Analisi delle patologie cardiache

Advanced Statistical Modelling For Big Data
Andrea Mauro [mat.0222400914]

Introduzione

Nel seguente progetto viene preso in analisi il fenomeno relativo alla cardiopatia. In particolare, il principale obiettivo del lavoro proposto è la ricerca dei fattori di rischio dominanti cui determinano l'aumento delle probabilità di contrarre malattie caridache, dunque in questo contesto saranno utilizzati strumenti di analisi esplorativa per catturare le peculiarità del fenomeno. Inoltre verrà proposta una sezione di modellistica, dove sarà stimato un modello in grado di cogliere e riprodurre le dinamche del fenomeno di interesse. Nell'ultima parte del lavoro verrà quindi valuta la capacità previsiva del modello stimato.

1 Dati

Le informazioni presenti nel dataset sono state raccolte attraverso un sondaggio sottoposto ad un campione di circa 320,000 persone. Nel dataset sono registrate 18 caratteristiche che riguardano generalità, abitudini, stile di vita e la presenza o meno di specifiche malattie. Prima di procedere con le analisi delle caratteristiche del set di dati, la fase di *Data Wrangling* si rende necessaria in quanto il set originale si presenta in forma grezza, dunque sono state formattate le tipologie delle variabili presenti nel campione, in particolare la formattazione utilizzata è la seguente:

- *Heart Disease*: variabile categoriale nominale, rappresenta la presenza o meno di una patologia cardiaca;
- *BMI*: variabile numerica, rappresenta l'indice di Massa Corporea;
- *Smoking*: variabile categoriale nominale, indica se il soggetto è fumatore;
- *Alcol Drinking*: variabile categoriale nominale, indica se il soggetto consuma regolarmente bevande alcoliche;
- *Stoke*: variabile categoriale nominale, indica se il soggetto ha subito ictus;
- *Physical Health*: variabile numerica, rappresenta quanti giorni sui 30 precedenti (rispetto a quando viene sottoposto il sondaggio) il soggetto non ha goduto di buona salute fisica;
- *Mental Health*: variabile numerica, rappresenta quanti giorni sui 30 precedenti (rispetto a quando viene sottoposto il sondaggio) il soggetto non ha goduto di buona salute mentale;
- *Diff Walking*: variabile categoriale nominale, indica se il soggetto ha difficoltà a camminare e/o salire le scale;
- *Sex*: variabile categoriale nominale, rappresenta il sesso del soggetto;
- *Age Category*: variabile categoriale nominale, rappresenta la fascia d'età cui appartiene il soggetto;
- *Race*: variabile categorica nominale, rappresenta l'etnia del soggetto;
- *Diabetic*: variabile categorica nominale, indica se il soggetto è affetto da diabete;
- *Physical Activity*: variabile categorica nominale, indica se il soggetto svolge regolarmente attività fisica;

- *GenHealth*: variabile categorica nominale, indica l'autovalutazione complessiva dello stato di salute del soggetto;
- *Sleep Time*: variabile numerica, rappresenta le ore di sonno al giorno (mediamente);
- *Asthma*: variabile categorica nominale, indica se il soggetto soffre di asma;
- *Kidney Disease*: variabile categorica nominale, indica se il soggetto è affetto da malattie renali;
- *Skin Cancer*: variabile categorica nominale, indica se il soggetto è affetto da nepolasia;

In seguito alla formattazione delle feature, vengono rimosse le osservazioni ridondanti e gli oggetti che presentano alcune caratteristiche non disponibili.

2 Analisi Esplorativa

In questa fase vengono approfondite le caratteristiche della varibile *Cardiopatia (Heart Disease)*, esplorando la struttura e le relazioni con altre feature presenti nel campione. Chiaramente questa fase è fondamentale per comprendere le dinamiche della Cardiopatia e fornire ottimi spunti informativi da cui partire per la successiva modellazione del fenomeno attraverso gli strumenti presenti nella letteratura statistica.
Il primo passo essenziale per comprendere il fenomeno della cardiopatia consiste nell'osservare e analizzare la struttura della variabile *Cardiopatia* all'interno del campione. Come già menzionato, questa caratteristica è definita mediante una variabile dicotomica, in cui il valore 0 denota l'assenza e il valore 1 indica la presenza di patologie cardiache. Di seguito viene mostrata la figura rappresentante la frequenza relativa della variabile *Cardiopatia*:

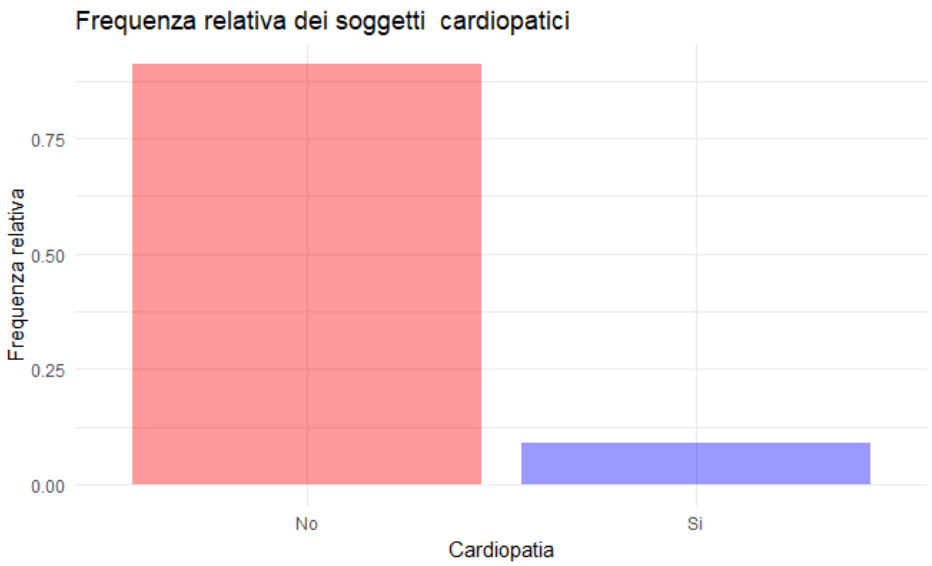

Figure 1: Freq. relativa *Cardiopatia*

Dal grafico si osserva chiaramente come la distribuzione delle classi della feature *Cardiopatia* risultino chiaramente sblianciate, difatti nel campione si registrano il 92% delle osservazioni relative a soggetti non cardiopatici. Dunque preso atto del netto sblianciamento delle classi, tale evidenza non può essere ignorata in fase di modellazione, in quanto ciò implicherebbe problemi di bias del modello, avendo così una visione distorta del fenomeno che si sta modellando.

La Figura 2 mette in evidenza il box-plot della variabile *Indice di Massa Corporea* condizionato per la variabile *Cardiopatia*:

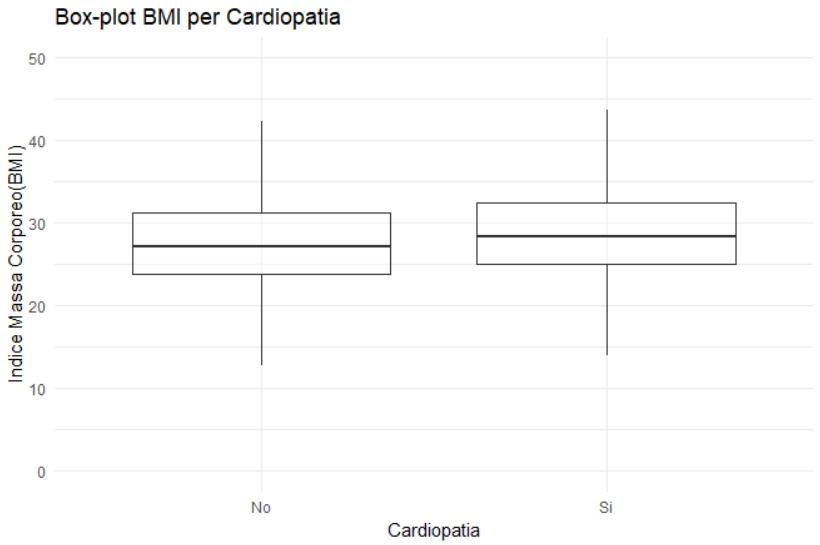

Figure 2: Box-plot *BMI* condizionato per la *Cardiopatia*

I due box-plot risultano essere molto simili, difatti analizzando le distribuzioni dell'indice di massa corporea condizionate su soggetti con e senza malattie cardiache, notiamo come i valori mediani risultano approssimativamente simili. Tuttavia, la distribuzione condizionata per i pazienti cardiopatici mostra un valore mediano leggermente superiore. Inoltre, osservando la distribuzione del BMI tra i soggetti con malattie cardiache notiamo una leggera asimmetria positiva.
Dalla figura 3, osserviamo come circa il 35% dei soggetti cardiopatici presenti nel campione non svolge alcun tipo di Attività fisica, mentre i soggetti che svolgono attività fisica non affetti da patologie cardiache risultano essere circa l'80%.

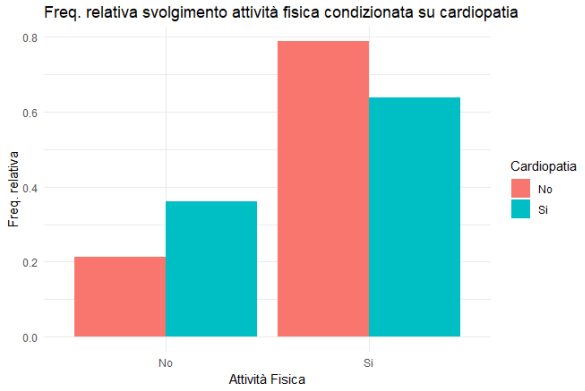

Figure 3: Freq. relativa *Attività Fisica condizionata su Cardiopatia*

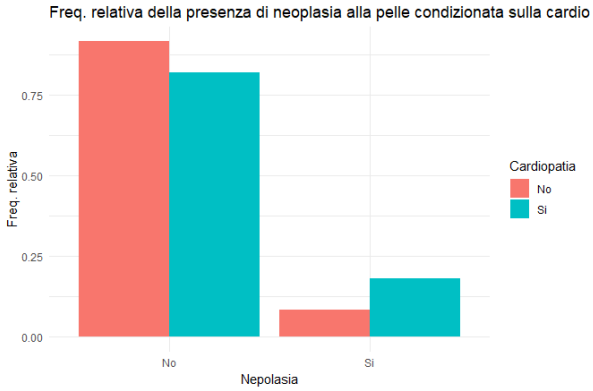

Figure 4: Freq. relativa *Nepolasia* condizionata su Cardiopatia

In generale, non emergono differenze significative nello svolgimento di attività fisica tra i soggetti affetti da cardiopatia e quelli non affetti, tuttavia approfondendo il discorso potrebbero sussistere differenze sul tipo di attività fisca svolta dai caridopatici, ai quali viene fortemente sconsigliato lo svolgimento di attività sportive intense.
Dall'analisi della figura 4, emerge che circa il 15% dei soggetti cardiopatici nel campione presenta una neoplasia cutanea. Inoltre, i soggetti cardiopatici affetti da neoplasia cutanea risultano essere il doppio rispetto ai soggetti non cardiopatici affetti da neoplasia.

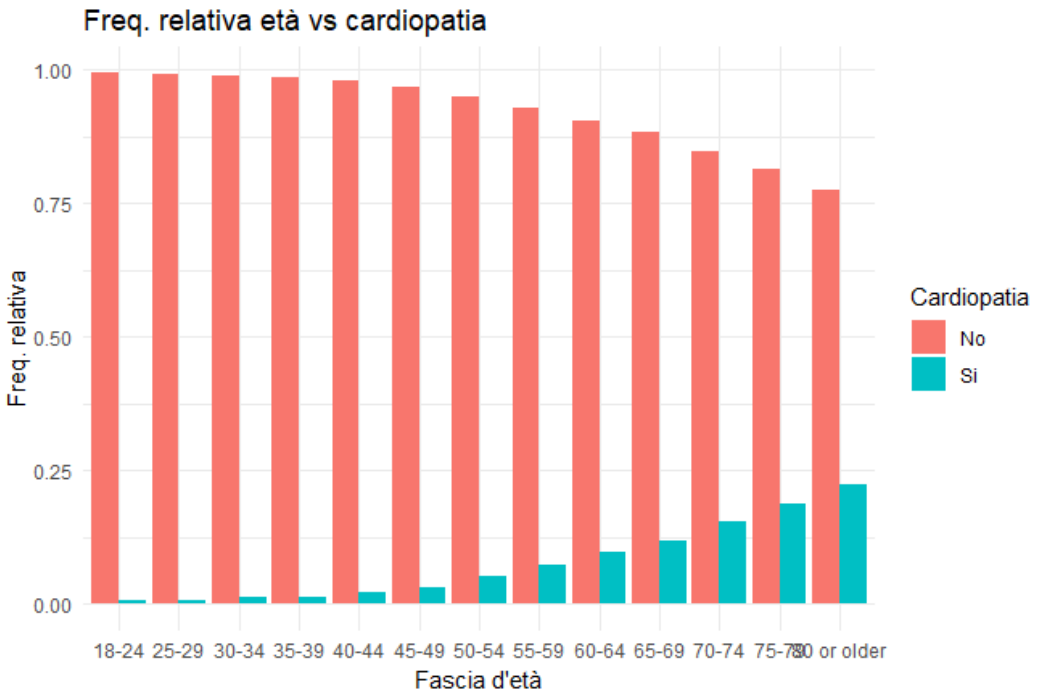

Figure 5: Freq. relativa *Cardiopatici per ogni fascia d'età*

Dall'analisi della Figura 5, emerge chiaramente un incremento del numero di cardiopatici nelle età più avanzate. Infatti, nelle fasce d'età comprese tra i 18 e i 39 anni, la percentuale di individui con malattie cardiache è approssimativamente del 1%, mentre si registra il picco massimo di casi nella fascia d'età di 80 anni e oltre.
Attraverso la matrice delle correlazioni (Figura 6), viene fornita una misura dell'intensità della relazione lineare che intercorre tra le variabili numeriche.

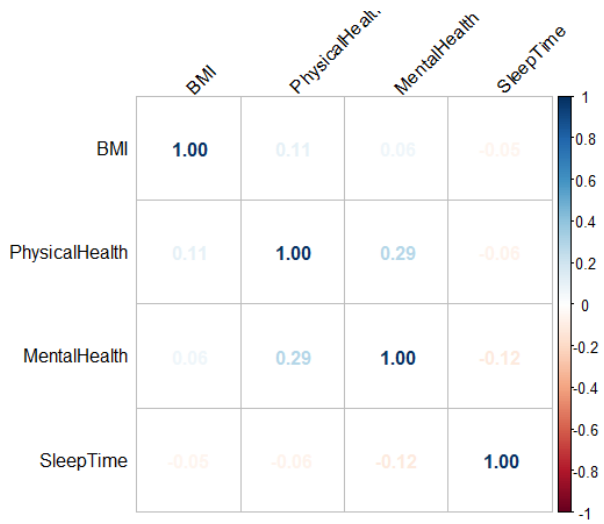

Figure 6: Matrice delle correlazioni

Dalla Figura 6 emerge chiaramente l'assenza di particolari relazioni lineari, le quali potrebbero apportare problemi in fase di stima del modello.

3 Modello

In base alle caratteristiche emerse nella fase precedente e considerando la natura dicotomica della variabile "Cardiopatia", nel contesto della modellazione del fenomeno si può ragionevolmente ipotizzare che segua approssimativamente una distribuzione di Bernoulli. In questo tipo di distribuzione, il fenomeno può assumere esclusivamente due valori: 1, indicante la presenza di cardiopatia, e 0, indicante l'assenza di cardiopatia.

In questo contesto, per la modellazione del fenomeno di interesse tramite modelli lineari generalizzati viene naturale adottatare un modello la cui componente stocastica si distribuisce come una Bernoulli, con funzione di collegamento logit.
Inoltre, considerando il problema dello sbilanciamento delle classi, dove è stato osservanto che il 92% delle istanze nel campione riguardano soggetti non cardiopatici, si è proceduto alla stima di due modelli distinti: una regressione logistica standard e una regressione logistica ponderata.
La determinazione dei pesi è stata effettuata considerando la proporzione dello sbilanciamento stesso. Dunque, sono stati assegnati pesi pari a 0.92 alle osservazioni relative a soggetti cardiopatici e 0.08 alle osservazioni relative a soggetti non affetti da cardiopatia.

Dopo aver convertito le diverse *variabili categoriali* presenti nel dataset in *variabili dummy*, il totale dei regressori disponibili nella fase di modellazione è risultato essere 38. Per la selezione delle variabili da incorporare nella regressione, abbiamo inizialmente adottato un approccio di stima del modello che comprendeva tutti i regressori a disposizione. Successivamente, abbiamo escluso le variabili i cui coefficienti non raggiungevano un livello di significatività pari a 0.01 (ossia, p-value > 0.01), generando così il modello ridotto. Per ciascuna regressione è stato svliuppato il test del rapporto di verosimiglianza(LRT), per verificare l'assunzione di equivalenza tra il modello completo e ridotto.

Analizzando la Figura 7, notiamo come la regressione logistica che include tutti i predittori e quella ridotta (esclude la variabile "Attività Fisica") secondo il test del rapporto di verosimiglianza risultano essere equivalenti, difatti non rifiutiamo l'ipotesi nulla per un valore di α pari a 0.01. Dunque per il principio di parsimonia le successive analisi verteranno sulla versione ridotta del modello logistico.

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	223819	101645			
2	223818	101640	1	5.5326	0.01867

Figure 7: Regressione logistica - Test LRT: modello ridotto e completo

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	223819	33477			
2	223818	33475	1	1.3555	0.2443

Figure 8: Regressione logistica pesata - Test LRT : modello ridotto e completo

Dalla figura 8, osserviamo il risultato del LRT sviluppato sul modello logistico pesato e il modello logistico pesato ridotto, questa mostra il non rifiuto dell'ipotesi nulla di equivalenza tra il modello completo e quello ridotto. Dunque per il principio di parsimonia le successive analisi verteranno sulla versione ridotta del modello logistico pesato.
In un contesto dove non si conoscono a pieno le dinamiche del fenomeno, può essere utile valutare e decretare il modello più adatto al problema prendendo in esame la bontà di adattamento dei modelli e valutarla attraverso particolari criteri di informazione. Nello specifico vengono considerati i criteri d'informazione Akaike (AIC) e Bayesian (BIC).

Model	**AIC**	**BIC**
logistico pesato -Rid	16959	17339
logistico -Rid	97554	97934

Table 1: Bontà di Adattamento

Orientando la selezione del modello più appropriato per descrivere il fenomeno di interesse attraverso i criteri di informazione AIC e BIC, emerge come la preferenza sia verso il modello logistico ponderato,

difatti risulta evidente come la bontà di adattamento del modello logistico pesato risulti nettamente migliore rispetto la logistico non pesato.
Di seguito vengono mostrati alcuni dei coefficienti associati alle variabili presenti tra i regressori del modello logistico pesato:

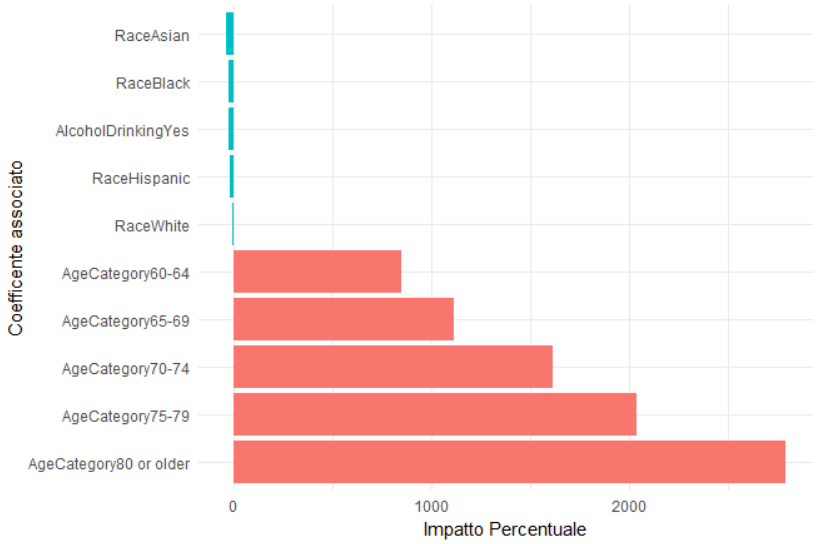

Figure 9: Impatto coefficienti regressione logistica pesata ridotta

La Figura 9 rappresenta una sintesi dei coefficienti più influenti, sia negativi che positivi, per la regressione logistica pesata. Questa rappresentazione è basata sulla variazione percentuale dell'odds in favore della cardipatia. In particolare, abbiamo considerato la trasformazione $(\exp(\beta_j) - 1) \times 100$, interpretabile come la variazione percentuale stimata sull'odds a favore di Y=1 (Cardiopatia), ciò avviene quando la variabile X_j viene aumentata di un'unità, se continua, o quando passa da 0 a 1 se binaria (ceteris paribus).
Osservando la figura, emerge chiaramente l'ampio impatto della fascia di età del soggetto sull'odds a favore della Cardiopatia, difatti i cinque coefficienti più significativi, che incidono sull'odds in modo positivo, sono tutti associati alle variabili legate all'età.
Al contrario, i coefficienti associati alle variabili etniche mostrano variazioni percentuali negative sull'odds. Tuttavia, è importante notare che tali variazioni sono prossime allo zero, e dunque non esercitano un impatto significativo sull'odds a favore della presenza di cardiopatia.

Capacità predittive
Per completezza viene quindi valutata la capacità previsiva dei due modelli considerati sin ora. In un momento antecedente alla fase di stima, il set di dati disponibili è stato suddiviso per valutare il comportamento dei due modelli su dati non osservati, in particolare la divisione del dataset è avvenuta come segue:

- il 75% delle osservazioni sono state destinate al *Train*
- il 25% delle osservaizone sono state destinate al *Test*

Così facendo, ottenuta la stima dei modelli sul train, la relativa capacità previsiva dei modelli viene valutata sul test.

Modelli	**Accuracy**	**Balanced Accuracy**	**Sensitivity**	**Specificity**
Reg. logistica -rid	0.9157	0.5476	0.1034	0.9918
Reg. logistica pesata -Rid	0.7359	0.7641	0.7980	0.7301

Table 2: Performance previsive

I due modelli stimati offrono perfromance chiaramente differenti, la regressione logistica riesce bene a individuare i soggetti che non risultano affetti da malattie cardiache e al contrario riesce a indentificare solo il 10% dei soggetti cardiopatici. Chiaramente tale performance è il risultato del netto sbilanciamento presente nelle classi delle variabile dipendente. In un contesto dove si vuole identificare un soggetto affetto da malattie cardiache, tale risultato non può essere ritenuto soddisfaciente.
Al contrario migliori performance si registrano con il modello logistico ponderato, dove nel 79% dei casi il modello è stato in grado di individuare il soggetto affetto da cardiopatia, rappresentando un notevole miglioramento rispetto al modello non ponderato.

Data le migliori capacità previsive, andiamo quindi a calcolare gli effetti marginali sfruttando le stime dei coefficienti ottenute con il modello logistico pesato.

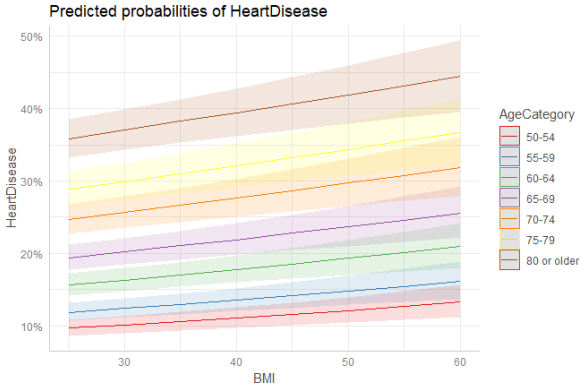

Figure 10: Effetti marginali dei coefficienti delle Fasce d'età e BMI

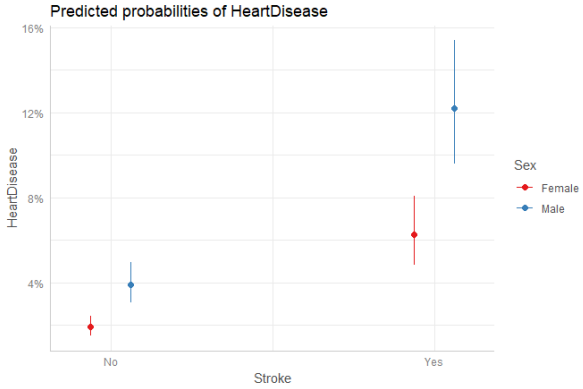

Figure 11: Effetti marginali dei coefficienti del Genere e Stroke

Per una questione di sintesi vengono mostrati in figura solo alcuni dei possibili effetti marginali calcolati con i coefficienti del modello logistico pesato ridotto. In particolare, seguendo le indicazioni fornite dalla Figura 9, sono stati calcolati gli effetti marginali dell'Indice di Massa Corporea (BMI) e della fascia d'età del soggetto sulla probabilità di avere una patologia cardiaca (fare riferimento alla Figura 10). Osserviamo come a parità di BMI la fascia d'età risulta essere determinante sulla probabilità di essere affetti dalla cardiopatia. D'altro canto anche se consideramo la medesima fascia d'età per BMI diversi, maggiore è l'indice di massa corporeo maggiore risulta essere la probabilità di essere affetti da cardiopatia.
Continuando con l'analisi degli effetti marginali, nella Figura 11, viene mostrata l'incidenza del genere e aver subito un Ictus sulla probabilità di riscontrare patologie cardiache. In particolare si registrano per i soggetti cui hanno subito un ictus probabilità più alte di cardiopatia, nello specifico per gli uomini, la probabilità risulta essere il triplo rispetto ai soggetti maschili che non hanno mai sofferto di ictus.

4 Conclusione

In questo progetto, sono state approfondite le peculiarità e i principali fattori sulla probabilità di riscontrarre malattie cardiache, usando informazioni incentrate sul benessere e stile di vita dei soggetti, mediante l'utilizzo di strumenti di analisi esplorativa e la modellizzazione della probabilità del fenomeno attraverso il modello logistico appartenente alla famiglia dei GLM. Attraverso tali strumenti, abbiamo identificato alcune delle sfumature e fattori che influenzano la probabilità di essere affetti da malattie cardiache.