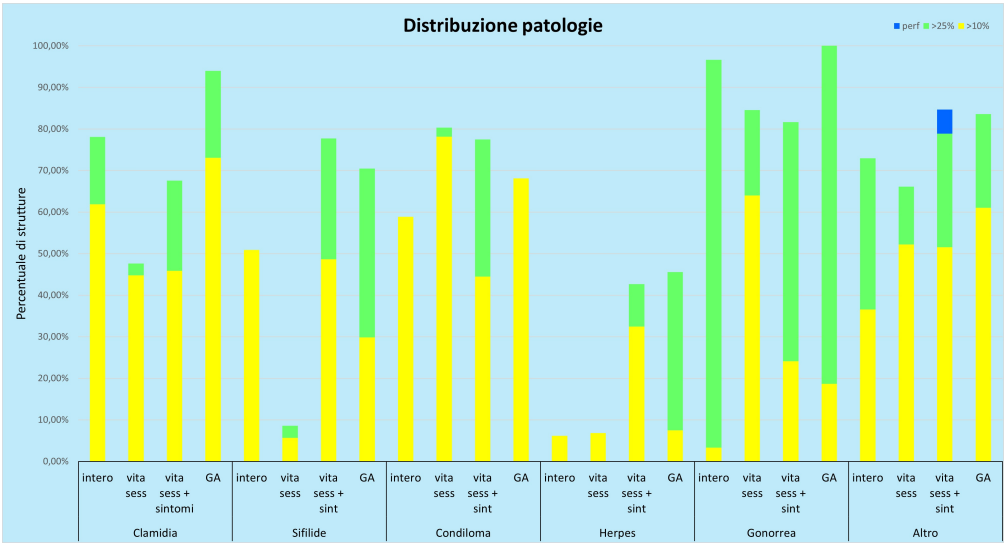


Figura 3.125: Grafico che indica la distribuzione delle patologie relative alle 5 ripetizioni.

.

Capitolo 4

Discussione dei risultati

Si procede ora alla discussione dei risultati delle ultime clusterizzazioni sotto diversi aspetti:

- i valori di variabilità intra-cluster e distanza inter-cluster;
- il primo indicatore di clusterizzazione, valutando quindi l'efficacia della *Feature Selection* ottenuta indipendentemente dal taglio e della numerosità dei *clusters*;
- il secondo indicatore di clusterizzazione, il quale, invece, permette di verificare la distribuzione di pazienti sani relativa alla numerosità del *cluster*;
- l'eventuale influenza della dimensione delle reti;
- la tipologia di tagli del dendrogramma ottenuti;
- il possibile collegamento tra alcune delle singole patologie.

Successivamente, verrà analizzata nel dettaglio la correlazione tra le domande e le singole *features* selezionate e i risultati della ricerca bibliografica riportati al Capitolo 2.

Valori dei due parametri considerati

Come viene mostrato in Figura 4.1, la variabilità intra-cluster peggiora in ognuna delle sezioni del questionario analizzate, inoltre il suo valore risulta legato al numero di *features* utilizzate durante le prove:

- 66 *features* per la sezione "intero questionario", che corrisponde al valore maggiore;
- 35 *features* per la sezione "vita sessuale";

(a)

	variabilità intra-cl		distanza inter-cl	
media primo tentativo	221,698	0,358	1,592	0,047
media tot	223,192	1,150	1,611	0,027

(b)

	variabilità intra-cl		distanza inter-cl	
media primo tentativo	33,701	0,405	1,262	0,098
media tot	33,961	0,293	1,206	0,088

(c)

	variabilità intra-cl		distanza inter-cl	
media primo tentativo	93,726	0,697	1,295	0,055
media tot	94,428	0,890	1,272	0,026

(d)

	variabilità intra-cl		distanza inter-cl	
media primo tentativo	36,984	0,071	1,005	0,054
media tot	37,010	0,123	1,015	0,024

Figura 4.1: Tabella con media e deviazione standard per entrambi i parametri per: (a) la sezione "intero questionario", (b) la sezione "vita sessuale", (c) la sezione "vita sessuale e sintomi", (d) la soluzione dell'Algoritmo Genetico.

.

- 17 *features* per la sezione "vita sessuale e sintomi", che corrisponde al valore minore;
- 20 *features* per la soluzione dell'Algoritmo Genetico che, infatti, possiede valori leggermente superiori alla sezione "vita sessuale".

La distanza inter-cluster, invece, migliora in tutti i casi ad eccezione della sezione "vita sessuale e sintomi", sebbene la diminuzione sia ridotta.
La diversa variazione dei due parametri viene confermata dai due grafici in Figura 4.2 e Figura 4.3, che derivano dalle tabelle mostrate al Paragrafo 3.4.2: la variabilità intra-cluster diminuisce solo nel 25% delle strutture (viene riportata anche la distribuzione nelle sezioni delle strutture con variabilità intra-cluster diminuita), mentre la distanza inter-cluster aumenta per il 55% delle strutture.

Risultati della clusterizzazione: primo e secondo indicatore

Come riportato in Figura 3.120 nel Paragrafo 3.4.2, il primo indicatore ha prodotto un leggero miglioramento rispetto alla sezione "vita sessuale e sintomi" che, con le ripetizioni, viene attenuata rispetto alle prime prove. La media superiore e la minore deviazione standard indicano anche una riduzione dei valori. Il motivo di questo percettibile miglioramento può essere dato dall'importanza dell'elenco dei sintomi e dall'utilizzo di dati simulati dai medici.
In Figura 3.122 vengono riportati i risultati del secondo indicatore, in cui la soluzione dell'Algoritmo Genetico non ha migliorato la clusterizzazione. Tale risultato è la conseguenza del maggior numero di *clusters* prodotti mediamente dalla sezione "vita sessuale e sintomi", in cui il dendrogramma veniva tagliato più in basso, rendendo più probabili le percentuali più alte di pazienti sani relative alla numerosità di un singolo *cluster*.

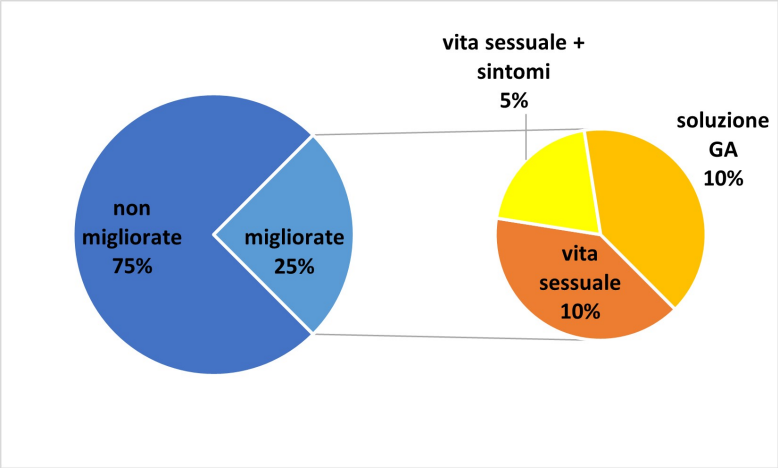


Figura 4.2: Grafico che riporta le strutture con variabilità intra-cluster diminuita, di cui si riporta anche la distribuzione nelle varie sezioni.

.

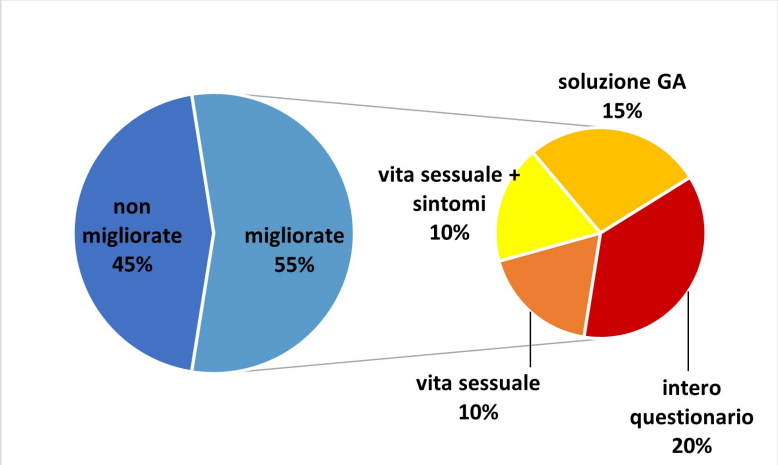


Figura 4.3: Grafico che riporta le strutture con distanza inter-cluster diminuita, di cui si riporta anche la distribuzione nelle varie sezioni.

.

Dimensione delle reti SOM

Si è deciso di verificare se alcune delle dimensioni delle reti SOM fossero selezionate con maggiore frequenza, si sono quindi considerate le dimensioni delle reti che hanno presentato i valori migliori dei due indicatori per ogni ripetizione della

clusterizzazione.

intero	5x5/6x6	6x6	4x4/5x5	5x5/4x4
vita sessuale	8x8	5x5	8x8	8x8
vita sessuale + sintomi	5x5	6x6	7x7	5x5/7x7
soluz. GA	6x6	7x7/4x4	4x4	4x4

Figura 4.4: Tabella che riporta la dimensione delle reti selezionate per i due indicatori della clusterizzazione, qualora la rete fosse la stessa è stata riportata una sola volta.

.

Dai risultati (riportati in Figura 4.4) si ottengono i seguenti risultati:

- Rete SOM 4x4 = selezionata nel 21,875% dei casi;
- Rete SOM 5x5 = selezionata nel 25% dei casi;
- Rete SOM 6x6 = selezionata nel 21,875% dei casi;
- Rete SOM 7x7 = selezionata nel 12,5% dei casi;
- Rete SOM 8x8 = selezionata nel 18,75% dei casi.

La dimensione delle reti quindi è ininfluente, con il dataset attuale, in quanto le percentuali sono di selezione sono vicine alla probabilità iniziale (20%, in quanto sono 5 reti).

Tipologia dei tagli ottenuti

La tipologia dei tagli del dendrogramma ottenuti in fase di clusterizzazione (Figura 4.5) indicano tutti i tagli del dendrogramma che indicano o il valore migliore di uno dei due parametri dell'intera prova, oppure il valore migliore di uno degli indicatori della clusterizzazione. Essi presentano un'ampia maggioranza di tagli a 2, aspetto positivo considerando che l'obiettivo principale, in questa fase del progetto, è la distinzione tra pazienti sani e pazienti affetti da MST (se si includono anche i tagli a 3 del dendrogramma, si raggiunge circa l'89% dei tagli).

Distribuzione delle singole patologie

Riguardo all'efficacia delle patologie, la Gonorrea risulta la migliore, ma il numero di distribuzioni evidenziate viene ridimensionato rispetto al grafico ottenuto al Paragrafo 3.2.3, probabilmente a causa dello scarso numero di prove eseguite.

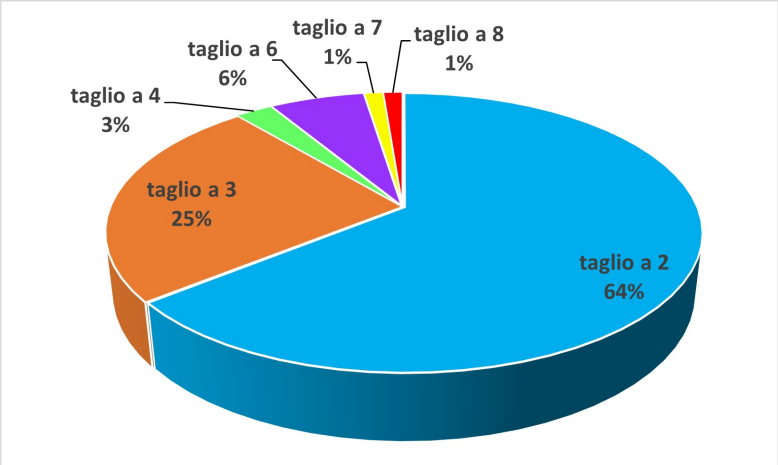


Figura 4.5: Grafico a torta che mostra la distribuzione dei tagli del dendrogramma nelle 4 prove per ognuna delle 5 ripetizioni.

.

Seguono Clamidia, Sifilide, Condiloma ano-genitale, la classe "Altro", ed Herpes genitale.

Collegamenti tra singole patologie

Al fine di verificare la possibile corrispondenza tra le distribuzioni delle singole patologie nei *clusters*, si è deciso di operare solo sui clusters selezionati mediante i valori dei parametri in fase di clusterizzazione, verificando che la coppia di patologie fossero tra quelle selezionate e con una distribuzione simile, valutata in percentuale in modo da considerare la diversa numerosità nel dataset di partenza. Si è deciso però di considerare le patologie principali: Clamidia, Condiloma ano-genitale, Gonorrea, Herpes genitale, Sifilide.
Si è quindi deciso di riportare (in Figura 4.6) le patologie che, globalmente, hanno presentato, tra i *clusters* adatti a questa analisi, una distribuzione simile almeno nel 50% dei casi. Di questi, le distribuzioni più correlate sono relative a Sifilide e Herpes genitale, Condiloma e Gonorrea ma, considerato che i *clusters* adatti a questa analisi variano tra il 19% e il 37%, questo risultato non può essere ritenuto rilevante.

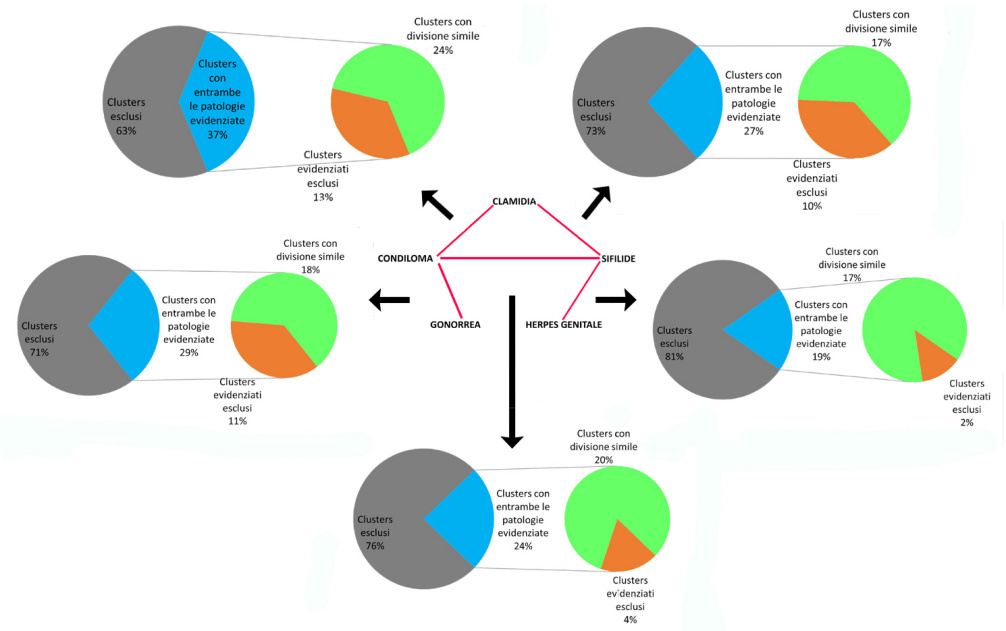


Figura 4.6: Grafici a torta che mostrano i collegamenti individuati tra le singole patologie.

.

Confronto tra le features selezionate e i risultati della ricerca bibliografica

Sesso	Clamidia	Condiloma ano-genitale	Gonorrea		Sifilide	
Etnia						
Gravidanza	Clamidia	Condiloma ano-genitale	Gonorrea	Herpes genitale		
Data dell'ultimo Pap Test		Condiloma ano-genitale				
Ultimo test per la Epatite B						
Ultimo test per la Clamidia						
Vaccinazione per HPV		Condiloma ano-genitale				
MST passate - Epatite B						
MST passate - Epatite C						
MST passate - Epatite A						
Tipi di rapporti			Gonorrea		Sifilide	
Vita promiscua del partner	Clamidia		Gonorrea		Sifilide	
Masturbazione	Clamidia	Condiloma ano-genitale	Gonorrea	Herpes genitale	Sifilide	
Sex toys	Clamidia	Condiloma ano-genitale	Gonorrea	Herpes genitale	Sifilide	HIV
Ruolo in rapporto orale	Clamidia	Condiloma ano-genitale	Gonorrea	Herpes genitale	Sifilide	
Ruolo in rapporto vaginale	Clamidia	Condiloma ano-genitale	Gonorrea	Herpes genitale	Sifilide	
Sintomo - eruzione cutanea				Herpes genitale	Sifilide	
Sintomo - linfonodi				Herpes genitale	Sifilide	HIV
Sintomo - congiuntivite	Clamidia		Gonorrea			
Sintomo - febbre	Clamidia		Gonorrea	Herpes genitale	Sifilide	HIV

Figura 4.7: Risultato del confronto tra la soluzione ottenuta con la Feature Selection e la ricerca bibliografica.

.

In Figura 4.7 viene proposto un dettaglio della soluzione ottenuta con l'Algoritmo Genetico, nel quale viene indicato per ogni *feature*, quale/i patologia/e essa permette di distinguere correttamente, in base alla ricerca bibliografica precedentemente svolta.
Tra 20 *features* selezionate, solo 6 (il 30%) non risultano utili per la classificazione di alcuna patologia, un risultato positivo considerando l'esiguo dataset di partenza, in particolare se si osserva la rilevanza delle restanti *features* per ognuna delle patologie:

- 50% per la Gonorrea e la Sifilide;
- 45% per la Clamidia;
- 40% per il Condiloma ano-genitale e l'Herpes genitale.

La *Feature Selection* ha quindi prodotto un risultato accettabile, in grado di distinguere le patologie con un'accuratezza simile, sebbene a livello teorico.

Capitolo 5

Conclusioni e sviluppi futuri

L'obiettivo di questo lavoro di tesi è stato sviluppare un classificatore per una iniziale distinzione di pazienti sani e pazienti affetti da IST/MST, in grado di fornire una base preliminare per lo sviluppo dell'algoritmo definitivo di riconoscimento del rischio di aver contratto una IST/MST. Il lavoro è stato svolto utilizzando dati simulati dai medici e/o specializzandi, e ottenuti come risposte alle domande del questionario fornito dal team Bridge.
I risultati iniziali ottenuti con un metodo di clusterizzazione sono ritenuti accettabili, considerata la dimensione del dataset di partenza (314 pazienti/dati simulati, costituiti ciascuno da un totale di 66 features/informazioni per ogni questionario): infatti, mediante l'utilizzo delle reti neurali SOM, si sono ottenute clusterizzazioni in grado di riconoscere un buon numero di pazienti sani. Questo risultato è stato percettibilmente migliorato a seguito della *Feature Selection*, mediante l'implementazione dell'Algoritmo Genetico, che ha permesso di ridurre le *features* da 66 a 20 (valore minimo scelto): ciò dimostra che il questionario di partenza conteneva informazioni ridondanti o poco significative per il problema di classificazione.
Il confronto tra le *features* quindi selezionate e una ricerca bibliografica ha permesso di verificare una buona distribuzione delle *features* relative alle 5 patologie considerate (ovvero patologie più diffuse e sufficientemente rappresentate nel dataset utilizzato): Clamidia, Condiloma ano-genitale, Gonorrea, Herpes genitale, Sifilide.
In futuro, il presente lavoro potrebbe essere ampliato con l'utilizzo di un metodo di classificazione supervisionato, in modo da valutare l'efficacia della *Feature Selection* ottenuta su un dataset più esteso di quello utilizzato nel presente lavoro. Inoltre, si verificherà l'efficacia del metodo citato al Paragrafo 3.3.1, (*A Tribe Competition-Based Genetic Algorithm for Feature Selection in Pattern Classification* [19]), il quale potrebbe risultare utile per l'obiettivo successivo del progetto, ovvero l'analisi del rischio di aver contratto una IST/MST.
Un aspetto importante sarà il passaggio dai dati simulati da medici e/o specializzandi ai dati ottenuti da pazienti che si sono presentati nei consultori delle 13 regioni

italiane che collaborano, ad oggi, con il team Bridge: infatti, potrebbe variare l'importanza delle *features* iniziali e, di conseguenza, il risultato della *Feature Selection*. Dall'analisi delle risposte attualmente ricevute da medici e/o specializzandi, infatti, è stata evidenziata l'importanza della domanda relativa ai sintomi manifestati dal paziente, le cui risposte potrebbero risultare meno chiare nel caso dei pazienti "reali".

Bibliografia

[1] Maria Giovanna Longo. «Progettazione e sviluppo di una Mobile App per affrontare le Malattie Sessualmente Trasmissibili». Tesi di laurea mag. Politecnico di Torino, Marzo 2021 (cit. alle pp. 4–7, 9–14, 16, 17, 22, 23).

[2] Istituto Superiore di Sanità ISS - Servizio Comunicazione Scientifica. «Notiziario ISS "Le Infezioni sessualmente trasmesse: aggiornamento dei dati dei due Sistemi di sorveglianza sentinella attivi in Italia al 31 dicembre 2018"». In: 33 n.7/8 (Luglio-Agosto 2020) (cit. alle pp. 4, 16).

[3] *Infezione da virus dell'immunodeficienza umana (HIV) - Malattie infettive*. Manuali MSD Edizione Professionisti. URL: https://www.msdmanuals.com/professionale/malattie-infettive/virus-dell-immunodeficienza-umana-hiv/infezione-da-virus-dell-immunodeficienza-umana-hiv?query=HIV (visitato il 31/07/2021) (cit. alle pp. 5, 6).

[4] AUSL Modena-HelpAIDS- www.helpaids.it. *Falsi Miti: non si può contrarre l'HIV da...* www.helpaids.it. URL: https://www.helpaids.it/flex/cm/pages/ServeBLOB.php/L/IT/IDPagina/391 (visitato il 13/04/2021) (cit. a p. 6).

[5] *Condilomi*. ISSalute. URL: https://www.issalute.it/index.php/la-salute-dalla-a-alla-z-menu/c/condilomi (visitato il 15/04/2021) (cit. a p. 6).

[6] «Condyloma Acuminatum (Genital Warts): Background, Pathophysiology, Epidemiology». In: (20 lug. 2021). Publication: Medscape - eMedicine. URL: https://emedicine.medscape.com/article/781735-overview (visitato il 21/04/2021) (cit. a p. 7).

[7] *Sintomi Condilomi*. URL: https://www.my-personaltrainer.it/Sintomi/Condilomi (visitato il 23/04/2021) (cit. a p. 8).

[8] *Sifilide - Infezioni*. Manuale MSD, versione per i pazienti. URL: https://www.msdmanuals.com/casa/infezioni/malattie-a-trasmissione-sessuale-mts/sifilide (visitato il 20/03/2021) (cit. alle pp. 8–10).

[9] European Centre for Disease Prevention e Control. «Syphilis. In: ECDC. Annual epidemiological report for 2017». In: (2019) (cit. a p. 9).

[10] EpiCentro. *Infezioni sessualmente trasmesse*. URL: https://www.epicentro.iss.it/ist/ (visitato il 25/04/2021) (cit. a p. 10).

[11] *Herpes genitale - Infezioni*. Manuale MSD, versione per i pazienti. URL: https://www.msdmanuals.com/it-it/casa/infezioni/infezioni-da-herpesvirus/herpes-genitale?query=herpes%5C%20genitale (visitato il 20/03/2021) (cit. alle pp. 11, 12).

[12] EpiCentro. *Herpes genitale, trasmissione, decorso, diagnosi, trattamento e prevenzione - EpiCentro*. URL: https://www.epicentro.iss.it/herpes-genitale/ (visitato il 20/04/2021) (cit. a p. 11).

[13] *Infezioni da clamidia - Infezioni*. Manuale MSD, versione per i pazienti. URL: https://www.msdmanuals.com/it-it/casa/infezioni/infezioni-da-clamidia-e-da-micoplasma/infezioni-da-clamidia (visitato il 20/03/2021) (cit. alle pp. 12–14).

[14] R. Chandran. «STD and contraception». In: *Malaysian Journal of Reproductive Health: A Publication of the Reproductive Research Centre of the National Population and Family Development Board, Malaysia* 11.1 (giu. 1993), pp. 1–7. ISSN: 0127-3213 (cit. a p. 14).

[15] Cesare Gentili. *Clamidia, sintomi e prevenzione di un'infezione silenziosa e pericolosa - Paginemediche*. 5 Giu. 2018. URL: https://www.paginemediche.it/medicina-e-prevenzione/disturbi-e-malattie/clamidia-sintomi-e-prevenzione-di-un-infezione-silenziosa-e-pericolosa (visitato il 16/08/2021) (cit. a p. 14).

[16] *Gonorrhea - Infezioni - Manuale MSD, versione per i pazienti*. URL: https://www.msdmanuals.com/it-it/casa/infezioni/malattie-a-trasmissione-sessuale-mts/gonorrhea?redirectid=3688 (visitato il 20/03/2021) (cit. alle pp. 15–17).

[17] *Gonorrea | SIAMS*. URL: https://www.siams.info/gonorrea/ (visitato il 17/08/2021) (cit. a p. 17).

[18] Benteng Ma e Yong Xia. «A tribe competition-based genetic algorithm for feature selection in pattern classification». In: *Applied Soft Computing* 58 (set. 2017), pp. 328–338. ISSN: 15684946. DOI: 10.1016/j.asoc.2017.04.042. URL: https://linkinghub.elsevier.com/retrieve/pii/S1568494617302247 (visitato il 08/04/2021) (cit. a p. 55).

[19] Zhongheng Zhang, Victor Trevino, Sayed Shahabuddin Hoseini, Smaranda Belciug, Arumugam Manivanna Boopathi, Ping Zhang, Florin Gorunescu, Velappan Subha e Songshi Dai. «Variable selection in Logistic regression model with genetic algorithm». In: *Annals of Translational Medicine* 6.3 (feb. 2018), pp. 45–45. ISSN: 23055839, 23055847. DOI: 10.21037/atm.2018.01.15. URL: http://atm.amegroups.com/article/view/18292/18559 (visitato il 08/04/2021) (cit. alle pp. 55, 101).

[20] Babatunde Oluleye, Armstrong Leisa, Jinsong Leng e Diepeveen Dean. «A Genetic Algorithm-Based Feature Selection». In: 5.4 (2014), p. 7 (cit. alle pp. 55, 60).