UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Scienze Economiche e Statistiche

Corso di Laurea Triennale in Statistica per i Big Data

Tesi in

Modelli Statistici

ABITUDINI E PROFILO DEI PAZIENTI DIABETICI: UN'ANALISI STATISTICA

Relatore:

Ch.mo Prof. Marialuisa Restaino

Candidato:

Aniello De Palma

matr. 0212800936

ANNO ACCADEMICO 2022/2023

INDICE

INTRODUZIONE

Il diabete mellito è una patologia cronica sempre più diffusa a livello globale, con impatti significativi sulla salute e sulla qualità della vita dei pazienti. La presente tesi si propone di investigare le abitudini e i comportamenti dei pazienti diabetici al fine di comprendere meglio i fattori determinanti per una gestione efficace della malattia e per promuovere uno stile di vita sano. Attraverso un'analisi dettagliata delle abitudini alimentari, dell'attività fisica, dell'aderenza alle terapie e delle abitudini di monitoraggio della glicemia, si mira a identificare correlazioni significative tra tali aspetti e il controllo glicemico. Questo studio si basa su dati raccolti attraverso questionari strutturati e interviste condotte su un campione rappresentativo di pazienti diabetici. I risultati ottenuti saranno utili per la definizione di strategie personalizzate di gestione del diabete, nonché per informare l'educazione sanitaria e le politiche volte a migliorare la salute e il benessere dei pazienti affetti da questa patologia cronica.

Pertanto, nella prima parte della trattazione si pone l'accento sul fenomeno oggetto di studio, illustrandone la complessità e l'incidenza crescente nella società contemporanea. Vengono presentate le caratteristiche fisiopatologiche del diabete mellito, i suoi diversi tipi e le principali complicanze associate, al fine di fornire una panoramica approfondita della malattia. Inoltre, vengono esposti i fattori di rischio e le possibili cause che contribuiscono all'insorgenza del diabete, delineando così il contesto in cui si inserisce l'analisi delle abitudini dei pazienti diabetici. Questa parte introduttiva è cruciale per una chiara contestualizzazione del lavoro di ricerca e per una comprensione preliminare delle basi scientifiche su cui si fonderanno le successive analisi e valutazioni.

Nella seconda parte della trattazione, l'attenzione si concentra sulle metodologie statistiche impiegate per condurre un'analisi approfondita. Essendo le variabili in gioco principalmente di natura categorica, compresa la variabile target, si opta per l'impiego di modelli lineari generalizzati appositamente progettati per trattare questo tipo di dati. Questi modelli permettono di affrontare efficacemente la natura delle variabili categoriche, assicurando un'analisi accurata delle relazioni e delle associazioni presenti nel dataset. In particolare, vengono adottati modelli come la regressione logistica, il modello probit e il c-loglog, appositamente adattati per gestire le peculiarità delle variabili in questione. Questo approccio metodologico si rivela essenziale per ottenere risultati affidabili e informativi sull'influenza delle variabili considerate all'interno dello studio sulle abitudini dei pazienti diabetici.

Nell'ultima parte della trattazione, è stata condotta un'analisi esplorativa dei dati, che ha permesso di visualizzare le caratteristiche salienti delle variabili considerate e di comprendere la loro frequenza, associazione e particolarità. Successivamente, sono stati confrontati i tre modelli statistici differenti: l'obiettivo è stato valutare la bontà di adattamento di ciascun modello e la loro capacità predittiva. Per fare ciò, è stata eseguita una cross-validation leave-one-out (LOOCV) al fine di verificare la stabilità e l'efficacia delle previsioni ottenute. Tale analisi ha consentito di selezionare il modello più appropriato per interpretare e predire correttamente le relazioni all'interno del dataset, fornendo indicazioni fondamentali per una migliore comprensione delle abitudini dei pazienti diabetici e delle variabili che le influenzano.

CAPITOLO 1
IL DIABETE MELLITO

1.1 Il diabete e le principali classificazioni

Il diabete è una patologia di tipo cronico dovuta ad una disfunzione a carico dell'insulina, l'ormone secreto dal pancreas che consente l'utilizzo del glucosio come fonte di energia per l'organismo. Nel momento in cui questo meccanismo subisce un'alterazione, come accade nei pazienti diabetici, il glucosio si accumula e i suoi livelli nel sangue aumentano (iperglicemia). I principali fattori causanti il diabete sono: l'aumento dei casi di obesità, l'aumento dell'età media e dell'aspettativa di vita più sedentaria, l'aumento dello stress e soprattutto, la genetica.

L'organo colpito è il pancreas, in particolare la parte endocrina, *le isole di Langerhans*, che costituiscono non più dell'1% della massa pancreatica complessiva. Queste sono popolate da *cellule beta* secernenti insulina (60-80%), *cellule alfa* che producono glucagone (20-30%), *cellule delta* che rilasciano somatostatina (3-10%), la quale modula l'immissione in circolo di glucagone ed insulina, e le *cellule F*, molto rare (1-2%), secernenti polipeptide pancreatico quando s'ingeriscono cibi essenzialmente proteici.

Esistono principalmente due tipologie di diabete. Da una parte, il diabete di tipo I, la cui genesi ha origine autoimmune; dall'altra, il diabete di tipo II, che può insorgere tipicamente in età adulta e in chi ha una familiarità (parenti diabetici) con questa patologia. Non è escluso, però, che questa forma di diabete possa essere sviluppata anche da bambini che soffrono di una condizione di obesità.

I. Il *prediabete* è una condizione nella quale i livelli di glucosio nel sangue sono troppo elevati per essere considerati nella norma ma non tanto da essere classificati come diabete. Si parla di prediabete quando il livello di glucosio nel sangue a digiuno è compreso tra 100 mg/dl (5,6 mmol/l) e 125 mg/dl (6,9 mmol/l) oppure se il livello di glucosio nel sangue, 2 ore dopo un test di tolleranza al glucosio, è compreso tra 140 mg/dl (7,8 mmol/l) e 199 mg/dl (11,0 mmol/l). Il prediabete determina un rischio più alto di incorrere in diabete e malattie cardiache in futuro. Una perdita di peso del 5-10% attraverso la dieta e l'attività fisica consente di ridurre in modo significativo il rischio di sviluppare il diabete.

II. Il *diabete di tipo I* è una forma patologica più rara e, infatti, riguarda solo il 10% dei pazienti che soffrono di questa condizione. Generalmente, si manifesta nel corso dell'infanzia o in età adolescenziale. La disfunzione riguarda il pancreas, incapace di produrre insulina. I pazienti con diabete di tipo I, di conseguenza, necessitano della somministrazione di insulina per tutto il corso della loro vita. In rari casi, questa forma può colpire anche gli adulti: si parla, in tale eventualità, di *Late Autoimmune Diabetes in Adults (LADA).*

   I pazienti presentano nel sangue anticorpi che attaccano gli antigeni situati sulle cellule responsabili della produzione di insulina. La causa di questo comportamento anomalo del sistema immunitario è ignota, ma si ipotizza possa

essere legata a fattori genetici, ambientali o ad una particolare predisposizione del soggetto a reagire in un certo modo ad agenti esterni.

III. Il *diabete di tipo II*, invece, è la forma più comune e, infatti, riguarda nove pazienti diabetici su dieci. A differenza del diabete di tipo I, il pancreas produce insulina, ma l'organismo non è in grado di utilizzarla. L'insorgenza della malattia, solitamente, avviene in età adulta, dopo i 30 anni ed è favorita da diversi fattori di rischio quali la familiarità con il diabete, obesità o condizione di sovrappeso e uno stile di vita scorretto, caratterizzato da sedentarietà e scarso esercizio fisico.

Pur essendo tipico dell'età adulta, in rari casi il diabete di tipo II può colpire anche i più giovani. Si parla in questo caso di *Maturity Onset Diabetes of the Young (MODY)*, la cui genesi sembrerebbe essere legata a difetti di tipo genetico nei meccanismi di azione dell'insulina tra una cellula e l'altra.

Può accadere che per la diagnosi di diabete di tipo II trascorrano anche molti anni. La condizione di eccesso di glucosio nel sangue (iperglicemia) si sviluppa, infatti, in maniera graduale e, nella maggior parte dei casi, la scoperta di essere diabetici avviene casualmente o in modo collaterale (in seguito a interventi chirurgici o infezioni, ad esempio).

Siccome le probabilità di ammalarsi di diabete di tipo II crescono con l'aumentare dell'età, ma anche in condizioni di obesità o in caso di eccessiva sedentarietà e alimentazione scorretta, la prevenzione gioca un ruolo fondamentale nella lotta a questa patologia.

IV. Un'altra forma di diabete, meno conosciuta delle altre, è quella che colpisce le donne gravide. Si parla, in questo caso, di *diabete gestazionale (DMG)*.

Il DMG si sviluppa quando, in seguito ai cambiamenti ormonali determinanti dalla gravidanza, le cellule sono meno sensibili all'azione dell'insulina, con la conseguenza che i livelli di glicemia aumentano. Il diabete gestazionale può riguardare circa il 18% delle donne in stato di gravidanza e se non diagnosticato tempestivamente, può comportare complicanze severe come una crescita eccessiva del feto o una nascita prematura e finanche l'aborto.

In questi ultimi anni l'attenzione dei ricercatori si è focalizzata sui due ormoni pancreatici, deputati al controllo del tasso glicemico nel sangue: insulina e glucagone. I due ormoni hanno ruoli principali opposti: l'insulina abbassa i livello di glucosio nel sangue, mentre il glucagone li aumenta.

1.2 Cellula alfa e glucagone

Le cellule alfa costituiscono, insieme alle cellule beta, la principale componente endocrina delle isole pancreatiche. Le cellule alfa sintetizzano e rilasciano il glucagone. Tale ormone è il risultato finale di una serie di modifiche a carico del proglucagone, che rappresenta il reale polipeptide derivante dalla trascrizione del gene corrispondente.

Il glucagone è un ormone ad azione *"iperglicemizzante"* in quanto ha il compito principale di promuovere il rilascio di glucosio dal fegato nei periodo lontani dai pasti, evitando quindi che le concentrazioni di tale substrato energetico scendano al di sotto dei livelli normali. Numerosi fattori regolano la secrezione del glucagone. Tra questi, i più importanti sono gli amminoacidi (in particolare l'*arginina*), che inducono il rilascio dell'ormone, il glucosio e l'insulina, che invece ne inibiscono la secrezione.

Nei pazienti con diabete mellito di tipo II le concentrazioni circolanti di glucagone si sono rivelate più elevate di quanto atteso in presenza dei corrispondenti livelli glicemici, particolarmente a digiuno[1] (Baron et al., 1987). Questo contribuisce a far aumentare la produzione epatica di glucosio in tali soggetti. Inoltre, nel diabete di tipo II è stata documentata una ridotta capacità, da parte delle cellule alfa, di riconoscere in maniera congrua l'effetto inibitorio dell'iperglicemia. Qualche autore ha suggerito che le cellule alfa diabetiche potrebbero essere resistenti all'azione dell'insulina, e quindi non riconoscere l'effetto inibitorio[2]. Rispetto ai soggetti di controllo, nei pazienti con diabete di tipo II si assiste inoltre a un'aumentata secrezione di glucagone in seguito a stimolo con arginina. Per quanto riguarda la quantità delle cellule alfa nelle isole pancreatiche, vi è un aumento del volume delle cellule nel diabete di tipo II relativamente al volume pancreatico.

1.3 Cellula beta e insulina

Come accennato prima, le cellule beta rappresentano fino all'80% delle cellule insulari. Esse producono e secernono insulina in maniera controllata, in modo da mantenere le concentrazioni circolanti di glucosio nel loro intervallo fisiologico. La normale funzione beta-cellulare dipende essenzialmente dall'integrità dei meccanismi che regolano la sintesi e il rilascio dell'insulina. Il regolatore più importante della secrezione insulinica è proprio il glucosio, anche se numerosi altri nutrienti, così come vari ormoni, neurotrasmettitori e farmaci possono influenzare il rilascio dell'ormone. In risposta all'aumento dei livello di glucosio nel sangue, le cellule rilasciano insulina nel flusso sanguigno. Questa si lega ai recettori specifici presenti sulla superficie delle cellule bersaglio, come le cellule muscolari e del fegato. Questa interazione con i recettori attiva una serie di processi cellulari che consentono alle cellule di assorbire il glucosio dal sangue e di immagazzinarlo sotto forma di glicogeno o di utilizzarlo immediatamente per l'energia.

Le cellule beta e l'insulina svolgono un ruolo fondamentale nella regolazione dei livelli di zucchero nel sangue. Quando i livelli di glucosio sono troppo alti, le cellule beta aumentano la produzione di insulina nel sangue; quando i livelli di glucosio sono bassi, la produzione di insulina diminuisce. Quanto a lungo sopravviva una cellula beta nell'uomo non è noto, ma si comincia a pensare che la "speranza di vita" normale possa essere di alcuni anni (da 2 a 5). Pertanto, è ben noto come nei primi anni di vita la massa beta-cellulare aumenti notevolmente, grazie a marcati fenomeni replicativi e di neogenesi. Successivamente, si raggiunge una sorta di equilibrio, che viene poi mantenuto, di solito, durante la vita adulta.[3] Con l'avanzare dell'età, i fenomeni apoptotici tendono a prevalere su quelli rigenerativi, e la massa cellulare si riduce leggermente. In caso di necessità (ad es. riduzione della sensibilità all'insulina, gravidanza), le cellule beta sono in grado di adattarsi alle nuove circostanze. In particolare, in caso di sovrappeso, l'insulino-resistenza che ne deriva viene compensata da un accentuato tasso di replicazione e neogenesi.

---

[1] Baron AD, Schaeffer L, Shragg P, Kolterman OG. Role of hyperglucagonemia in maintainance of increased rates of hepatic glucose output in type II diabetics. Diabetes 1987; 36:274-83.

[2] Hamaguchi T, Fukushima H, Uehara M, Wada S, Shirotani T, Kishikawa H, et al. Abnormal glucagon response to arginine and its normalization in obese hyperinsulinemic patients with glucose intolerance: importance of insulin action onpancreatic alpha cells. Diabetologia 1991; 34:801-6.

[3] Rhodes CJ. Type 2 diabetes - a matter of beta-cell life and death? Science 2005; 307:380-4.

Nei pazienti con diabete mellito di tipo II, fattori genetici e acquisiti, concomitano nel determinare il danno di funzione e di massa della cellula beta.

La *glucotossicità* (vale a dire i danni indotti da elevate concentrazioni di glucosio) e la *lipotossicità* (vale a dire i danni dovuti ad alte concentrazioni di acidi grassi) sono, tra i vari fattori acquisiti, quelli che maggiormente sono stati studiati[4]. Numerosi studi hanno dimostrato che entrambe le condizioni inducono alterazioni della secrezione insulinica, aumentata apoptosi e interferenza con i processi rigenerativi beta-cellulari. In sintesi, anche se le cellule beta producono insulina, le cellule del corpo non rispondono adeguatamente all'ormone. Per compensare questa resistenza all'insulina, il pancreas può aumentare la produzione di insulina per cercare di far assorbire il glucosio dalle cellule. Tuttavia, nel tempo, la produzione di insulina dal pancreas può diminuire progressivamente a causa del carico eccessivo di lavoro.

Nel diabete di tipo I, invece, il sistema immunitario del corpo attacca erroneamente e distrugge le cellule beta nel pancreas. Questo processo autoimmune è responsabile della mancanza completa di insulina nel corpo. Di conseguenza, le persone con diabete di tipo 1 dipendono dalla somministrazione esterna di insulina tramite iniezioni multiple al giorno o l'uso di pompe per insulina. Le cellule beta sono gravemente danneggiate o distrutte, quindi la produzione di insulina endogena è ridotta o completamente assente.

1.4 Terapie di gestione del diabete

Nonostante il diabete mellito sia ancora oggi una malattia incurabile, alcune terapie permettono ai soggetti affetti di condurre uno stile di vita quanto più normale possibile. Infatti la scoperta dell'insulina, nel 1921, ha cambiato la storia della malattia diabetica, trasformandola da acuta e fatale, a cronica, in quanto la vita media di una persona di tipo I era di pochi mesi e la morte sopraggiungeva in seguito a chetoacidosi o infezioni.

Nei primi anni della sua scoperta, la terapia insulinica, anche se in grado di salvare la vita alle persone diabetiche, non era in grado di proteggerle dallo sviluppo delle complicanze diabetiche. Con la diabetologia moderna si comprese che l'obiettivo di un'adeguata terapia del diabete era di raggiungere livelli glicemici simili al soggetto non diabetico.

1.4.1 Terapia per il diabete mellito di tipo I

Come detto precedentemente, il diabete di tipo I (T1DM), è causato dalla distruzione totale delle cellule beta pancreatiche, da parte dei linfociti T, provocando un aumento incontrollato della glicemia. Dunque nei pazienti T1DM la terapia insulinica sostitutiva, mediante somministrazione esogena dell'ormone, diventa l'unica forma di trattamento per controllare le alterazioni metaboliche acute e la comparsa di complicanze d'organo periferiche. Il compito dell'insulina esogena è quello di simulare il più possibile, l'azione di quella prodotta dall'organismo, sia per quanto riguarda la concentrazione d'insulina basale che di quella rilasciata in acuto dopo i pasti, permettendo un normale utilizzo del glucosio da parte delle cellule sia a digiuno che dopo aver mangiato. Nei casi più severi di diabete di tipo I si ricorre al trapianto del pancreas in toto, con rilascio d'insulina endogena e normalizzazione sia dei valori della glicemia, che di altri importanti prodotti

4 Poitout V, Robertson RP. Minireview: secondary beta-cell failure in type 2 diabetes - a convergence of glucotoxicity and lipotoxicity. Endcorinology 2002; 143:339-42.

del metabolismo, definiti metaboliti intermedi. Per quei pazienti che non possono sottoporsi al trapianto, a causa per esempio di problemi cardiovascolari, un'alternativa è il trapianto d'isole pancreatiche, che consiste nel prelevare dal pancreas di un donatore le beta-cellule e trapiantarle nel paziente; dopo un processo di separazione e purificazione vengono impiantate, attraverso un'iniezione, nella *vena porta*, nel fegato, dove attecchiscono e iniziano a produrre insulina.

1.4.2 Terapia per il diabete mellito di tipo II

Per i pazienti T2DM si prevede, almeno inizialmente, il rispetto di una dieta e la pratica costante di attività fisica, poiché la maggior parte di loro è in sovrappeso o obeso. Se però la sola dieta e l'esercizio fisico, non sono sufficienti a riportare i valori di glicemia a livelli ottimali, i diabetici di tipo II vengono sottoposti a trattamento farmacologico, attraverso l'uso d'ipoglicemizzanti orali, che hanno l'effetto di ridurre la concentrazione di glucosio nel sangue, sia favorendone la captazione periferica e inibendo la gluconeogenesi, che stimolando la produzione d'insulina da parte del pancreas.

Tra i farmaci antidiabetici di ultima generazione, ci sono gli inibitori *alfa-glucosidasi*, che inibiscono l'attività dell'omonimo enzima, riducendo così il livello di glucosio nel sangue, e nello tesso tempo agiscono potenziando l'azione dell'ormone *glucagon-like-peptide* 1 (GLP-1)[5], stimolatore della sintesi d'insulina.

Questa molecola, rilasciata dalle cellule L intestinali[6], è in grado di potenziare la secrezione d'insulina glucosio-stimolata, aumentare la crescita e la sopravvivenza delle cellule β pancreatiche, riducendone l'apoptosi, così da aumentarne la massa complessiva, inibire il rilascio di glucagone da parte delle cellule α e rallentare lo svuotamento gastrico, riducendo l'assunzione di cibo. Dunque gli effetti antidiabetici di questo ormone, uniti al fatto che il GLP-1 si trovi ridotto nei pazienti con diabete di tipo 2, hanno incrementato l'interesse verso questa incretina come un possibile trattamento proprio per i pazienti colpiti da tale patologia.

1.5 Monitoraggi dei livelli di glucosio nel sangue

I livelli di glucosio nel sangue possono essere misurati facilmente a casa o altrove. Il test mediante *pungidito* è il più utilizzato per monitorare il glucosio nel sangue. La maggior parte degli apparecchi per monitorare il glucosio nel sangue (glucometri) utilizza una goccia di sangue prelevata dal polpastrello con una lancetta. La lancetta è costituita da un ago sottile che può essere conficcato nel dito o introdotto in un dispositivo a scatto che attraversa la pelle in modo semplice e rapido. In seguito, si depone una goccia di sangue su una striscia reattiva. La striscia contiene sostanze chimiche che subiscono alterazioni in base al livello di glucosio. Un glucometro rileva queste variazioni e riporta il risultato su un piccolo schermo digitale. Alcuni dispositivi consentono al campione di sangue di essere ottenuto da altri siti, come ad esempio il palmo, l'avambraccio, la parte superiore del braccio, la coscia o il polpaccio. I glucometri domestici sono più piccoli di un mazzo di carte.

[5] Drucker DJ. Glucagon-like peptides: regulators of cell proliferation, differentiation, and apoptosis. Mol Endocrinol. 2003 Feb;17(2):161-71. Review

[6] Gareth E. Lim and Patricia L. Brubaker. Glucagon-Like Peptide 1 Secretion by the L-Cell. The View from Within. Doi: 10.2337/db06-S020 Diabetes December 2006 vol. 55 no. Supplement 2 S70-S77.

I sistemi di monitoraggio continuo del glucosio (*Continuous Glucose Monitoring, CGM*) utilizzano un piccolo sensore per il glucosio posizionato sotto la pelle. Il sensore misura il livello di glucosio a intervalli di pochi minuti.

Ci sono due tipi di CGM, con scopi diversi:

- Professionali.
- Personali.

I CGM *professionali* raccolgono costantemente informazioni sulla glicemia in un determinato arco di tempo (da 72 ore a 14 giorni). Gli operatori sanitari utilizzano queste informazioni per formulare raccomandazioni terapeutiche. I CGM professionali non forniscono dati al soggetto diabetico.

I CGM *personali* sono utilizzati dal paziente e forniscono i dati sulla glicemia in tempo reale su un piccolo monitor portatile o su uno smartphone collegato. È possibile impostare il sistema CGM in modo da attivare un allarme se i livelli glicemici sono troppo bassi o troppo alti, in modo che il dispositivo possa aiutare i soggetti a identificare rapidamente variazioni preoccupanti del glucosio nel sangue.

I sistemi CGM sono particolarmente utili in determinate circostanze, come ad esempio nelle persone con diabete di tipo 1 che presentano variazioni rapide e frequenti del glucosio nel sangue (in particolare quando il glucosio raggiunge a volte livelli molto bassi) che sono difficili da identificare con il test mediante pungidito.

1.5.1 Emoglobina A1C

Il medico può monitorare il trattamento mediante un esame del sangue chiamato emoglobina A1C. Quando i livelli di glucosio nel sangue sono elevati, si verificano variazioni dell'emoglobina, la proteina che trasporta l'ossigeno nel sangue. Questi cambiamenti sono direttamente proporzionali ai livelli di glucosio nel sangue su un lungo periodo di tempo. Maggiore è il livello di emoglobina A1C, maggiore è stata la glicemia del soggetto. Pertanto, a differenza della misurazione del glucosio nel sangue, che rivela un valore puntuale, la misurazione dell'emoglobina A1C dimostra se i livelli di glucosio nel sangue siano stati controllati in modo adeguato nei mesi precedenti.

I soggetti diabetici devono avere un livello di emoglobina A1C inferiore al 7%. A volte è difficile raggiungere questo livello, ma quanto minore è il livello dell'emoglobina A1C, tanto minore è la probabilità di avere complicanze. I medici possono raccomandare un target leggermente inferiore o superiore ad alcuni soggetti, in base alle particolari condizioni di salute. Valori superiori al 9% dimostrano scarso controllo e livelli superiori al 12% un pessimo controllo. La maggior parte dei medici specializzati in diabetologia raccomanda di controllare l'emoglobina A1C ogni 3-6 mesi.

CAPITOLO 2
LA METODOLOGIA

In questo capitolo vengono presentate tutte le principali metodologie usate per poter effettuare un'analisi statistica sulle abitudini dei pazienti diabetici.

2.1 Modelli con variabile dipendente binaria

I modelli con variabile dipendente limitata sono utilizzati quando la variabile di risposta è limitata in modo naturale e rappresenta un conteggio o ha un intervallo specifico di valori possibili. Nel nostro caso la variabile target è una variabile dicotomica che assume, quindi, due valori possibili:

$$Y_i = \begin{cases} 1 & \text{se si sceglie l'alternativa 1} \\ 0 & \text{se NON si sceglie l'alternativa 1} \end{cases}$$

L'alternativa 1 equivale a *"soggetto diabetico"*.

Gli esiti sono mutuamente esclusivi, cioè si sceglie un'alternativa o l'altra.

Possiamo costruire un modello che ci aiuti nello spiegare quali sono i fattori di rischio, tra le abitudini dei pazienti, che influenzano l'insorgenza della patologia; quindi, siamo più interessati a capire la probabilità che ciò si manifesti.

L'obiettivo, quindi, è stimare p_i, ovvero la probabilità legata all'evento successo, ed individuare i fattori che possono influenzare l'insorgenza della patologia.

Sia *n* la dimensione campionaria, e sia Y_i la scelta per l'unità i-esima. Ciascun individuo avrà una probabilità diversa di contrarre il diabete, per cui la componente casule nei modelli per dati binari è costituita da una v.c. Bernoulliana

$$Y_i \sim Be(p_i)$$

dove la probabilità p_i è data da

$$P(Y_i = 1) = p_i$$

che è la probabilità di contrarre il diabete.

La funzione di probabilità per la variabile indicatrice binaria Y, essendo una variabile casuale bernoulliana è:

$$f(Y_i) = {p_i}^{Y_i} * (1 - p_i)^{(1-Y_i)}$$

Con valore atteso e varianza pari a

$$E(Y_i) = p_i$$
$$Var(Y_i) = p_i * (1 - p_i)$$

La funzione di legame che lega la variabile risposta al predittore lineare

$$n_{ij} = \sum \beta_j x_{ij}$$

deve essere tale da assicurare che per qualsiasi valore delle variabili esplicative la risposta Y sia compresa nell'intervallo [0,1]. Per questo motivo, un *modello di probabilità lineare* del tipo

$$\Pr(Y_i = y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad i = 1,2,\dots,n$$

è totalmente inadeguato. L'adozione del *modello di probabilità lineare* comporta che la variabile risposta (probabilità di successo) possa assumere anche valori negativi o superiori a 1, il che è in contraddizione con i postulati del *Calcolo delle Probabilità.* In passato, comunque, esso è stato spesso utilizzato in mancanza di software efficienti per gli altri modelli che, invece, risultano più coerenti.

Infatti, occorre individuare funzioni di legame che trasformano l'intervallo unitario [0,1] nella retta reale $(-\infty, \infty)$. Tale esigenza ha indotto gli studiosi ad adottare come legame funzionale, funzioni di ripartizione di v.c., in particolare della v.c. Normale e Logistica.

Le funzioni di ripartizione assicurano, infatti, i seguenti due limiti

$$\begin{cases} \lim_{\mathbf{x}_i'\beta \to \infty} F(\mathbf{x}_i'\beta) = 1 \\ \lim_{\mathbf{x}_i'\beta \to -\infty} F(\mathbf{x}_i'\beta) = 0 \end{cases}$$

in quanto

$$p_i = P(Y_i = 1|x_i') = F(x_i'\beta) \in [0,1] \quad \forall i$$

A seconda della v.c. associata alla funzione di ripartizione, si ha:

- Modello logit/logistico.
- Modello probit.
- Modello cloglog/loglog.

2.1.1 Modello logistico

Consideriamo la variabile casuale logistica.

$$p_i = P(Y_i = 1|x_i') = F(x_i'\beta) + \varepsilon_i = \Lambda(x_i'\beta) + \varepsilon_i$$

Con $\Lambda(\cdot)$ la funzione di ripartizione della distribuzione logistica.

Una variabile casuale X con funzione di densità

$$\lambda(x) = \frac{e^{-\frac{x-\lambda}{\delta}}}{\delta\left(1 + e^{-\frac{x-\lambda}{\delta}}\right)^2}$$

con:

- $-\infty < \lambda < \infty$, parametro di posizione
- $\delta > 0$, parametro di variabilità

è detta *Logistica* e si scrive $X \sim \Lambda(\lambda, \delta)$.

La distribuzione è simmetrica rispetto al parametro di posizione, ha code più pesanti rispetto ad una distribuzione normale; pertanto, si dice *leptocurtica*.

Di seguito la *Figura 1* mostra la distribuzione di densità della variabile al variare dei parametri.

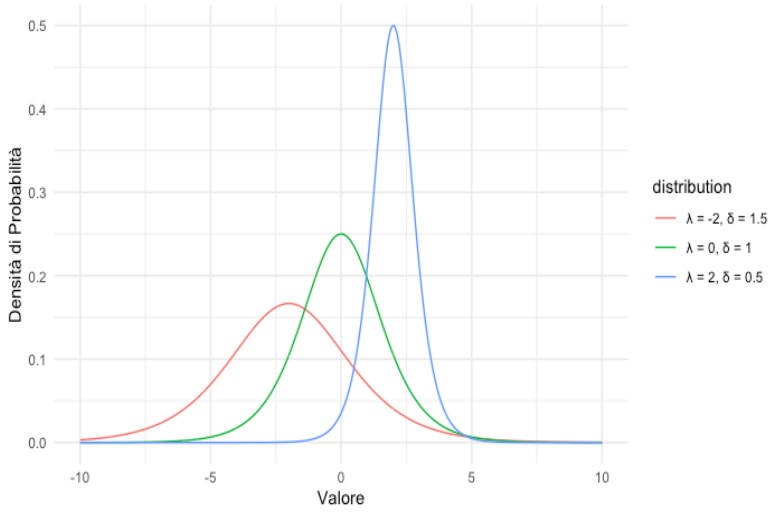

Figura 1: Distribuzione logistica

Nel caso della regressione logistica, i parametri sono fissi

$$\lambda = 0 \quad e \quad \delta = 1$$

Pertanto la funzione di densità sarà data da:

$$f(x) = \frac{e^{-x}}{(1+e^{-x})^2}$$

e la funzione di ripartizione da:

$$F(x) = \int f(x)\, dx = \int \frac{e^{-x}}{(1+e^{-x})^2}\, dx = \frac{e^{x}}{1+e^{x}}$$

Siccome la probabilità di successo è associata a una funzione non lineare nelle variabili esplicative e nei parametri, è possibile considerare una quantità che si chiama *odds*, che è dato dal rapporto tra la probabilità di successo e quella di insuccesso.

$$odds_i = \frac{p_i}{1-p_i} = \frac{P(Y_i = 1)}{P(Y_i = 0)}$$

Sostituendo le rispettive espressioni si ha:

$$odds_i = odd(Y_i = 1) = \frac{\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}}{1-\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}} = \frac{\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}}{\frac{1}{1+e^{x_i'\beta}}} = e^{x_i'\beta}$$

Notiamo che la quantità ottenuta non è ancora lineare nelle variabili esplicative e nei parametri; pertanto, si considera il logaritmo ottenendo, così, i logit della probabilità di successo per ogni osservazione.

$$\log(odds_i) = \log\left(\frac{p_i}{1-p_i}\right) = \log\left(e^{x_i'\beta}\right) = x_i'\beta$$

La quantità ottenuta si chiama *logit*(p_i).

- $Odds_i$ varia tra 0 e ∞.
- $Logit(p_i)$ varia tra -∞ e +∞.

Il motivo per cui si vuole ottenere una quantità lineare è soprattutto a scopo interpretativo.

Ovviamente, tornando alla linearità, otteniamo sempre il problema riscontrato con l'adozione del *modello di probabilità lineare*.

"Come si interpretano gli odds e i logit?"

- *Se* $odds_i$ *= 1 /* $logit(p_i) = 0 \rightarrow$ *eventi equiprobabili*
- *Se* $odds_i > 1$ */* $logit(p_i) > 0 \rightarrow$ *evento 1 più probabile dell'evento 0*
- *Se* $odds_i < 1$ */* $logit(p_i) < 0 \rightarrow$ *evento 1 meno probabile dell'evento 0*

Per quanto riguarda l'interpretazione dei coefficienti stimati, bisogna distinguere se la covariata è numerica o discreta

- Se la covariata è numerica, β_j misura il cambiamento nel logit per un incremento unitario in x_j, mentre $\exp(\beta_j)$ misura il cambiamento nell'odds per un incremento unitario in x_j.

- Se la covariata è dicotomica, un coefficiente positivo sta a significare che la probabilità di successo è più alta per la classe 1, se invece è negativo, p_i è più alta per la classe 0.

  Considerando una variabile binaria D_i a cui è associato il parametro δ, possiamo definire l'odds per $D_i = 1$ e l'odds per $D_i = 0$.

$$odds(Y_i = 1|D_i = 1) = \frac{P(Y_i = 1|D_i = 1)}{P(Y_i = 0|D_i = 1)} = e^{\beta_0 + \delta}$$

$$odds(Y_i = 1|D_i = 0) = \frac{P(Y_i = 1|D_i = 0)}{P(Y_i = 0|D_i = 0)} = e^{\beta_0}$$

  Si può definire una nuova quantità, chiamata *Odds ratio per* D_i, che è dato dal rapporto tra gli odds per $D_i = 1$ e $D_i = 0$.

$$odds\ ratio(D_i) = \frac{odds(Y_i = 1|D_i = 1)}{odds(Y_i = 1|D_i = 0)} = \frac{e^{\beta_0 + \delta}}{e^{\beta_0}} = e^{\delta}$$

Quindi se $e^{\delta} = 1$ significa che D_i non è statisticamente significativa poiché la presenza o assenza della categoria non influenza la probabilità.

Per valutare, invece, l'impatto sulla probabilità di successo della variazione di una covariata continua, si ricorre agli effetti marginali.

$$\frac{\partial}{\partial x_i'}\hat{p}_i = \frac{\partial}{\partial x_i'}F(x_i'\beta) = f(x_i'\beta)\hat{\beta} = \hat{p}_i(1 - \hat{p}_i)\hat{\beta} = \Lambda\left(x_i'\hat{\beta}\right)\left(1 - \Lambda\left(x_i'\hat{\beta}\right)\right)\hat{\beta}$$

Come notiamo dipendono dal valore delle variabili per l'i-esima unità e dal segno del parametro; pertanto, per eliminare la dipendenza dall'i-esima istanza si possono considerare gli effetti marginali medi e la media degli effetti marginali.

La media degli effetti marginali si ottiene attraverso una media di tutti gli effetti marginali calcolati per ciascuna istanza.

$$EMM\left(\hat{\beta}\right) = \frac{1}{n}\sum EM_i = \frac{1}{n}\sum \hat{p}_i(1 - \hat{p}_i)\hat{\beta}$$

Gli effetti marginali medi si ottengono calcolando normalmente gli effetti marginali, però, in questo caso, si considera il vettore delle medie delle variabili, che non dipende dall'i-esima istanza.

$$\frac{\partial}{\partial x_i'}\hat{p}_i = \frac{\partial}{\partial x_i'}F(\bar{x}\beta) = f(\bar{x}\beta)\hat{\beta} = \Lambda(\bar{x}\hat{\beta})\left(1 - \Lambda(\bar{x}\hat{\beta})\right)\hat{\beta}$$

2.1.2 Modello probit

Consideriamo la variabile casuale normale

$$p_i = P(Y_i = 1|x_i') = F(x_i'\beta) + \varepsilon_i = \Phi(x_i'\beta) + \varepsilon_i$$

con $\Phi(\cdot)$ la funzione di ripartizione della distribuzione normale standardizzata.

Una variabile casuale X con funzione di densità

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

con:

- media $-\infty < \mu < \infty$
- varianza $\sigma^2 > 0$

è detta *Normale* e si scrive $X \sim N(\mu\ ,\ \sigma^2)$. Essa è simmetrica rispetto alla media, è *mesocurtica*, e l'asimmetria è zero.

La media, mediana e moda coincidono e sono al centro della distribuzione. Circa il 68% dei dati si trova entro una deviazione standard dalla media (μ±σ), il 95% entro due deviazioni standard (μ±2σ) e il 99.7% entro tre deviazioni standard (μ±3σ).

Ponendo la media pari a zero e la varianza pari a 1 si ottiene la distribuzione *standardizzata* $X \sim Z(0\ ,\ 1)$. La distribuzione normale standardizzata ha una forma a campana simmetrica, centrata in x=0, e la sua area totale sotto la curva è 1.

La sua funzione di densità è data da

$$\phi(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}$$

Mentre la funzione di ripartizione è data da

$$\Phi(x) = \int \phi(u)\, du = \int \frac{1}{\sqrt{2\pi}}\, e^{-\frac{u^2}{2}}\, du$$

Nella *Figura 2 si* confrontano la distribuzione logistica con quella normale standardizzata.

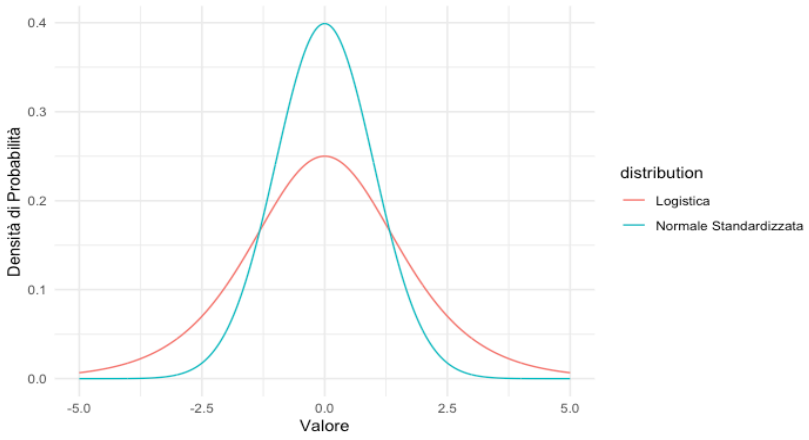

Figura 2: Confronto tra la distribuzione logistica e normale std.

Le code della distribuzione logistica sono più pesanti rispetto alla distribuzione normale standardizzata. Questo significa che la distribuzione logistica assegna una probabilità relativamente più alta agli eventi estremi rispetto alla distribuzione normale standardizzata. Importante notare che se $x_i'\beta \in [-1.2, 1.2]$, allora le probabilità stimate con il modello logit e probit sono molto simili; ciò che cambia è la varianza.

Quindi anche il modello probit non è lineare nei parametri e nelle variabili esplicative.

La *Figura 3* mostra il confronto tra le funzioni di ripartizione.

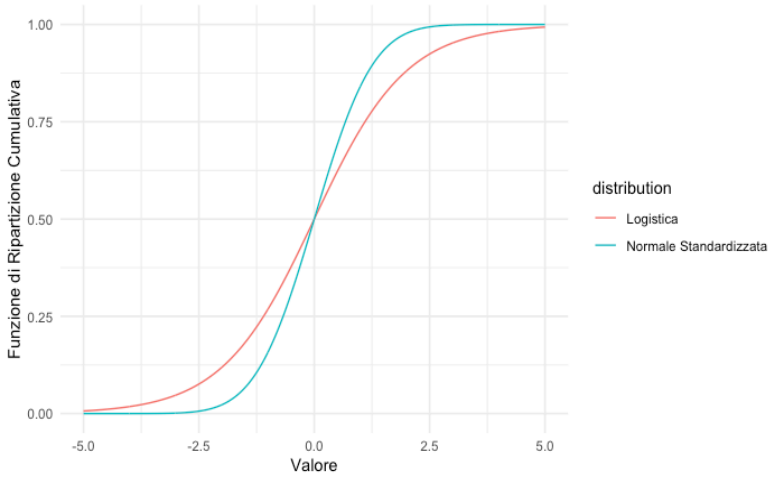

Figura 3: Confronto tra le funzioni di ripartizioni

La CDF della distribuzione normale standardizzata inizia lentamente, quindi aumenta gradualmente fino a raggiungere 1. La CDF della distribuzione logistica ha una forma a "S" più accentuata e inizia ad aumentare più rapidamente rispetto alla normale standardizzata. Questo indica code più pesanti rispetto alla normale.

La distribuzione logistica assegna una probabilità relativamente più alta agli eventi estremi rispetto alla distribuzione normale standardizzata. Questo è indicato dalla pendenza più elevata nella CDF della distribuzione logistica.

Al contrario del modello logistico, la cui funzione di ripartizione può essere scritta sottoforma di funzione chiusa, per il modello probit non è possibile quantificare l'effetto delle covariate sulla variabile di risposta.

Per quanto riguarda gli effetti marginali, si calcolano allo stesso modo di quelli del modello logit.

2.1.3 Modello C-loglog

Per l'ultimo modello analizzato, si considera una variabile causale che appartiene alle *extreme values distributions*[7], ovvero la distribuzione *Gumbel*[8].

Una variabile casuale X con funzione di densità

$$f(x) = \frac{1}{\delta} e^{-(z+e^{-z})}, \qquad z = \frac{x-\lambda}{\delta}$$

con:

- $-\infty < \lambda < \infty$, parametro di posizione
- $\delta > 0$, parametro di variabilità

è detta distribuzione di Gumbel e si scrive $X \sim P\ (\lambda\ ,\ \delta\ )$. È una delle distribuzioni più comuni utilizzate per modellare eventi estremi, come ad esempio il massimo di una serie di osservazioni. È ampiamente utilizzata nell'ambito della teoria delle code estreme e trova applicazioni in idrologia, meteorologia, ingegneria, finanza e altri campi.

La potenziale applicabilità della distribuzione di Gumbel per rappresentare la distribuzione dei massimi si riferisce alla *teoria dei valori estremi*, che indica che è probabile che sia utile se la distribuzione dei dati campione sottostanti è di tipo normale o esponenziale.

- Il parametro λ rappresenta la posizione e indica il punto in cui la coda della distribuzione inizia ad alzarsi.
- Il parametro δ rappresenta la scala ed è associato alla "larghezza" della coda della distribuzione

[7] La "extreme value distribution" (distribuzione delle variabili estreme) è una classe di distribuzioni di probabilità utilizzate per modellare eventi estremi o rari.

[8] È comunemente utilizzata per modellare il massimo delle osservazioni in un dato periodo di tempo o spazio. È appropriata quando il massimo è influenzato da molti fattori indipendenti.

Nella *Figura 4 si* confrontano la distribuzione di Gumbel con quella della normale standardizzata.

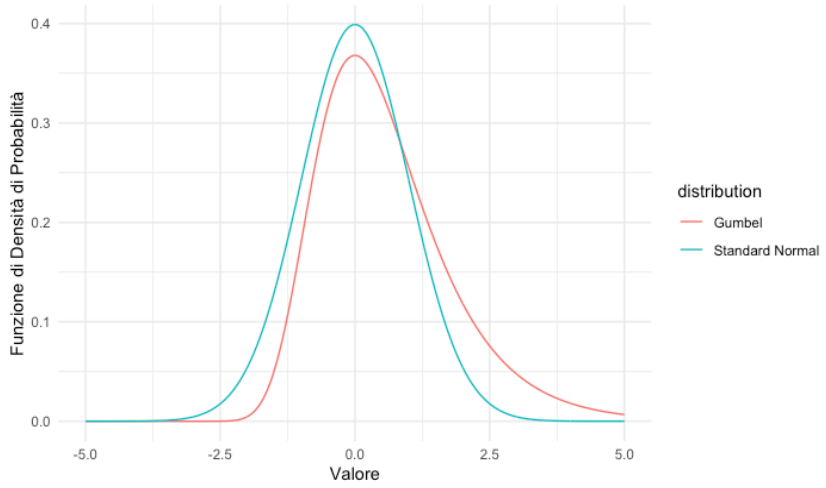

Figura 4: confronto tra la distribuzione di Gumbel e la normale standard

Da come si può notare, la distribuzione di Gumbel ha una coda più pesante rispetto alla distribuzione normale standardizzata; inoltre, la sua asimmetria le consente di trattare meglio eventi non equiprobabili.

In questo caso si ha

$$p_i = P(Y_i = 1|x_i') = F(x_i'\beta) + \varepsilon_i = e^{-e^{x_i'\beta}} + \varepsilon_i$$

con $e^{-e^{x_i'\beta}}$ funzione di ripartizione della *Gumbel distribution.*

Per valori piccoli di p_i, si avvicina al logit. Se la probabilità aumenta, si avvicina più lentamente rispetto al logit e probit.

I coefficienti associati ai predittori $(\beta_1, \beta_2 \dots)$ indicano come un'unità di cambiamento della variabile predittiva influisce sul $\log(-\log(p_i))$. Come nel caso logit, possiamo considerare l'esponenziale dei coefficienti ma, in questo caso, non si ottiene più l'effetto sugli odds, bensì sul $-\log(p_i)$.

Per ottenere invece l'effetto sulle probabilità

$$\hat{p}_i = e^{-e^{\hat{\beta}}}$$

2.2 Variable Selection

La variable selection è un processo che identifica le varabili più informative e rilevanti per un determinato modello o analisi. Questa pratica è fondamentale per ottenere modelli più semplici, interpretabili ed efficienti.

Considerando il caso in cui $k < n$, ovvero il numero di variabili minore del numero di osservazioni, si vuole ottenere un sottoinsieme di variabili che tiene conto del cosiddetto *principio di parsimonia.* Questo criterio stabilisce di preferire sempre il modello con il minor numero di variabili, a parità di efficienza.

I predittori non necessari aggiungeranno rumore (*"noise")* alla stima di altre quantità a cui si è interessati; inoltre, l'eliminazione delle variabili ridondanti permette di risparmiare tempo e costi, se il modello è utilizzato a scopo previsivo.

Esistono diverse procedure di selezionamento delle variabili rilevanti, tra cui si ricorda la *stepwise*. La Stepwise è un approccio che coinvolge l'aggiunta o la rimozione sequenziale delle variabili in base a criteri predefiniti, come la significatività statistica e il miglioramento delle metriche di adattamento del modello. Tra queste, si ricordano i Pseudo-R^2 i criteri di validazione che validano la bontà di adattamento di un modello, ampiamente descritti nel capitolo successivo.

I principali approcci per la stepwise selection sono la forward selection (selezione in avanti), backward elimination (eliminazione all'indietro), ed una combinazione dei due.

- *Forward selection:* il processo inizia senza alcuna variabile nel modello; quindi, si parte col considerare il modello nullo, compreso di sola intercetta. Ad ogni iterazione, viene selezionata e aggiunta una variabile predittiva che, dal confronto, sembra avere il maggiore impatto o contributo al modello. Si inizia col predittore più significativo, come determinato dai criteri di selezione definiti, ad esempio il più basso p-value o il miglioramento più significativo nei criteri di validazione come l'*AIC e BIC*[9].

  Alla fine di ciascuna iterazione, il modello viene valutato e il predittore selezionato viene mantenuto se, dal confronto con il modello ottenuto allo step precedente, porta un miglioramento dei criteri di selezione. Il processo termina quando, l'aggiunta di una nuova variabile nel modello non provoca un miglioramento nei criteri, per cui si ritiene ottimale il modello ottenuto con la variabili scelte all'iterazione precedente. Una riflessione deve essere fatta sulla regola di stop della procedura. La regola di stop è soddisfatta quando tutte le variabili rimanenti da considerare hanno un p-value più alto della soglia specificata, quando viene aggiunta al modello.

  Quando si arriva a questo stato, la forward selection si stoppa, e restituisce un modello che contiene solo le variabile con p-value minore della soglia. La soglia viene determinata attraverso un valore prefissato, che deve essere uguale per tutte le variabili, oppure determinato dai criteri di validazione.

[9] Vedere Paragrafo 3.4.4

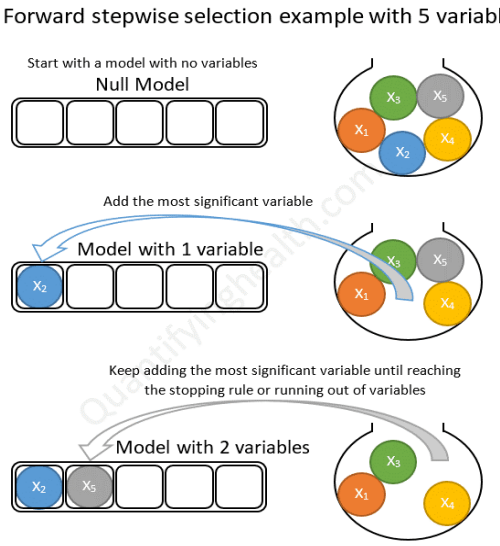

Figura 5: Forward selection

- *Backward selection:* il processo inizia con tutte le variabili predittive nel modello.
  Ad ogni iterazione, viene valutata la significatività statistica di ciascuna variabile e, se una variabile non è significativa o non contribuisce in modo significativo al modello, viene rimossa.
  Si inizia con tutte le variabili nel modello e successivamente si procede rimuovendo una variabile alla volta se, dal confronto con il modello ottenuto allo step precedente, l'eliminazione della variabile considerata provoca un miglioramento dei criteri di selezione.

  Il processo termina quando, la rimozione di una nuova variabile dal modello non provoca un miglioramento nei criteri, per cui si ritiene ottimale il modello ottenuto con la variabili scelte all'iterazione precedente. Per individuare le variabili meno significative a ciascuno step, si può guardare al più alto p-value, al più basso drop negli indici R^2 o all'incremento più basso in *RSS (Residuals Sum of Squares).*

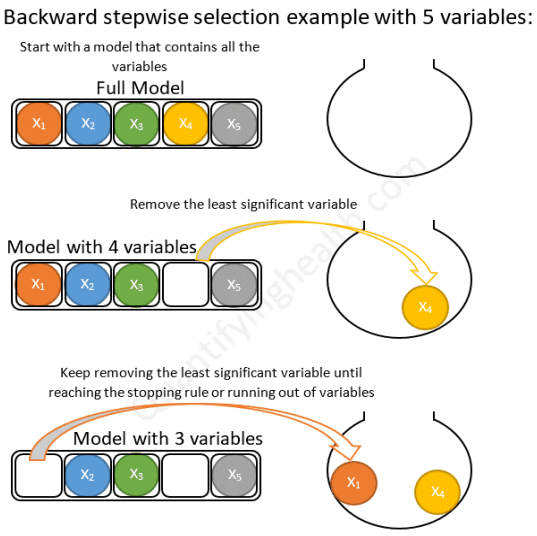

Figura 6: Backward selection

- *Stepwise completa o di tipo "both":* in questo contesto, vengono utilizzati entrambi gli approcci descritti precedentemente. Questo significa che durante il processo di selezione delle variabili, si considerano sia l'aggiunta progressiva di variabili significative, che la rimozione di variabili non più significative.

  Ad ogni iterazione, viene aggiunta una variabile se incrementa la spiegabilità della variabile di risposta e poi, nell'iterazione successiva, la stessa viene eliminata per verificare se è effettivamente rilevante.

Tra i vantaggi riconducibili a queste tecniche troviamo sicuramente quelli legati alla semplicità e automazione: attraverso una procedura facilmente riproducibile, la stepwise automatizza il processo di selezione delle variabili, rendendo più agevole la creazioni di modelli composti da features più dissimili tra loro. Ciò contribuisce a prevenire l'*overfitting* del modello, migliorando la stima sui dati outsample.

La stepwise, però presenta anche alcuni svantaggi soprattutto legati alla sovrastima dell'importanza di determinate variabili. Quest'ultime, infatti, vengono valutare in modo incrementale, senza considerare l'effetto complessivo sul modello.

Tende a produrre, inoltre, coefficienti, intervalli di confidenza, p-values e indici R^2 distorti. In particolare, i coefficienti potrebbero essere più larghi, gli intervalli di confidenza potrebbero essere più stretti, p-values potrebbero essere più piccoli, e l'R^2 più grandi. Gli s.e. potrebbero essere distorti, in particolare più bassi.

Per risolvere alcuni di questi problemi, si potrebbe procedere con un processo di selezionamento di determinate unità statistiche, per la creazione di due nuovi dataset:

- *Train set*
- *Test set*

I passaggi da svolgere sono indicati nel *Paragrafo 3.4.2.*

CAPITOLO 3
ANALISI STATISTICA SULLE ABITUDINI DEI PAZIENTI DIABETICI

3.1 Veridicità delle fonti

Affidarsi a enti attendibili, come istituti di ricerca accreditati o organizzazioni governative, è fondamentale per prendere decisioni ed evitare la diffusione di informazioni errate o disinformazione, che può avere impatti negativi sulle analisi.

Essa si riferisce alla qualità delle informazioni o dei dati stessi, ossia alla loro accuratezza e alla loro corrispondenza con la realtà o con i fatti veri. Per questo motivo si parla di campionamento rappresentativo, ovvero una tecnica utilizzata nella raccolta e nell'analisi dei dati che mira a selezionare un gruppo di individui o elementi in modo tale che rappresentino accuratamente l'intera popolazione o il fenomeno in studio.

Il *Centers for Disease Control and Prevention*[10] *(CDC)*,in questo contesto, è un'agenzia federale degli Stati Uniti d'America, situata ad Atlanta, in Georgia, specializzato nella promozione della salute, nella prevenzione delle malattie e nel controllo delle epidemie.

L'ente si occupa nel gestire un programma di sorveglianza comportamentale, chiamato "*Behavioral Risk Factor Surveillance System*"[11], composto da dati dettagliati sulla salute e i comportamenti legati alla salute della popolazione statunitense attraverso interviste telefoniche.

La crescita del programma di sorveglianza è stata confermata da notevoli successi: gli Stati lo usano per identificare problemi di salute emergenti, stabilire e seguire obiettivi, sviluppare e valutare programmi e politiche di sanità pubblica. Proprio per il fatto di essere organizzato a livello statale è stato usato da molti Stati anche per supportare iniziative legislative in materia di salute. Per esempio il Delaware l'ha sfruttato per creare un fondo per la prevenzione delle malattie legate al tabacco, il Nevada per mettere una tassa sulla vendita all'ingrosso di alcol distillato, e l'Illinois per approvare due atti: uno impone la presenza di aree non fumatori negli edifici pubblici, e l'altro l'inclusione degli screening mammografici in tutte le coperture assicurative sanitarie. Inoltre, sia i Cdc sia singoli Stati, analizzando i dati del sistema, hanno pubblicato articoli e rapporti, la maggior parte dei quali hanno avuto un'ampia disseminazione in letteratura e sui media.

Il BRFSS utilizza un campione probabilistico, specificamente un campionamento casuale stratificato, per raccogliere dati rappresentativi dalla popolazione adulta degli Stati Uniti.

3.1.1 Campionamento stratificato

Il campionamento stratificato è una tecnica di campionamento utilizzata nell'ambito delle indagini statistiche per ottenere un campione rappresentativo e accurato di una popolazione. Questo metodo coinvolge la suddivisione della popolazione in sottogruppi omogenei chiamati "*strati*" in base a determinate caratteristiche rilevanti, come età, sesso, livello di istruzione, occupazione, regione geografica, o altre variabili pertinenti.

[10] Center for disease control and prevention, website https://www.cdc.gov/about/

[11] *Behavioral Risk Factor Surveillance System*, website https://www.cdc.gov/brfss/index.html

Il processo di campionamento può essere sintetizzato in quattro fasi:

I. *Identificazione degli strati*: la prima fase coinvolge l'identificazione delle caratteristiche o variabili significative che possono influire sui risultati dello studio. Queste variabili sono scelte in base all'obiettivo dello studio e all'omogeneità che ci si aspetta all'interno di ciascun strato.

II. *Suddivisione della popolazione in strati*: Una volta identificate le variabili significative, la popolazione viene suddivisa in gruppi omogenei o strati in base a queste variabili. Ad esempio, se l'età è una variabile significativa, la popolazione viene suddivisa in gruppi di età come 18-30, 31-45, 46-60, oltre 60.

III. *Selezione casuale all'interno degli strati*: All'interno di ciascuno strato, viene eseguita una selezione casuale indipendente. Questo significa che i partecipanti vengono selezionati casualmente all'interno di ogni strato. Ciò consente di garantire una rappresentazione accurata e imparziale di ogni strato nella popolazione.

IV. *Campione estratto*: Una volta selezionati casualmente i partecipanti all'interno di ciascuno strato, essi costituiranno il campione rappresentativo che sarà oggetto di studio o indagine.

Questo tipo di campionamento si rivela particolarmente utile quando gli strati, all'interno della popolazione, sono di numerosità molto diversa. In una tale situazione di elevata variabilità del fenomeno, infatti, sarebbe necessario un campione molto ampio. Stratificando la popolazione, invece, è possibile ottenere una adeguata copertura degli strati meno numerosi anche con un campione di dimensioni ridotte, con un apprezzabile risparmio di tempi e costi di rilevazione.

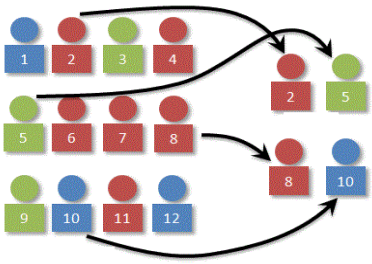

Figura 7: Campionamento stratificato

3.2 Fenomeno di studio

La presente tesi si propone di svolgere un'indagine statistica sulle abitudini dei pazienti diabetici. I dati reperiti si riferiscono alle risposte di un'intervista telefonica del 2015, promulgata dal *Cdc*, che verranno successivamente analizzate per ottenere informazioni utili e insight su varie questioni legate alla salute degli Stati americani.

L'intervista è composta da un totale di 22 domande, le cui risposte aperte sono state successivamente codificate per rendere il set di dati più chiaro e omogeneo possibile.

Il numero dei cittadini che ha partecipato all'intervista è 253.680, che corrisponde, quindi, al numero di righe del dataset.

In seguito viene riportato l'elenco delle domande poste ai candidati.

Tabella 1: Domande dell'intervista

Se il candidato è diabetico
Età del candidato
Genere del candidato
Pressione sanguigna
Livello di colesterolo
Se il candidato è stato sottoposto a un check per il colesterolo negli ultimi cinque anni
BMI: indice di massa corporeo
Se il candidato ha fumato almeno cinque pacchetti di sigarette in tutta la vita
Se il candidato ha avuto l'ictus
Se il candidato ha avuto attacchi di malattie cardiache come malattie coronariche o infarti al miocardio
Se il candidato ha svolto attività fisica negli ultimi 30 giorni
Se il candidato consuma frutta e verdura una o più volte al giorno
Se i candidati maschi e femmine bevono rispettivamente più di 14 e 7 drink a settimana
Se il candidato ha un'assicurazione sanitaria
Se il candidato nell'ultimo anno aveva bisogno di andare dal medico ma non l'ha fatto per mancanza di soldi
Salute percepita dal candidato
Giorni di cattiva salute mentale nell'ultimo mese
Giorni di cattiva salute fisica o infortuni nell'ultimo mese
Se il candidato ha problemi a camminare o a salire le scale
Livello di reddito
Livello di istruzione

3.2.1 Codifica dei dati

Prima di procedere alle analisi, viene effettuata la fase di codifica dei dati mediante il software statistico R.

La codifica dei dati in un dataset è un processo fondamentale nella preparazione dei dati per l'analisi e l'elaborazione da parte di algoritmi o modelli statistici. Questo processo trasforma i dati grezzi in un formato che può essere utilizzato in modo efficiente e accurato dall'algoritmo o dal modello. La modalità di codifica dei dati dipende dal tipo di dati che si sta elaborando e dall'obiettivo dell'analisi: i dati qualitativi sono trasformati in *fattori* o *fattori ordinati*, attraverso il costrutto *ifelse*[12] preceduto da *as.factor*[13], mentre i dati quantitativi, essendo già espressi da un numero, presentano una codifica automatica.

Per quanto riguarda le domande dicotomiche, cioè quelle che prevedono una sola risposta tra due modalità come sì/no, vero/falso, d'accordo/non d'accordo, vengono fattorizzate attraverso il procedimento indicato in precedenza.

L'intervista, inoltre, era basata anche da domande politomiche che prevedevano una sola risposta tra più di due modalità, come per esempio la domanda inerente alla salute percepita dal candidato: quest'ultimo poteva scegliere un valore tra 1 e 5[14] per valutare oggettivamente la propria salute; in questo caso la variabile è stata fattorizzata ed ordinata.

Le domande che prevedevano risposte aperte, come ad esempio quelle che richiedevano esattamente l'età, il livello di istruzione o il reddito, sono state, in un primo momento codificate per livelli, poi successivamente si è sentita la necessità di ridurne il numero; pertanto, alcuni di essi sono stati creati attraverso la fusione di più livelli.

3.2.2 Codifica della variabile Age

La variabile Age, che si riferisce alla domanda in cui si chiede l'età del candidato, è un esempio di variabile codificata due volte.

La prima codifica è stata realizzata per rendere la variabile discreta, inizialmente numerica, categoriale. Sono stati creati 14 livelli, ognuno di essi corrisponde a un intervallo di ampiezza 4, che va dai 18 anni di età[15] fino a un limite superiore, fissato a 89 anni.

Successivamente la variabile è stata nuovamente codificata con l'obiettivo di diminuire i livelli per ridurre la complessità delle analisi, facilitando l'interpretazione dei risultati e la visualizzazione empirica. Infatti da 18 si passa a 7 livelli di ampiezza 10, con il primo livello che raggruppa tutti i candidati con età inferiore ai 30 anni.

[12] In R, *ifelse* è una funzione condizionale in R che consente di eseguire operazioni condizionali element-wise su vettori o frame di dati.

[13] In R, *as.factor* è una funzione usata per convertire l'oggetto passato in un fattore, ovvero una variabile che può assumere un numero limitato di livelli o categorie discrete. Questo tipo di dato è molto utile per rappresentare informazioni qualitative come categorie, gruppi o livelli di un attributo.

[14] 1 = salute eccellente; 2 = salute buona; 3 = buona; 4 = discreta; 5 = scarsa.

[15] L'intervista è indirizzata ai soli maggiorenni, pertanto non è possibile individuare informazioni inerenti al diabete infantile e alle abitudini dei minorenni.

Tabella 2: Codifica della variabile Age

1: LESS THAN 30
2: AGE 30 TO 39
3: AGE 40 TO 49
4: AGE 50 TO 59
5: AGE 60 TO 69
6: AGE 70 TO 79
7: AGE 80 TO 89

3.3 Analisi esplorativa dei dati

L'*EDA (Exploratory data analysis)* è un approccio all'analisi di set di dati per riassumere le loro caratteristiche principali, spesso con metodi di visualizzazione e grafici, a prescindere da processi di modellizzazione formale o di verifica di ipotesi. L'analisi esplorativa dei dati è stata promossa da *John Tukey*[16] per incoraggiare gli statistici a esplorare i dati e possibilmente a formulare ipotesi che potrebbero portare a nuove raccolte di dati e nuovi esperimenti. I metodi non grafici generalmente comportano il calcolo di statistiche riassuntive, mentre i metodi grafici riassumono i dati in modo diagrammatico o pittorico. La maggior parte delle tecniche EDA sono grafiche, con alcuni indici e misure di tipo numerico. La ragione della forte dipendenza dalla grafica è che, per sua stessa natura, il ruolo principale di EDA è quello di esplorare i dati con una forte apertura mentale. La grafica permette di guardare ai dati in modo da rilevare gli aspetti strutturali per ottenere informazioni nuove, spesso insospettate e sorprendenti.

Per le *variabili categoriali* vengono calcolate le tabelle di frequenza e tabelle di contingenza, rappresentate con diagrammi a barre (di vario tipo).

Per le *variabili a quantitative* si calcolano misure di posizione o tendenza centrale come la media, la mediana e i quartili, misure di dispersione come la deviazione standard, la varianza e la MAD e, infine, misure di forma come l'indice di simmetria e curtosi.

Fra gli strumenti grafici associati a variabili numeriche vi sono i boxplots[17], stime di densità kernel e normal probability plots.

[16] John Tukey è stato un matematico e statistico americano , meglio conosciuto per lo sviluppo dell'algoritmo veloce della trasformata di Fourier (FFT) e del boxplot. Il test dell'intervallo di Tukey , la distribuzione lambda di Tukey , il test di additività di Tukey e il lemma di Teichmüller-Tukey portano tutti il suo nome. A lui viene anche attribuita la coniazione del termine " bit " e il primo utilizzo pubblicato della parola " software ".

[17] Un boxplot è un grafico che mostra la distribuzione e la variabilità di un insieme di dati numerici. Mostra i quartili, la mediana e identifica eventuali valori anomali, aiutando a comprendere la forma e la dispersione dei dati in modo sintetico e informativo.

Ciascun metodo può essere classificato come:

- *univariato,* esamina una variabile alla volta.

- *multivariato,* esamina due o più variabili alla volta per esplorare le relazioni e il condizionamento.

Il *condizionamento* è l'anima della statistica. *Condizionamento* è uno strumento potente (e semplice) per l'analisi dei dati esplorativi, soprattutto se associato alla colorazione e al faceting. Le relazioni multivariate possono essere facilmente individuate utilizzando questi strumenti. In R questi strumenti sono prontamente disponibili sia nel sistema base che, a un livello più avanzato, in diversi pacchetti come ggplot2[18]. Le statistiche riassuntive possono anche essere calcolate utilizzando il condizionamento. Anche in questo caso il linguaggio R offre strumenti efficaci per calcolare le statistiche sui sottogruppi di dati (ottenuti condizionando uno o più valori delle variabili incluse nel set di dati).

È quasi sempre una buona idea eseguire un EDA univariato su ciascuno dei componenti della matrice di dati prima di eseguire l'EDA multivariato.

3.3.1 Analisi esplorativa univariata

Segue, in questo paragrafo, un'analisi approfondita delle risposte dei candidati in relazione a ciascuna domanda, utilizzando gli strumenti statistici analizzati al paragrafo precedente. Innanzitutto, ci si sofferma sulle domande relative alle informazioni generali di tipo demografico, per conoscere genere, età, educazione e reddito degli intervistati.

Dalla *Figura 8* e dalla *Tabella 3* si evince che la maggior parte dei candidati ha un età compresa tra i 60 e i 69[19] anni, circa il 26 %. Nel complesso possiamo dire che la ricerca è abbastanza distribuita in base all'età, la prevalenza non è estremamente dominante o significativa rispetto alle altre categorie. Come abbiamo detto precedentemente, le interviste sono rivolte a soli maggiorenni; pertanto, non è possibile individuare informazioni sulle abitudini dei diabetici in età infantile.

Tabella 3: Frequenze percentuali

Categoria	*Frequenza*
1	*5,2 %*
2	*9,8 %*
3	*14,2 %*
4	*22,5 %*
5	*25,8 %*
6	*15,6 %*
7	*6,8 %*

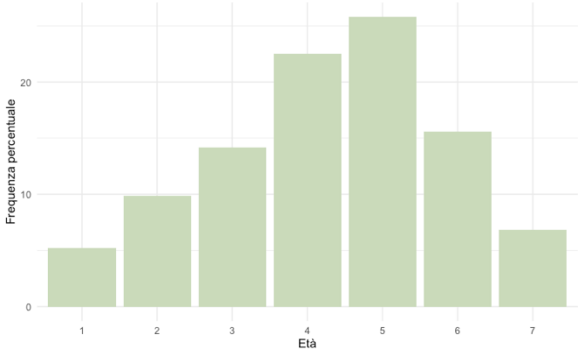

Figura 8: Bar plot relativo all'età dei candidati

[18] *ggplot2 (Grammar of Graphics plot 2)* è un popolare pacchetto di visualizzazione dei dati per il linguaggio di programmazione R.

[19] Vedi paragrafo 3.2.2 per la codifica della variabile Age

Dalla *Figura 9* e nella *Tabella 4* si evince una prevalenza da parte dei candidati di sesso femminile, pari al 56%.

Tabella 4: Frequenze percentuali

Categoria	*Frequenza*
Femmina	*56 %*
Maschio	*44 %*

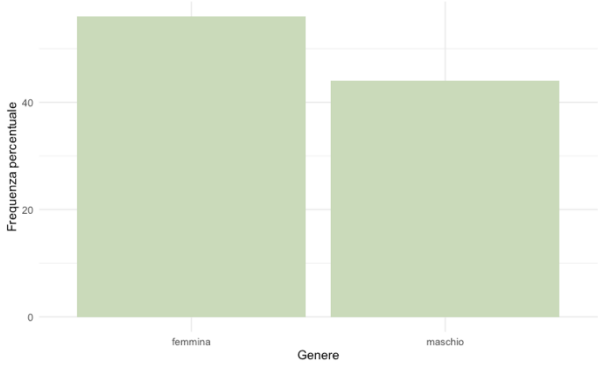

Figura 9: Bar plot relativo al genere dei candidati

La variabile Educazione riporta le risposte dei candidati alla domanda relativa all'istruzione del rispondente. In seguito, nella *Tabella 5,* viene riportata la codifica della variabile categoriale.

Tabella 5: codifica variabile educazione

1: Mai frequentato la scuola o infanzia
2: Grades 1 through 8 (Elementary)
3: Grades 9 through 11 (Some high school)
4: Grade 12 or GED (High school graduate)
5: College 1 year to 3 years (Some college or technical school)
6: College 4 years or more (College graduate)

La *Figura 10* e la *Tabella 6* mostrano una netta prevalenza di candidati laureati, circa il 42 % dei candidati.

Tabella 6: Frequenze percentuali

Categoria	*Frequenza*
1	*0,1 %*
2	*1,6 %*
3	*3,7 %*
4	*24,7 %*
5	*27,6 %*
6	*42,3 %*

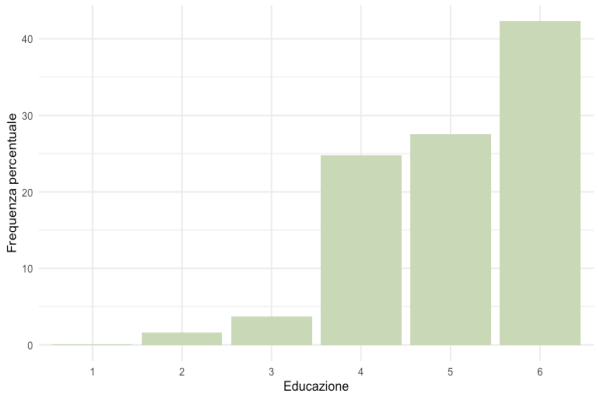

Figura 10: bar plot relativo all'educazione dei candidati

Siccome le prime due categorie comprendono una bassa percentuale di candidati, si può pensare di unirle così da ridurre il numero di classi della variabile.

In seguito, la F*igura 11* e la *Tabella 7* mostrano la distribuzione delle classi ricodificate. Ancora una volta si nota una netta prevalenza della sesta categoria, val al dire la maggioranza dei candidati laureati.

Tabella 7: Frequenze percentuali

Categoria	*Frequenza*
2	*1,7 %*
3	*3,7 %*
4	*24,7 %*
5	*27,6 %*
6	*42,3 %*

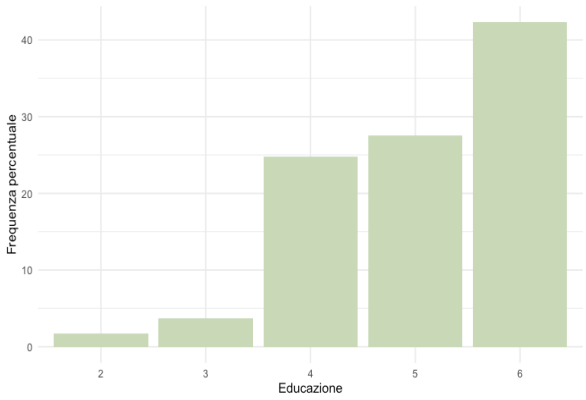

Figura 11: bar plot relativo all'educazione ricodificata

La *Figura 12* e la *Tabella 9* mostrano il reddito dei candidati. Anche in questo caso viene effettuata una seconda codifica della variabile per diminuire il numero di classi, riportata dalla *Tabella 8*. Essa è composta da sette livelli

Tabella 8: codifica tabella Reddito

2: meno di 15.000 $
3: dai 15 ai 20.000 $
4: dai 20 ai 25.000 $
5: dai 25 ai 35.000 $
6: dai 35 ai 50.000 $
7: dai 50 ai 75.000 $
8: più di 75.000 $

Il bar plot mostra come il reddito della maggior parte dei rispondenti è oltre ai 75 mila dollari. Basse, invece, le percentuali nei livelli inferiori di reddito.

Tabella 9: Frequenze percentuali

Categoria	*Frequenza*
2	*9 %*
3	*6 %*
4	*8 %*
5	*10 %*
6	*14 %*
7	*17 %*
8	*36 %*

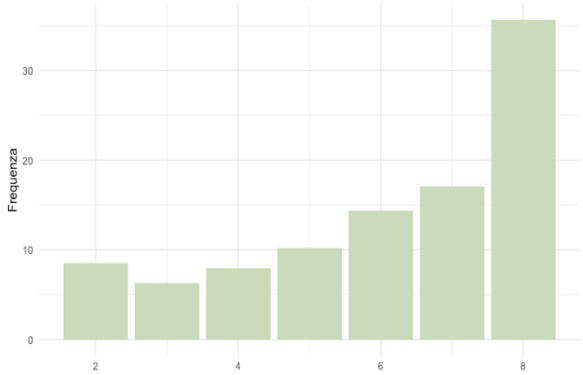

Figura 12: bar plot relativo al reddito dei candidati

Dopo aver analizzato il profilo demografico dei candidati, si è proceduto con l'analisi della variabile *target*, il diabete, che riporta le risposte dei candidati riguardo alla presenza della patologia.

La *Figura 13* e la *Tabella 10* mostrano una netta prevalenza di candidati senza il diabete; ciò ci porta a considerare che la scelta del campione sia giusta, in quanto abbastanza rappresentativo della popolazione. I diabetici nel campione sono circa il 14 %.

Tabella 10: Frequenze percentuali

Categoria	*Frequenza*
Diabetico	*14 %*
Non diabetico	*86 %*

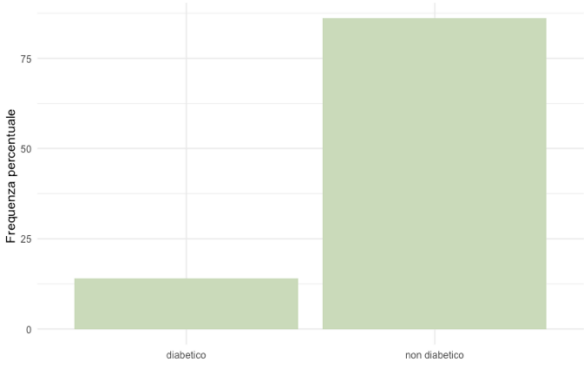

Figura 13: bar plot relativo al diabete

In seguito l'analisi di tutte le componenti che potrebbero incidere sulla patologie.

Iniziamo con l'analisi della variabile *BMI*, unica variabile continua del dataset che tiene conto dell'indice di massa corporeo di ciascun candidato.

Il BMI (*Body Mass Index*) è un indicatore comunemente utilizzato per valutare la composizione corporea di un individuo in rapporto al proprio peso e altezza. Si ottiene dividendo il peso (in kg) per il quadrato dell'altezza (in metri).

La formula matematica è la seguente:

$$BMI = \frac{peso(Kg)}{altezza^2(m^2)}$$

Interpretazione:

- $16{,}5 \leq BMI < 18{,}5$, Indica che il peso corporeo è inferiore rispetto a quello considerato sano per l'altezza.
- $18{,}5 \leq BMI < 25$, Indica che il peso corporeo è nella fascia considerata sana per l'altezza.
- $25 \leq BMI < 30$, Indica un eccesso di peso rispetto alla fascia considerata sana.
- $BMI \geq 30$, Indica un'eccessiva adiposità corporea, con ulteriori suddivisioni in:
  - Obesità di classe I (30-34,9).
  - Obesità di classe II (35-39,9).
  - Obesità di classe III (≥ 40).

La *Figura 14* mostra l'istogramma della variabile, mentre la *Tabella 11* presenta la sintesi statistica[20] del BMI.

Tabella 11: sintesi statistica

	BMI
MIN	*12,00*
Q_1	*24,00*
MEDIANA	*27,00*
MEDIA	*28,38*
Q_3	*31,00*
MAX	*98,00*

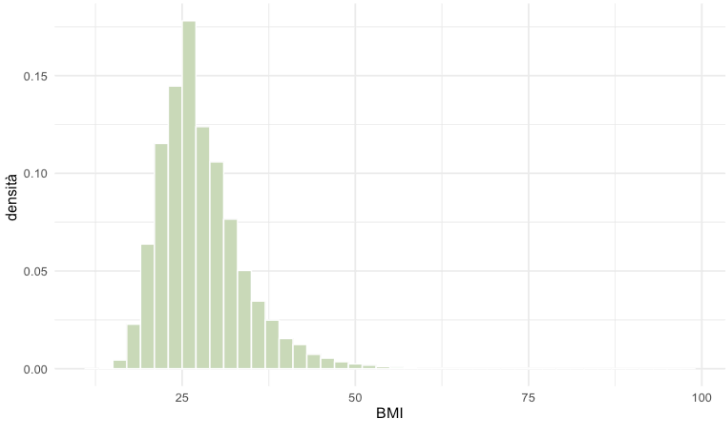

Figura 14: Istogramma del BMI

[20] La *sintesi statistica* fornisce una visualizzazione concisa e utile delle principali misure statistiche, come minimo, massimo, mediana, quartili e conteggio dei dati mancanti, per aiutarti a comprendere la distribuzione e le proprietà dei tuoi dati in modo rapido ed efficace.

Dalla visualizzazione risulta che il BMI mediano è più basso di quello medio, questo perché la media è un indice di posizione sensibile a valori anomali e all'asimmetria positiva della distribuzione. Un BMI pari a 98 ed a 12, rispettivamente valori massimo e minimo della distribuzione, sono ottimi candidati ad *outlier*[21] pertanto sarebbe opportuno attenzionarli.

Eseguo uno screening più approfondito s(*Figura 15)* per la visualizzazione dei valore anomali visto che, nel grafico dell'istogramma, non sembrano comparire.

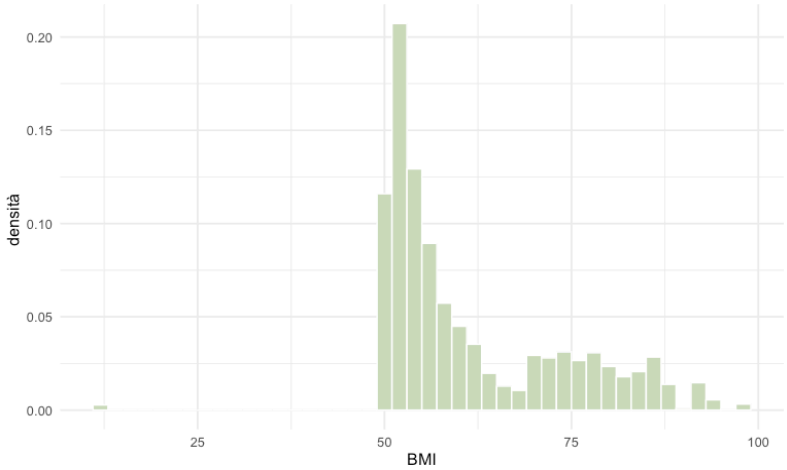

Figura 15: Visualizzazione dei presunti outliers

La *Figura 15* mostra tutti i possibili valori anomali (circa 2170). Secondo alcune ricerche valori di BMI che superano 50 o addirittura 60, rientrano nella categoria *super super obeso*. Questa, però, ha un'incidenza statistica talmente bassa da non essere stata ancora inclusa nella suddivisione di base. Pertanto proseguo con la rimozione dei valori anomali.

La *Figura 16* mostra la nuova distribuzione di densità della variabile. Data l'asimmetria positiva, si può determinare una maggiore presenza di candidati normopeso e sottopeso; ciò è in accordo con la proporzione di persone non diabetiche nel campione. Si ricorda che l'obesità rientra tra i primissimi fattori di rischio per la patologia.

Tabella 12: sintesi statistica

	BMI
MIN	14,00
Q_1	24,00
MEDIANA	27,00
MEDIA	28,22
Q_3	31,00
MAX	59,00

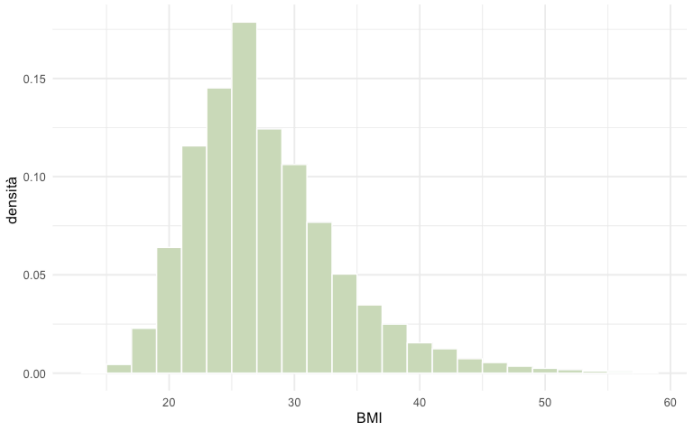

Figura 16: Istogramma relativo al BMI aggiustato

[21] Gli *outliers* sono punti dati che si discostano in modo significativo dal resto del set di dati. Sono osservazioni che presentano un comportamento insolitamente alto o basso rispetto alla maggior parte degli altri valori nel dataset.

Tra i fattori di rischio per il diabete viene analizzato il livello di colesterolo[22] dei rispondenti. Il controllo del colesterolo è fondamentale perché il diabete è spesso associato a un aumento del colesterolo nel sangue, in particolare del LDL (lipoproteine a bassa densità) comunemente noto come "colesterolo cattivo".

La *Figura 17* mostra il bar plot relativo alla percentuale dei rispondenti con un alto livello di colesterolo. Circa il 58% dei rispondenti non ha un elevato livello di colesterolo.

Tabella 13: Frequenze percentuali

Categoria	*Frequenza*
Alto colesterolo	*42 %*
Non alto	*58 %*

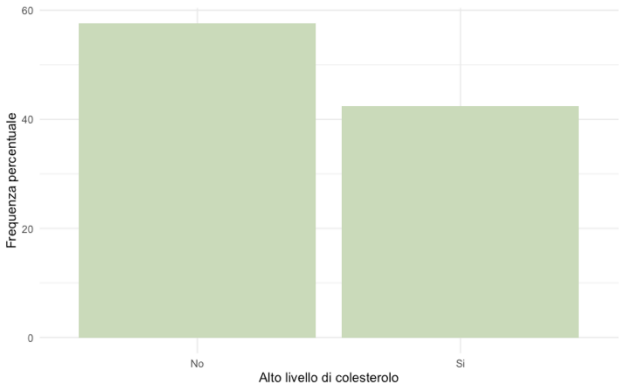

Figura 17: bar plot relativo al livello di colesterolo

La *Figura 18* segnala che la maggioranza dei candidati non soffre di pressione alta. Circa il 43% dei rispondenti soffre di alta pressione sanguigna.

Tabella 14: Frequenze percentuali

Categoria	*Frequenza*
Alto pressione	*43 %*
Non alta	*57 %*

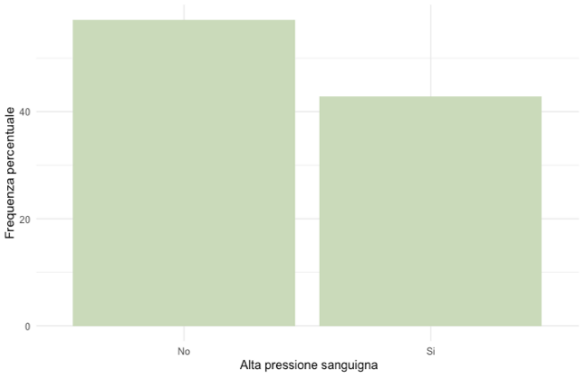

Figura 18: bar plot relativo alla pressione sanguigna

La *Figura 19* mostra la percentuale di candidati che si è sottoposto a un check per il colesterolo negli ultimi 5 anni. Vi è una netta prevalenza di rispondenti che si è sottoposta a una visita almeno una volta, circa il 96 %. La valutazione regolare del colesterolo è fondamentale per le persone col diabete, volta a prevenire malattie cardiache e vascolari.

[22] Il colesterolo è una sostanza cerosa, grassa e simile a un alcol presente in tutte le cellule del corpo umano. È essenziale per la costruzione delle membrane cellulari, la produzione di ormoni, la sintesi della vitamina D e la produzione di acidi biliari necessari per la digestione dei grassi.

Tabella 15: Frequenze percentuali

Categoria	*Frequenza*
Si	*96 %*
No	*4 %*

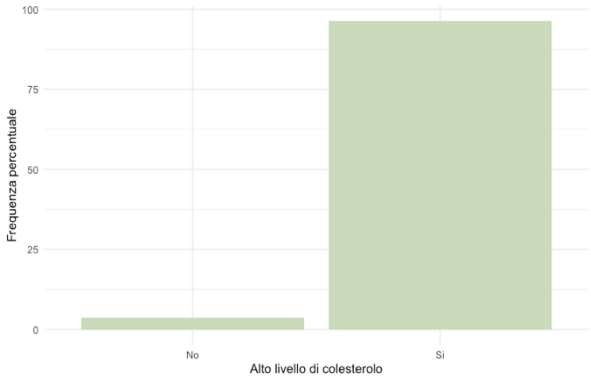

Figura 19: bar plot relativo al check per il colesterolo

Dalla *Figura 20* si evince che il 44 % dei candidati ha fumato almeno cinque pacchetti in tutta la sua vita.

Tabella 16: Frequenze percentuali

Categoria	*Frequenza*
Si	*44 %*
No	*56 %*

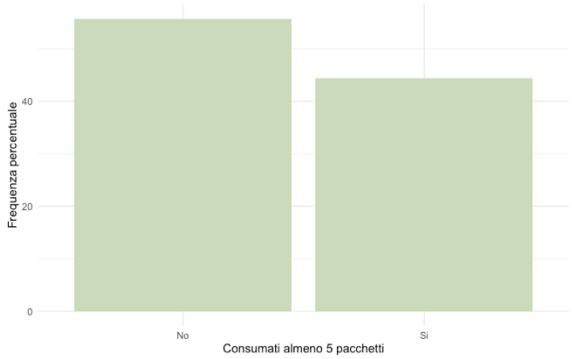

Figura 20: bar plot sul consumo di sigarette

Nella *Tabella 17* si analizzano le risposte in merito all'ictus. Solo il 4% dei rispondenti ha avuto un ictus.[23]

Tabella 17: Frequenze percentuali

Categoria	*Frequenza*
Si	*4 %*
No	*96 %*

[23] *L'ictus è* una condizione medica che si verifica quando il flusso sanguigno verso una parte del cervello viene improvvisamente interrotto o ridotto. Questa interruzione del flusso sanguigno può essere causata da un coagulo di sangue che blocca un vaso sanguigno nel cervello (ictus ischemico) o da una rottura di un vaso sanguigno nel cervello (ictus emorragico).

Nella *Figura 21* si analizzano le risposte in merito ad altre malattie coronariche o infarti al miocardio. Anche in questo caso la percentuale è bassa, circa il 9% dei rispondenti ha avuto un infarto.

Tabella 18: Frequenze percentuali

Categoria	*Frequenza*
Si	*9 %*
No	*91 %*

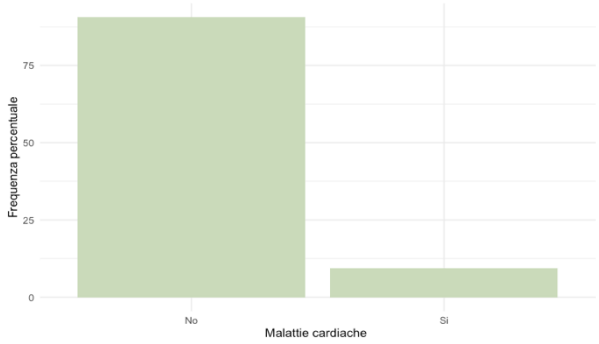

Figura 21: bar plot sull'insorgenza di malattie cardiache

La *Figura 22* mostra se il candidato ha svolto attività fisica negli ultimi 30 giorni. Il 75 % dei rispondenti ha risposto positivamente alla domanda.

Tabella 19: Frequenze percentuali

Categoria	*Frequenza*
Si	*9 %*
No	*91 %*

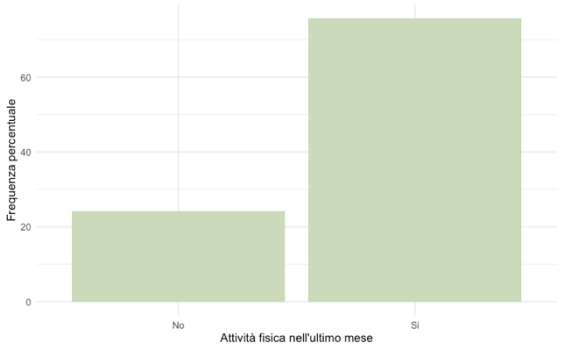

Figura 22: bar plot relativo alle risposte sull'attività fisica

La *Figura 23* e *24* analizzano le risposte dei candidati in merito al consumo di Frutta e Verdura. Per quanto riguarda la frutta il 63 % dei rispondenti ha dichiarato di consumarla frequentemente; le percentuali si alzano per la verdura. Circa l'81% dei rispondenti integra la verdura nella propria dieta.

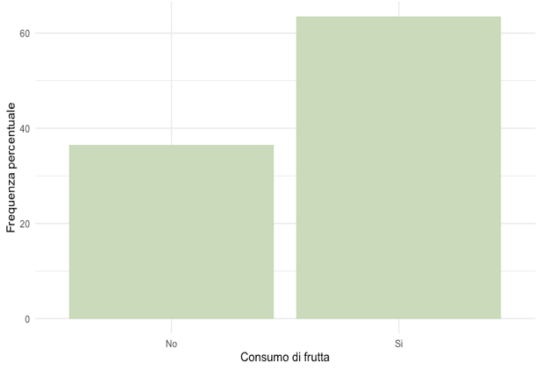

Figura 23: bar plot sul consumo di frutta

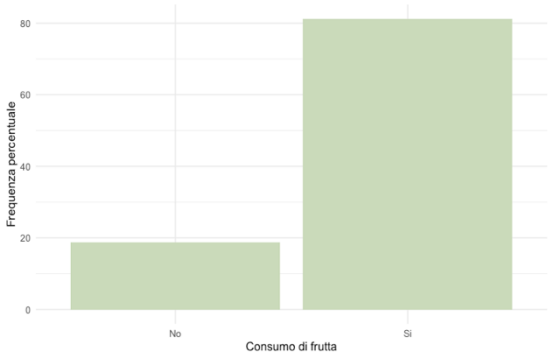

Figura 24: bar plot sul consumo di verdura

La *Tabella 20* mostra la frequenza dei candidati in possesso di un'assicurazione sanitaria[24]. Il 95% dei rispondenti risponde positivamente alla domanda.

Tabella 20: Frequenze percentuali

Categoria	*Frequenza*
Si	*95 %*
No	*5 %*

Nella *Tabella 21* si evince che più dell'8 % dei candidati aveva bisogno di andare dal medico ma non l'ha fatto per mancanza di soldi. Si ricorda che la maggior parte dei rispondenti ha un reddito superiore ai 75 mila dollari (tabella 9).

Tabella 21: Frequenze percentuali

Categoria	*Frequenza*
Si	*8 %*
No	*92 %*

Nella *Tabella 22* si evince che solo il 6% dei candidati bevono più di 14 drink a settimana (nel caso di rispondenti donne il numero di drink cala a 7).

Tabella 22: Frequenze percentuali

Categoria	*Frequenza*
Si	*6 %*
No	*94 %*

[24] Negli Stati Uniti, il sistema sanitario è basato principalmente sul sistema di assicurazione sanitaria privata, dove le persone pagano premi assicurativi per ottenere copertura medica.

La *Figura 25* mostra, attraverso un istogramma, i giorni di cattiva salute mentale nell'ultimo mese dei singoli rispondenti.

Tabella 23: sintesi statistica

	BMI
MIN	0
Q_1	0
MEDIANA	0
MEDIA	3,2
Q_3	2
MAX	30

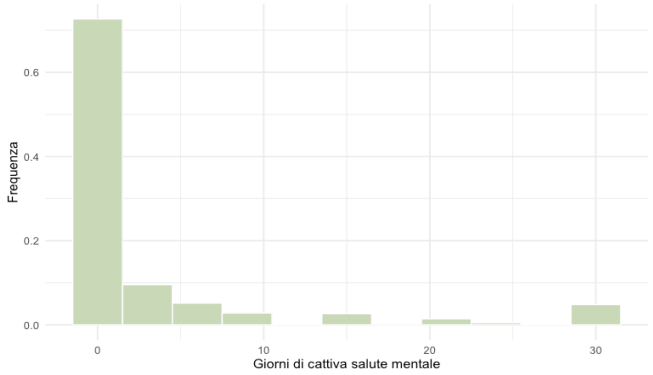

Figura 25: istogramma relativo ai giorni di cattiva salute mentale

Si nota una forte asimmetria positiva; pochi sono i candidati con instabilità mentale.

Decido di trasformare la variabile numerica discreta in categoriale composta da tre livelli differenti:

- 1: se i giorni d'instabilità mentale sono 0
- 2: se i giorni d'instabilità mentale sono minori di 15
- 3: se i giorni d'instabilità mentale sono maggiori di 15

In seguito, viene riportata la frequenza percentuale nelle classi

Tabella 24: Frequenze percentuali

Categoria	*Frequenza*
1	*69,3 %*
2	*21,5 %*
3	*9,2 %*

La *Tabella 25* mostra che il 17% circa dei rispondenti fatica a camminare o a salire le scale.

Tabella 25: Frequenze percentuali

Categoria	*Frequenza*
Si	*17 %*
No	*83 %*

Si analizza il numero di giorni di cattiva salute fisica o infortunio nell'ultimo mese.

Dalla *Figura 26* si nota che, come nella *Figura 25*, vi è una forte asimmetria positiva; pertanto, la maggior parte dei canditati è risultata essere fisicamente sana negli ultimi trenta giorni.

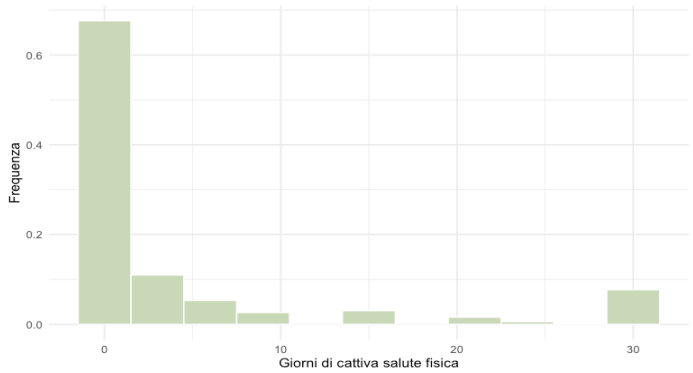

Figura 26: istogramma relativo ai giorni di cattiva salute fisica

Anche in questo caso si è deciso di rendere la variabile di tipo categoriale composta da tre livelli:

- 1: se i giorni di cattiva salute fisica sono 0
- 2: se i giorni di cattiva salute fisica sono minori di 15
- 3: se i giorni di cattiva salute fisica sono maggiori di 15
-

In seguito, viene riportata la frequenza percentuale nelle classi

Tabella 26: Frequenze percentuali

Categoria	*Frequenza*
1	*63,1 %*
2	*24,7 %*
3	*12,2 %*

Infine, si analizzano le risposte circa alla domanda posta ai candidati nel fare una valutazione oggettiva alla propria salute.

La variabile categoriale è composta da cinque livelli:

- 1: salute eccellente
- 2: molto buona
- 3: buona
- 4: discreta
- 5: scarsa

La *Tabella 27* riporta le frequenze percentuali della variabile. Secondo i dati, vi è una prevalenza dei candidati che reputa la propria salute *"molto buona",* la percentuale diminuisce nelle ultimi classi. In generale i candidati credono di avere una buona salute fisica.

Tabella 27: Frequenze percentuali

Categoria	*Frequenza*
1	*17,9 %*
2	*35,1 %*
3	*29,8 %*
4	*12,5 %*
5	*4,7 %*

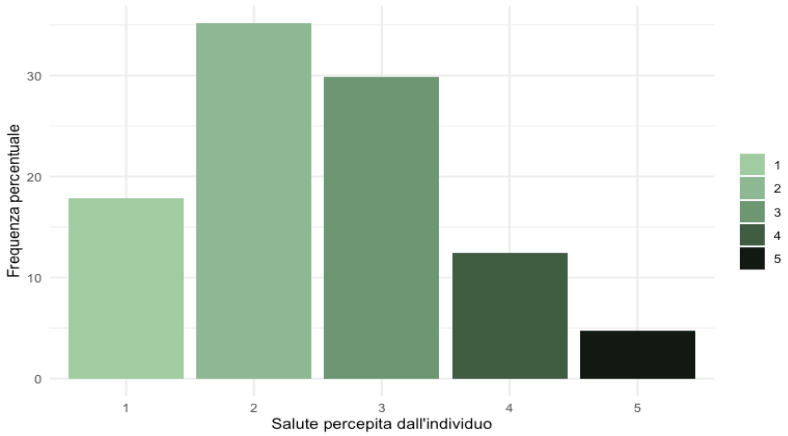

Figura 27: bar plot relativo alla salute percepita dall'individuo

3.3.2 Analisi esplorativa multivariata

Terminata l'analisi esplorativa delle singole variabili, adesso si passa all'analisi bivariata e multivariata per comprendere le relazioni e le associazioni tra le variabili, al fine di cogliere degli aspetti che, non sarebbero noti.

La prima area che si intende analizzare è quella demografica. L'obiettivo è capire se informazioni generali di tipo demografico influenzano la presenza del diabete.

Si vuole rispondere a domande del tipo:

- *"L'età e il sesso condizionano la comparsa del diabete?"*
- *"Basso livello di reddito può essere considerato un fattore di rischio per la patologia?"*

La *Figura 28* mostra una proporzionalità diretta tra età e il diabete, quindi una reale associazione tra le due variabili. All'aumentare dell'età, infatti, aumenta la prevalenza diabetica.

L'invecchiamento può portare a una diminuzione della sensibilità all'insulina (insulino-resistenza) e a una ridotta capacità del pancreas di produrre insulina in modo efficiente, contribuendo all'insorgenza del diabete di tipo 2[25].

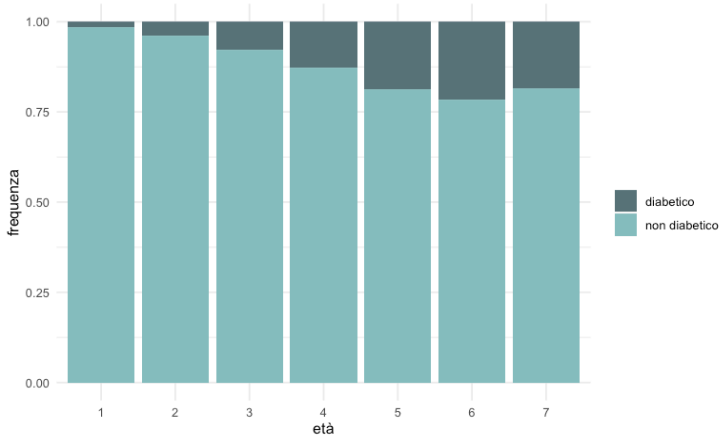

Figura 28: bar plot sull'età e diabete

[25] Vedi cap. I, par. I

Dalla *Figura 29* si evince che il sesso del soggetto non è un fattore condizionante per la presenza del diabete. Si nota, nella *Tabella 28,* un leggero aumento della frequenza di diabetici nel genere femminile molto probabilmente a causa della gravidanza: le cellule sono meno sensibili all'azione dell'insulina, con la conseguenza che i livelli di glicemia aumentano. In questo caso si parla di diabete gestazionale.[26]

Tabella 28: tabella di contingenza

	Diabetico	*Non diabetico*
Femmina	7,3 %	48,%
Maschio	6,7 %	37,%

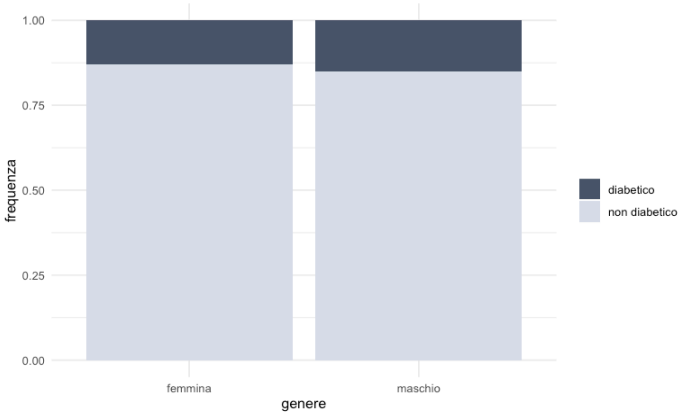

Figura 29: bar plot relativo al genere e al diabete

Infine si analizza la frequenza di diabetici per ogni livello di reddito dei rispondenti.

La *Figura 30* mostra una relazione inversa tra reddito e diabete. In generale, un reddito più elevato può essere associato a uno stile di vita più sano e ad abitudini più salutari, come una dieta equilibrata, l'accesso a cibo di qualità, l'opportunità di partecipare a attività fisiche regolari e una migliore gestione dello stress. Questi fattori possono contribuire a ridurre il rischio di sviluppare il diabete, in particolare il diabete di tipo 2, che è spesso influenzato da fattori di stile di vita.

Alcuni studi hanno dimostrato che l'accesso a una migliore assistenza sanitaria e a controlli medici regolari, che spesso è correlato a un reddito più alto, può contribuire a una migliore gestione e prevenzione del diabete.

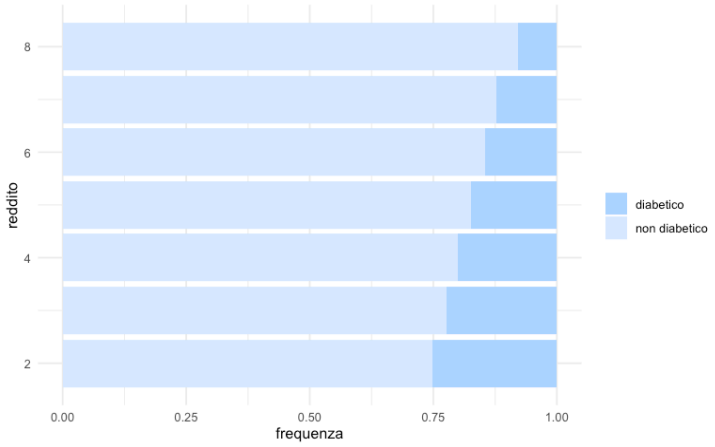

Figura 30: bar plot relativo al reddito e diabete

[26] Vedi cap. I, par. I

Un'altra area da prendere in considerazione riguarda l'analisi dei fattori di rischio per altre patologie come l'ictus o infarti al miocardio.

Si verifica se il BMI influenza il rischio di sviluppare infarti cardiaci o altri condizioni cardiache. Eseguo un confronto delle sintesi statistiche per rispondenti colpiti da gravi condizioni cardiache e rispondenti sani.

Dai risultati mostrati nella *Tabella 29* possiamo dire che in media il BMI dei candidati colpiti da infarto al miocardio è maggiore rispetto al BMI dei candidati sani.

Tabella 29: sintesi statistiche per il BMI condizionato alle condizioni cardiache

	Min	Q_1	***Mediana***	***Media***	Q_3	***Max***
Heart disease	*14*	*25*	*28*	*29,33*	*33*	*59*
No Heart disease	*14*	*24*	*27*	*28,11*	*31*	*59*

Ciò trova conferma anche dalla visualizzazione del box plot in *Figura 31*. Un BMI più alto, spesso associato all'obesità, può aumentare il rischio di infarti cardiaci. Mantenere un peso corporeo sano attraverso una dieta equilibrata e l'esercizio fisico può contribuire a ridurre il rischio di infarti.

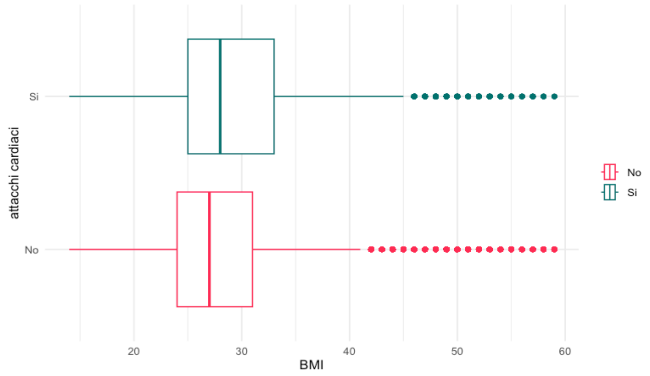

Figura 31: Box plot relativo al BMI condizionato alle condizioni cardiache

Continuando ad utilizzare l'indice BMI, questa volta si vuole analizzare se la salute percepita è coerente con il proprio indice di massa corporeo, consapevole del fatto che un valore discostante dall'intervallo centrale può portare a reali disturbi del comportamento alimentare come l'anoressia o, al contrario, l'obesità.

La *Figura 32* mostra che, effettivamente, i rispondenti che sostengono di avere una scarsa salute, hanno un BMI superiore alla media. Tuttavia, la distribuzione nell'indice di massa corporeo non si discosta in maniera netta tra le classi.

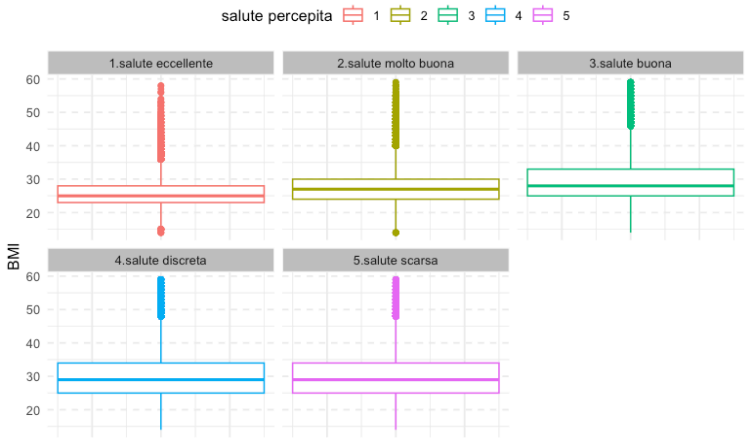

Figura 32 : box plot relativo al BMI condizionato alla salute percepita

Nella *Figura 33* si nota un elevata variabilità nei valori BMI soprattutto per alti livelli di salute.

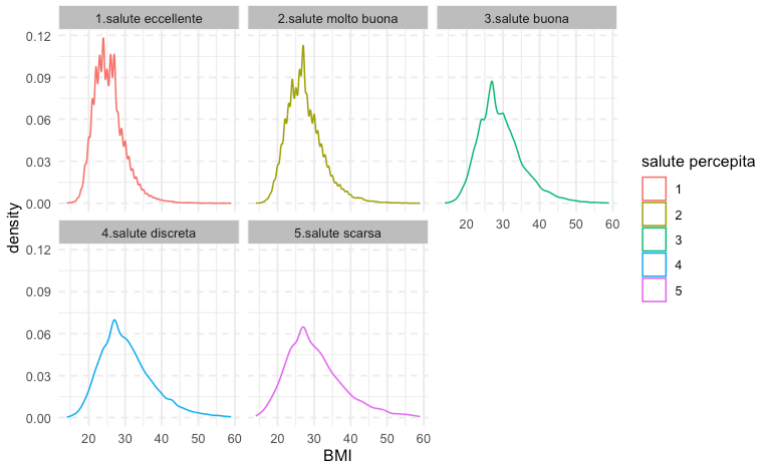

Figura 33: Density plot del BMI condizionato alla salute percepita

Anche in questo caso, si può dire che nelle diverse classi non si assiste a cambiamenti della distribuzione di densità del BMI. Tutte hanno in comune l'asimmetria positiva; inoltre, più la salute percepita si abbassa, più diminuisce la variabilità dei valori di BMI. Molto probabilmente a causa dell'aumento della consapevolezza nell'avere un'alimentazione sbagliata e quindi un disturbo del comportamento alimentare, i

rispondenti con un indice BMI elevato percepiscono oggettivamente di avere una scarsa salute fisica.

Come detto precedentemente, un reddito elevato può consentire un migliore accesso alle cure mediche e ai controlli regolari. Si verifica adesso se i controlli per il colesterolo vengono svolti regolarmente dai candidati, indipendentemente dall'ipercolesterolemia.[27]

La *Figura 34* mostra che i check vengono svolti anche da persone non soggette a sbalzi di colesterolo. Si ricorda che la maggioranza dei candidati ha un reddito elevato; pertanto, ha un'assicurazione medica che gli consente di svolgere controlli periodici.

Tuttavia si evince che la maggioranza delle persone che effettuano check al colesterolo soffre di ipercolesterolemia.

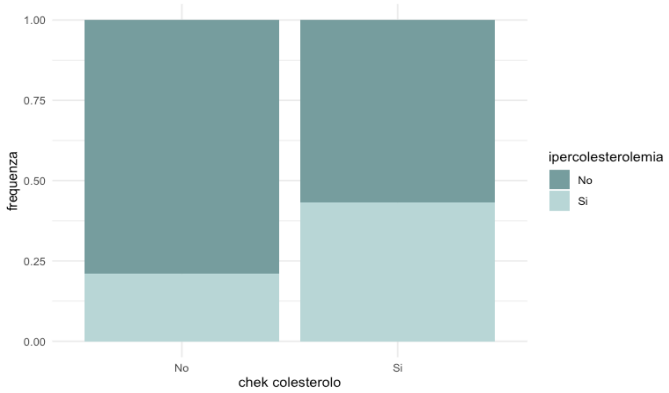

Figura 34: bar plot relativo al check per il colesterolo e ipercolesterolemia

Si verifica se un basso consumo di frutta e verdura può essere considerato un fattore di rischio per l'insorgenza dell'ictus. Si ricorda, inoltre, che solo il 4 % del campione ha contratto la patologia almeno una volta nella vita.

Dalla *Figura 35* si evince che, in generale, vi è un alto consumo di frutta e verdura tra tutti gli intervistati. Importante è analizzare invece i rispondenti che non integrano nelle loro diete la frutta e la verdura; in questo caso il numero di casi patologici aumenta.

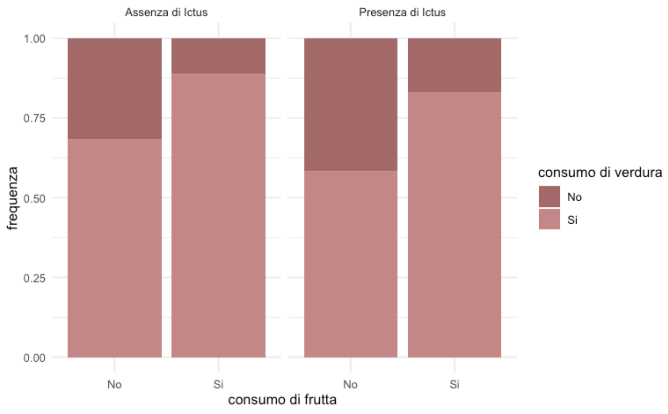

Figura 35 : bar plot relativo al consumo di frutta e verdura per pazienti colpiti da ictus

[27] L'*ipercolesterolemia* è una condizione caratterizzata dalla presenza di livelli elevati di colesterolo nel sangue.

Infine, si analizza se la carenza di denaro per pagare le visite mediche possa essere considerato un fattore di rischio per le patologie.

Costruisco una nuova variabile *patologia* che tiene conto di tutte le malattie analizzate precedentemente; la codifica è la seguente:

- Patologia = *"Si"* , se il candidato è diabetico, ha mai avuto un ictus o una qualsiasi malattia cardiaca.
- Patologia = *"No"*, altrimenti.

Dalla *Figura 36* si nota che la carenza di risorse finanziarie per pagare le visite mediche sembra correlare ad un leggero aumento nell'incidenza della patologia.

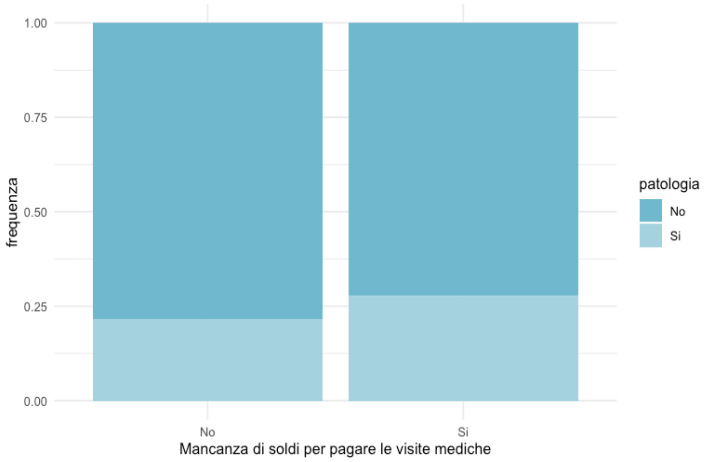

Figura 36: bar plot relativo alla carenza di denaro in relazione alla patologia

I risultati hanno mostrato che i pazienti che hanno difficoltà a sostenere i costi delle visite mediche tendono ad avere una leggera ma significativa aumento dell'incidenza della patologia rispetto a coloro che possono permettersi le spese mediche.

Nell'ambito di questa ricerca, abbiamo anche esaminato le abitudini dei pazienti diabetici. Attraverso un'analisi univariata, abbiamo valutato diversi fattori quali l'alimentazione, l'attività fisica, e l'aderenza alle terapie raccomandate. Questo ci ha permesso di ottenere una visione dettagliata delle caratteristiche e degli stili di vita dei pazienti con diabete. Successivamente, nell'analisi multivariata, abbiamo esaminato come queste abitudini interagiscano con la presenza di diabete, considerando anche altre variabili di interesse come l'età, il sesso e la presenza di altre condizioni mediche. Questa analisi ci ha fornito informazioni più approfondite sulle dinamiche complesse che coinvolgono il diabete e ha evidenziato eventuali associazioni significative tra le abitudini dei pazienti e la manifestazione del diabete.

In conclusione, l'analisi univariata e multivariata delle abitudini dei pazienti diabetici ci ha permesso di comprendere meglio il ruolo cruciale che lo stile di vita e l'aderenza alle terapie possono avere nel manifestarsi e nel gestire questa condizione medica.

3.3.3 Analisi dell'associazione tra le variabili

L'associazione tra due variabili categoriali, anche chiamata associazione o relazione tra variabili qualitative, riguarda la dipendenza o la connessione statistica tra due variabili che rappresentano categorie o classi discrete. Questa associazione può essere di grande importanza nell'ambito dell'analisi statistica, poiché fornisce informazioni sul legame tra i diversi fattori e può guidare decisioni e azioni informate. Per valutare l'associazione tra due variabili categoriali, è comune utilizzare una *tabella di contingenza* (o tabella a doppia entrata). Questa tabella mostra le frequenze con cui le diverse categorie delle due variabili si verificano contemporaneamente.

- *Le frequenze osservate* nella tabella di contingenza rappresentano il numero reale di casi che ricadono in ogni combinazione di categorie delle variabili.

- Le *frequenze attese* sono calcolate sotto l'ipotesi di indipendenza tra le variabili. In altre parole, rappresentano le frequenze che ci aspetteremmo di vedere se le variabili fossero indipendenti.

Il test che generalmente è utilizzato per valutare se l'associazione tra due variabili sia statisticamente significativa è il *test del chi-quadrato*, chiamato in questo modo dall'impiego della statistica $\mathcal{X}^2$ di Pearson. L'indice viene calcolato confrontando le frequenze osservate con le frequenze attese, sotto l'ipotesi di indipendenza tra le variabili. Maggiore è il valore dell'indice del chi-quadrato, maggiore è l'associazione tra le variabili.

Dati due caratteri qualitativi *X* e *Y*, rispettivamente con *k* e *h* modalità, essa è definita nel modo seguente:

$$\mathcal{X}^2 = \sum_{i=1}^{k}\sum_{j=1}^{h}\frac{\left(n_{ij} - n_{ij}^*\right)^2}{n_{ij}^*}$$

ossia è pari alla somma dei quadrati delle differenze tra le frequenze osservate n_{ij} e le frequenze attese n_{ij}^* , rispetto alle frequenze attese. Le frequenze attese sono date da:

$$n_{ij}^* = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$$

dove $n_{i\bullet}$ e $n_{\bullet j}$ sono le frequenze marginali.

- L'indice $\mathcal{X}^2$ è nullo se i due caratteri sono indipendenti.
- È positivo se vi è dipendenza tra i due caratteri.
- Il valore di $\mathcal{X}^2$ aumenta all'aumentare della numerosità del dataset n.

Il test si basa sull'ipotesi nulla che afferma che le variabili sono indipendenti.

Si dimostra che sotto l'ipotesi H_0 di indipendenza tra i caratteri e se $n\to\infty$, la statistica di Pearson converge ad una distribuzione del chi-quadrato con g=(k−1)(h−1) gradi di libertà:

$$\mathcal{X}^2 = \sum_{i=1}^{k}\sum_{j=1}^{h}\frac{\left(n_{ij} - n_{ij}^*\right)^2}{n_{ij}^*} \xrightarrow{d} \mathcal{X}^2{}_{(\mathrm{k}-1)(\mathrm{h}-1)}$$

Se il valore del chi-quadrato supera il valore critico (solitamente al livello di significatività del 5% o del 1%), si rifiuta l'ipotesi nulla in favore dell'ipotesi alternativa, indicando che c'è un'associazione significativa tra le variabili categoriali.

Per avere una misura di distanza che non dipenda da n viene introdotta una versione normalizzata dell'indice $\mathcal{X}^2$ ottenuta dividendo il valore dell'indice per il suo valore massimo.

$$\phi^2 = \frac{\mathcal{X}^2}{n \cdot (\min\{h,k\} - 1)}$$

Da questo indice si ottiene l'indice V di Cramér come

$$V = \sqrt{\phi^2}$$

Entrambi gli indici variano nell'intervallo [0,1]. Assumono il valore 0 per indipendenza perfetta, il valore 1 per dipendenza perfetta. In particolare:

i. $V = 0$ se i due caratteri sono indipendenti
ii. $V < 0{,}3$ se c'è una bassa dipendenza tra i caratteri
iii. $V \geq 0{,}3$ se c'è un'apprezzabile dipendenza tra i due caratteri.

L'indice di Cramér è particolarmente utile quando si confrontano tabelle di contingenza di dimensioni diverse o quando si vogliono confrontare associazioni in tabelle di dimensioni diverse. È una misura di associazione standardizzata che facilita il confronto tra diversi contesti.

Un efficace strumento per visualizzare l'associazione tra due o più variabili categoriali è rappresentato dal *Mosaic Plot legato ai Residui di Pearson.*

I residui di Pearson sono una misura utilizzata nell'analisi statistica per valutare quanto i dati osservati si discostino dai valori attesi sotto un modello specifico.

Si ottengono con la seguente formula :

$$e_{ij} = \frac{n_{ij} - n_{ij}^*}{\sqrt{n_{ij}^*}}, \quad i = 1{,}2, \ldots., k, \quad j = 1{,}2, \ldots., h$$

Un residuo positivo indica che l'osservazione è più alta di quanto ci si aspetti, mentre un residuo negativo indica che è più bassa di quanto ci si aspetti. Grandi residui di Pearson possono indicare che l'ipotesi di indipendenza non è soddisfatta e che c'è un'associazione significativa tra le variabili in esame. Ritornando al Mosaic plot, l'interpretazione dei colori è la seguente:

- *Colori freddi:* indicano una deviazione positiva significativa rispetto ai valori attesi. Ciò significa che le frequenze osservate sono maggiori di quanto ci si aspetterebbe sotto l'ipotesi di indipendenza
- Colori *caldi:* indicano una deviazione negativa significativa rispetto ai valori attesi. Le frequenze osservate sono inferiori a quanto ci si aspetterebbe secondo l'ipotesi di indipendenza.

- *Colori neutri:* indicano una deviazione non significativa rispetto ai valori attesi. Le frequenze osservate coincidono con quelle teoriche e, dunque, vi è indipendenza tra i due caratteri.

Si analizzano principalmente le coppie di variabili utilizzate precedentemente per

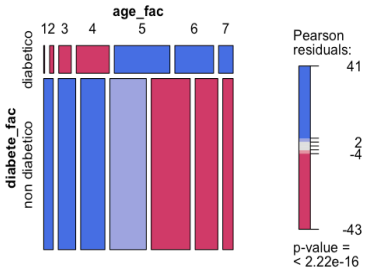

Figura 37: mosaic plot relativo al diabete ed età

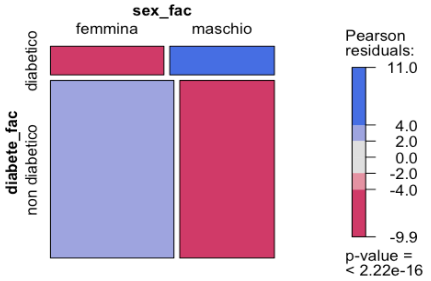

Figura 38: mosaic plot relativo al diabete e al genere

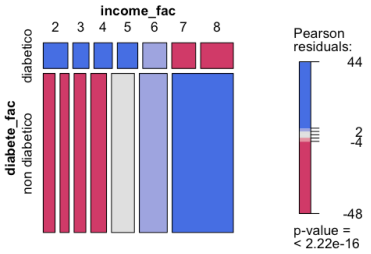

Figura 39: mosaic plot relativo al diabete e al reddito

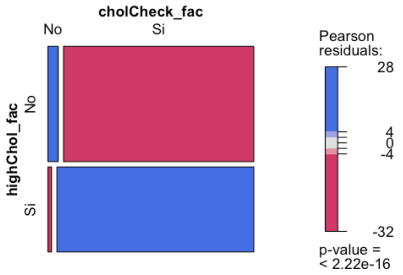

Figura 40: mosaic plot relativo al controllo medico per il colesterolo e ipercolesterolemia

l'analisi multivariata.

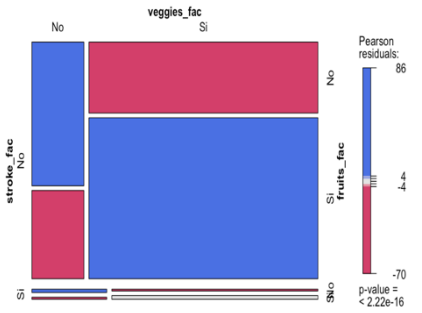

Figura 41: mosaic plot relativo all'ictus condizionato per il consumo di frutta e verdura

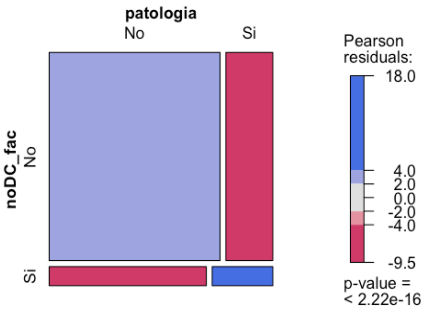

Figura 42: mosaic plot relativo alla presenza della patologia e alla scarsa quantità di denaro per pagare le visite

Tutte le variabili utilizzate per le analisi multivariate sono legate tra di loro. Infatti, osservando sia i valori di p-value, che dei residui, si può rifiutare l'ipotesi H_0 di indipendenza tra le variabili.

La *Tabella 30* mostra i valori dell'indice V di Cramér per confermare i risultati.

Tabella 30: indici V di Cramér per ogni coppia di variabili prese in considerazione

Variabili	*V di Cramer*
Diabete-età	0,18
Diabete-genere	0,16
Diabete-reddito	0,17
Check medico-livello colesterolo	0,19
Ictus-consumo frutta-verdura	0,15/0,19
Patologia- carenza di soldi	0,16

Si notano valori positivi, ma comunque inferiori alla soglia *0.3*; pertanto, secondo l'indice di Cramér vi è una bassa dipendenza tra le variabili.

3.4 Analisi del modello logit

In questo paragrafo si pone l'interesse verso l'analisi del modello logit, la cui metodologia è descritta nel *Paragrafo 2.1.1,* per modellare e analizzare la relazione tra le *features* e la variabile *target*. In questo contesto, il modello logit è uno strumento che ci permette di valutare l'effetto di variabili indipendenti, i fattori di rischio e le abitudini dei rispondenti, sulla probabilità di contrarre il diabete, consentendo di comprendere quali variabili influiscono maggiormente sulla patologia.

Prima di analizzare i risultati, viene svolta un fase preliminare essenziale, ovvero quella della *one-hot encoding* delle variabili categoriali composte da più livelli.

3.4.1 Pre-processing: gestione delle variabili categoriali

Nell'affrontare complessi problemi di machine learning, spesso ci troviamo a dover manipolare dati rappresentati attraverso variabili categoriche, le quali non possono essere direttamente trattate come dati numerici. Molti algoritmi di apprendimento automatico non possono operare direttamente sui dati categoriali, molti altri peccano di efficienza; pertanto, si prosegue con la codifica *one-hot* delle variabili.

La *one-hot encoding* è una tecnica che trasforma una variabile categoriale, composta da k livelli, in k variabili binarie. Questo processo consente di rappresentare ogni livello come un vettore binario. Ogni variabile dummy rappresenta una delle categorie, e solo una di queste variabili avrà il valore 1 per ogni osservazione, indicando a quale categoria appartiene. La one-hot encoding non solo ci permette di rappresentare chiaramente le categorie come vettori binari, ma ci consente anche di evitare possibili ambiguità interpretative o attribuzioni di erronea importanza tra le categorie. Ogni categoria è trattata in modo equo, senza implicazioni gerarchiche o ordini arbitrari.

La *"trappola delle dummy"* è una situazione indesiderata che può verificarsi quando si utilizzano variabili dummy, come quelle generate con la tecnica della One-Hot Encoding, in modelli di regressione. Essa si verifica quando le variabili dummy sono altamente correlate o linearmente dipendenti, creando problemi nella stima dei coefficienti nel modello di regressione. Questa dipendenza lineare tra le variabili è il risultato del fatto che, per una variabile categorica con k categorie, basta includere $k - 1$ variabili dummy per rappresentare completamente l'informazione.

La categoria non considerata prende il nome di *gruppo di riferimento*, scelto arbitrariamente, rispetto al quale verranno confrontati gli altri livelli.

In seguito viene riportato un esempio con la variabile *Sex*, la quale riporta informazioni sul sesso dei candidati.

$$Sex = \begin{cases} 1 = Maschio \\ 0 = Femmina \end{cases}$$

Supponiamo di voler considerare una nuova variabile dummy, costruita come il complemento ad 1 della variabile Sex_i

$$Sex_i^* = 1 - Sex_i$$

Si considera adesso un modello che tiene conto delle due variabili come predittori

$$Y_i = \beta_0 + \beta_1 X_1 + \delta Sex_i + \lambda Sex_i^* + e_i \quad \forall_i$$

In questo modo si avrà β_0 e λ che catturano lo stesso effetto, ovvero $Sex_i = 0$. Siccome si ha la stessa informazione due volte, siamo in presenza di *collinearità perfetta*, e lo stimatore dei minimi quadrati[28] non è definito.

Esistono, dunque, due modi per evitare la trappola delle variabili dummy:

i. Utilizzare $k - 1$ variabili dummy per una variabile categorica con k categorie

ii. Escludere dal modello statistico l'intercetta, che cattura l'informazione della variabile di riferimento.

3.4.2 Divisione del dataset: Train Set e Test Set

La divisione del dataset è un concetto fondamentale nell'ambito del machine learning e dell'analisi dei dati, utilizzati per addestrare e valutare i modelli predittivi in modo accurato e affidabile.

- Il *train set* è la parte del dataset utilizzata per addestrare il modello. Quest'ultimo apprende dalle caratteristiche presenti nel train set e cerca di stabilire relazioni tra gli input e la variabile target binaria al fine di predire correttamente su dati futuri.

- Il *test set* è la parte del dataset separata e non vista dal modello durante la fase di addestramento. Pertanto, viene utilizzato per valutare le prestazioni del modello dopo l'addestramento. Il test set serve a misurare quanto bene il modello generalizza su dati non osservati, fornendo una valutazione obiettiva delle prestazioni del modello.

La separazione tra train e test set aiuta a prevenire l'*overfitting*, un problema in cui il modello si adatta troppo bene al train set ma non generalizza bene su dati nuovi e non visti. Se la dimensione del train set aumenta, il modello ha più informazioni per apprendere e quindi aumenta l'adattabilità ai dati osservati; pertanto, consente una stima più affidabile dei parametri del modello. Ciò comporta però a una riduzione del test set; si hanno, quindi, meno dati per valutare le prestazioni del modello, rendendo la valutazione meno affidabile.

In generale, nasce un *trade-off* tra affidabilità del modello e valutazione delle prestazioni. Un train set più grande consente al modello di imparare meglio e adattarsi ai dati, ma può portare a una valutazione meno attendibile a causa del test set più piccolo. Al contrario, un test set più grande offre una valutazione più attendibile, ma può limitare la quantità di dati utilizzati per l'addestramento del modello. L'obiettivo è trovare un equilibrio appropriato tra queste due dimensioni in modo da garantire un buon apprendimento del modello e una valutazione affidabile delle sue prestazioni su dati non visti.

28 Lo stimatore dei minimi quadrati è una tecnica utilizzata per stimare i parametri di un modello statistico, generalmente un modello di regressione lineare. È uno dei metodi più comuni e fondamentali utilizzati per stimare i parametri in diversi tipi di modelli.

La suddivisione più comune quando si dispone di un numero sufficiente di dati, come nel caso nostro, è in percentuali fisse, generalmente 80% per il train set e 20% per il test set. La suddivisione del dataset in percentuali fisse tra il train set e il test set è spesso eseguita in modo tale che la proporzione della variabile target sia approssimativamente la stessa in entrambi i set.

3.4.3 Interpretazione dei coefficienti

Il modello addestrato sul train set comprende le features estratte dalla stepwise di tipo *"both",* la cui metodologia è descritta nel *Paragrafo 2.2.*

La *Tabella 31* mostra il vettore β dei coefficienti stimati per ogni predittore del modello. Sappiamo che con $\hat{\beta}$ si vede l'effetto delle variabili sul logit della probabilità $Y = 1$.

Tabella 31: Sintesi dei risultati del modello logit

	Estimate	*Std.Error*	*Z value*	$Pr(>\|z\|)$	
(Intercept)	-8.499378	0.119714	-70.997	< 2e-16	***
BMI	0.073381	0.001171	62.673	< 2e-16	***
I(Age = 2) = 1	0.538296	0.087125	6.178	6.47e-10	***
I(Age = 3) = 1	1.061904	0.081600	13.014	< 2e-16	***
I(Age = 4) = 1	1.441001	0.079712	18.078	< 2e-16	***
I(Age = 5) = 1	1.766076	0.079432	22.234	< 2e-16	***
I(Age = 6) = 1	1.861146	0.080178	23.213	< 2e-16	***
I(Age = 7) = 1	1.677363	0.082585	20.311	< 2e-16	***
I(Income = 5) = 1	-0.125594	0.024235	-5.182	2.19e-07	***
I(Income = 6) = 1	-0.210354	0.022927	-9.175	< 2e-16	***
I(Income = 7) = 1	-0.225732	0.023370	-9.659	< 2e-16	***
I(Income = 8) = 1	-0.360619	0.022683	-15.898	< 2e-16	***
I(Education = 4) = 1	-0.066709	0.017583	-3.794	0.000148	***
I(Education = 6) = 1	-0.090572	0.018369	-4.931	8.19e-07	***
Sex = maschio	0.254231	0.014967	16.986	< 2e-16	***
High BP = Si	0.691186	0.016582	41.684	< 2e-16	***
I(Gen Hlth = 2) = 1	0.686349	0.037187	18.457	< 2e-16	***
I(Gen Hlth = 3) = 1	1.356920	0.036353	37.326	< 2e-16	***
I(Gen Hlth = 4) = 1	1.790390	0.038865	46.067	< 2e-16	***
I(Gen Hlth = 5) = 1	1.951615	0.044588	43.770	< 2e-16	***
Diff Walk = Si	0.106475	0.018404	5.785	7.24e-09	***
Heart DS = Si	0.248202	0.019942	12.446	< 2e-16	***
High Chol = Si	0.550835	0.015304	35.992	< 2e-16	***
Hvy AC = Si	-0.757321	0.043015	-17.606	< 2e-16	***
I(Ment Hlth = 2) = 1	-0.125207	0.018870	-6.635	3.24e-11	***
I(Ment Hlth = 3) = 1	-0.114396	0.024312	-4.705	2.54e-06	***
Stroke = Si	0.201032	0.028092	7.156	8.29e-13	***

Per quanto riguarda il *BMI*, unica variabile continua del dataset, si evince che, per ogni incremento unitario vi è un aumento nel logit della probabilità di successo (diabete = Si) di 0,07, a parità di tutte le altre condizioni.

Una variazione positiva sebbene minima, confermata ampiamente dall'analisi esplorativa eseguita nel paragrafo precedente, in cui si evidenzia che valori alti di BMI sono considerati fattori di rischio per il diabete.

Si vedono adesso i coefficienti associati alle variabili binarie. Il coefficiente associato alla variabile binaria indica se la probabilità di $Y_i = 1$ sia più alta per $D_i = 1$ o per $D_i = 0$. Se il coefficiente è positivo, p_i è più alta per $D_i = 1$, se invece è negativo, p_i è più alta per $D_i = 0$. Per quanto riguarda la variabile categoriale *Age*, si notano coefficienti crescenti all'aumentare delle classi; tutti i coefficienti sono positivi ma più il candidato è anziano più il logit di p_i aumenta, ovvero c'è una propensione maggiore allo sviluppo del diabete.

Stesso discorso viene fatto con la variabile *Gen Hlth*[29], livelli alti sono associati a una salute precaria percepita. All'aumentare delle classi, si notano coefficienti crescenti, ovvero una variazione maggiore del logit di p_i.

Discorso contrario per le variabili *Income* e *Education*. Tutte le variabili indicatrici associate alle due variabili hanno un coefficiente negativo, ovvero la loro presenza porta un decremento nel logit di p_i; tale decremento cresce all'aumentare dei livelli. Livelli alti di reddito e di educazione, infatti, favoriscono allo sviluppo di una vita più sana.

La variabile *Sex,* anch'essa statisticamente significativa, ha un coefficiente positivo; ciò viene interpretato come l'essere maschio porta un incremento positivo nel logit della probabilità di successo. Discorso analogo per le altre variabili che riguardano alcuni de fattori di rischio per lo sviluppo del diabete, come *High BP* e *High Chol*[30]. La presenza delle due variabili porta un aumento del logit di p_i, come si poteva immaginare.

Ai fini dell'interpretazione, quando si lavora con variabili binarie, è importante far riferimento anche agli odds e odds-ratio. Per ottenere l'odds, bisogna considerare l'esponenziale dei coefficienti $\exp\{\hat{\beta}\}$. Se il coefficiente è positivo, $\exp\{\hat{\beta}\} > 1$ e quindi $p_i > 1 - p_i$; se il coefficiente è negativo, $0 \leq \exp\{\hat{\beta}\} < 1$ e, quindi, $p_i < 1 - p_i$.

Per quanto riguarda le variabili binarie, ricordiamo che

$$exp(\delta) = \frac{P(Y_i = 1|D_i = 1)}{P(Y_i = 1|D_i = 0)}$$

Se $exp(\delta) = 1$ significa che D_i non è statisticamente significativa poiché la presenza o l'assenza della categoria non influenza la probabilità.

[29] La variabile Gen Hlth racchiude le informazioni circa la salute percepita da parte dei candidati. Vedere l'analisi esplorativa

[30] High BP e High Chol racchiudono rispettivamente le informazioni circa l'alta pressione sanguigna e alti livelli di colesterolo dei candidati. Vedere l'analisi esplorativa

La *Tabella 32* mostra gli odds, in seguito si riportano le interpretazioni dei coefficienti.

Tabella 32: odds

	Exp(Estimate)
(Intercept)	0.0002
BMI	1.0761
I(Age = 2) = 1	1.7131
I(Age = 3) = 1	2.8919
I(Age = 4) = 1	4.2249
I(Age = 5) = 1	5.8479
I(Age = 6) = 1	6.4311
I(Age = 7) = 1	5.3514
I(Income = 5) = 1	0.8820
I(Income = 6) = 1	0.8103
I(Income = 7) = 1	0.7979
I(Income = 8) = 1	0.6972
I(Education = 4) = 1	0.9355
I(Education = 6) = 1	0.9134
Sex = maschio	1.2895
High BP = Si	1.9961
I(Gen Hlth = 2) = 1	1.9865
I(Gen Hlth = 3) = 1	3.8842
I(Gen Hlth = 4) = 1	5.9918
I(Gen Hlth = 5) = 1	7.0401
Diff Walk = Si	1.1123
Heart DS = Si	1.2817
High Chol = Si	1.7347
Hvy AC = Si	0.4689
I(Ment Hlth = 2) = 1	0.8823
I(Ment Hlth = 3) = 1	0.8919
Stroke = Si	1.2227

Si ottengono risultati coerenti con quanto detto prima.

Gli oddsratio associati alle variabili indicatrici di Age sono tutti maggiori di uno, in particolare I(Age = 6) presenta un oddsratio pari a 6,43; ciò significa che, a parità di altre condizioni, la probabilità dell'insorgere del diabete è ben 6 volte superiore nei candidati con età compresa tra i 70 e 79 anni[31] rispetto ai candidati con età minore di 30 anni (gruppo di riferimento). Ancora, si evince che la pressione sanguigna alta provoca un aumento dello sviluppo del diabete di quasi il 100%, ovvero la probabilità di contrarre il diabete nei pazienti con pressione alta raddoppia rispetto ai pazienti con pressione sanguigna normale.

Alti livelli di colesterolo e attacchi cardiaci provocano un aumento della probabilità di contrarre il diabete rispettivamente del 73% e 28%. Da ciò si nota quanto il monitoraggio del colesterolo sia importante per prevenire patologie di questo tipo.

[31] Vedere la codifica della variabile Age Par. 3.2.2

La presenza di ictus provoca un aumento del 22% circa della probabilità di contrarre il diabete, mentre una variabile che provoca, invece, un decremento della probabilità è quella legata al reddito. Anche in questa circostanza, alti livelli di reddito superiori a 75 mila dollari riducono la probabilità di contrarre il diabete di circa il 31%.

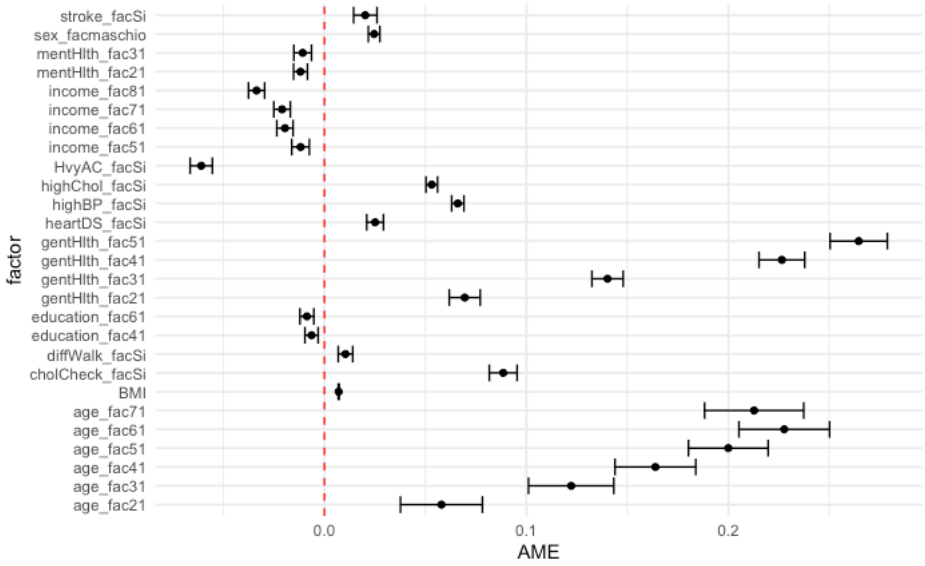

Figura 43: plot degli effetti marginali medi (AME)

Gli effetti marginali medi rappresentano la variazione media nella variabile dipendente (risposta) rispetto a una variazione di una singola unità nelle variabili indipendenti, mantenendo le altre variabili indipendenti costanti. Nel caso di variabili categoriali, gli effetti marginali medi rappresentano la variazione media nella variabile dipendente dovuta a un cambio dalla categoria di riferimento a una specifica categoria nella variabile categorica, mantenendo le altre variabili costanti.

- Se l'effetto marginale medio è positivo, un aumento di una unità nella variabile indipendente, o la presenza di una determinata categoria per variabili binarie, è associato a un aumento medio dell'outcome.

- Se l'effetto marginale medio è negativo, un aumento di una unità nella variabile indipendente, o la presenza di una determinata categoria per variabili binarie, è associato a una diminuzione media dell'outcome.

Anche in questo caso si nota come la maggior parte delle variabili provocano un aumento della probabilità di successo.

3.4.4 Validazione del modello

Dopo aver stimato i coefficienti, si procede con la validazione del modello. Per validazione del modello si intende una serie di tecniche volte a valutare l'adeguatezza e l'utilità prima di utilizzarlo con dati indipendenti per fini previsivi. In primo luogo, l'importanza di ciascun predittore viene valutato effettuando test statistici sulla significatività dei coefficienti. Dal momento in cui abbiamo preso in considerazione soltanto le variabili giudicate statisticamente significative dalla stepwise, ci aspetteremo un risultato analogo per i test statistici.

Il primo test è quello sul singolo parametro, finalizzato a valutare l'ipotesi che un particolare parametro abbia un valore specifico, generalmente pari a zero.

Sia $\hat{\beta}_k$ la stima di *massima verosimiglianza* del parametro β_k, e sia var($\hat{\beta}_k$) nota.

$$\begin{cases} H_0: \beta_k = \beta_c \\ H_1: \beta_k \neq \beta_c \end{cases}$$

La statistica test è data da

$$Z_c = \frac{\hat{\beta}_k - \beta_k}{s.e.(\hat{\beta}_k)} \sim N(0,1)$$

La regione critica associata al livello di significatività α è tale per cui se $z_c \geq z_{\frac{\alpha}{2}}$ si rifiuta l'ipotesi nulla al livello di significativa α.

Si può, inoltre, considerare un altro test statistico che valuta la significatività sul singolo parametro, la cui statistica test è ottenuta elevando al quadrato Z_c

$$W_c^2 = \frac{(\hat{\beta}_k - \beta_k)^2}{\left(s.e.(\hat{\beta}_k)\right)^2} \sim \mathcal{X}^2_{(1)}$$

Il test prende il nome di *test di Wald*, la cui statistica test tende a distribuirsi come una chi-quadrato con 1 grado di libertà.

La regione critica associata al livello di significatività α è tale per cui se $w_c^2 > \mathcal{X}^2_{(1),\alpha}$ si rifiuta l'ipotesi nulla al livello di significatività α.

La *Tabella 29,* del paragrafo precedente, raccoglie la statistiche test Z e i p-value per ogni coefficiente associato alle variabili. Come già previsto, le variabili sono tutte statisticamente significative; lo stesso risultato è ottenuto guardando al test di Wald, che si ottiene elevando al quadrato le statistiche test Z, i cui p-value rimangono invariati.

Per quanto riguarda la *bontà di adattamento*, siccome le stime sono ottenute attraverso la massima verosimiglianza, l'approccio dei minimi quadrati non può essere applicato. Per valutare l'adattabilità di un modello, sono stati sviluppati diversi indici chiamati *Pseudo-*R^2*,* i quali variano tra 0 e 1. Alti valori indicano un buon adattamento al modello, ma non è possibile interpretare i valori come nel caso dei minimi quadrati.

R^2 di Efron:

$$R_E^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$

poiché $\hat{y}_i$ è una probabilità mentre e y_i è binaria, la loro differenza non può essere letta come in un modello di regressione lineare multiplo. Esso, infatti, viene letto in termini di variabilità spiegata.

- Al denominatore c'è la variabilità totale della variabile dipendente.
- Al numeratore c'è la variabilità che non può essere spiegata dal modello.

Quindi, abbiamo la proporzione di variabilità non spiegata dal modello che, sottratta a 1, ci fornisce proprio la proporzione di variabilità spiegata. Minore è la variabilità non spiegata dal modello, più piccolo sarà il rapporto e migliore il modello.

R^2 di Mc-Fadden:

$$R_{mf}^2 = 1 - \frac{l_{modello}}{l_0}$$

dove $l^{modello}$ è la log-verosimiglianza del modello stimato, e l_0 è quella del modello nullo. Esso misura quanto bene il modello spiega la variabilità nei dati rispetto a un modello nullo. Un valore più vicino a 1 indica che il modello spiega meglio la variabilità rispetto al modello nullo; ovvero $l_0 > l_{modello}$. Più il rapporto è alto, cioè si avvicina a 1, più il modello non ha un buon adattamento.

R^2 di Cox-Snell:

$$R_{cs}^2 = 1 - (\frac{L_0}{L_{modello}})^{\frac{2}{n}}$$

In questo caso si considerano le verosimiglianze.

Se il rapporto è piccolo, vuol dire che il modello ha un buon adattamento rispetto al modello con solo l'intercetta. Il massimo di questo indice è circa

$$1 - (L_0)^{\frac{2}{n}} < 1$$

perché se il modello $L_{modello}$ prevede perfettamente la variabile di risposta, la verosimiglianza sarà 1

R^2 di Cox-Snell normalizzato (Nagelkerke):

$$R_n^2 = \frac{R_{cs}^2}{\max(R_{cs}^2)}$$

non è altro che l'R_{cs}^2 rapportato al suo valore massimo; utile per valutare l'*R^2 di Cox-Snell* tra 0 e 1.

Esso è uguale a 1, se $L_{modello} = 1$, quindi il modello si adatta bene.

Esso sarà uguale a 0, se $L_0 = L_{modello}$ e quindi il modello non si adatta bene.

La *Tabella 33* riporta i valori dei *Pseudo-R^2* calcolati sul modello stimato.

Tabella 33: Pseudo-R^2

Pseudo-R^2	
Efron	*0.19*
McFadden	*0.22*
Nagelkerke	*0.29*
CoxSnell	*0.16*

Nella pratica, raramente si assiste a livelli di Pseudo-R^2 elevati; Già vicino a 0,30/0,40 possiamo ritenerci soddisfatti. In questo caso abbiamo valori abbastanza bassi.

Tuttavia, la valutazione complessiva della bontà di adattamento del modello non dovrebbe basarsi esclusivamente sull'R^2 o sui suoi equivalenti. È sempre consigliabile valutare il modello utilizzando una combinazione di metriche di adattamento, come l'*AIC* (Criterio di Informazione di Akaike) e il *BIC* (Criterio Informativo Bayesiano).

- L'*AIC* è definito come

$$AIC = -2 \cdot \log(L) + 2 \cdot k$$

esso è un indice che permette il confronto tra più modelli. In modello ottimale è quello che presenta un AIC più piccolo.

Non tiene conto della parsimonia, ovvero il principio di preferire modelli più semplici senza compromettere la capacità del modello di spiegare i dati. Per questo motivo si guarda al BIC

- Il *BIC* è definito come

$$BIC = -2 \cdot \log(L) + k \cdot \log(n)$$

anche in questo caso il modello ottimale è quello che presenta un BIC più piccolo.

La *Tabella 34* mostra i valori di AIC e BIC del modello logit.

Tabella 34: Valori AIC e BIC modello logit

AIC	*127558.4*
BIC	*127844.5*

3.4.5 Valutazione della capacità predittiva

In questo paragrafo vengono valutate le prestazioni del modello sul test set. Si parte col verificare se il modello è in grado di prevedere adeguatamente l'appartenenza dei casi al gruppo diabete Si/No, attraverso la tabella di continenza. Per la creazione della tabella, si deve scegliere il *cutoff ottimale k* tale che

$$se\ \hat{p}_i > k \longrightarrow\ \hat{y}_i = 1\ ('Si')$$

In genere *k = 0,50*, ma la scelta del k è fortemente influenzato dallo sbilanciamento della variabile target. In seguito viene rappresentata la tabella di classificazione usando il generico *k = 0,50.*

Tabella 35: Confusion matrix k = 0.5

	Classi Vere	
Classi Previste	No	Si
No	**TN** 42627	**FN** 5894
Si	**FP** 900	**TP** 1113

In genere, la categoria 0 (*No*) è chiamata *"negativa",* mentre la categoria 1 (Si) è chiamata *"positiva".* Quindi è possibile suddividere i valori in

- *TN,* ovvero la porzione di Negativi correttamente classificati come Negativi. In termini probabilistici, si ha la probabilità di avere un vero negativo

$$P(\hat{Y}_i = 0|Y_i = 0)$$

- *FN,* ovvero la porzione di Positivi erroneamente classificati come Negativi. In termini probabilistici, si ha la probabilità di avere un falso negativo

$$P(\hat{Y}_i = 0|Y_i = 1)$$

- *FP,* ovvero la porzione di Negativi erroneamente classificati come Positivi. In termini probabilistici, si ha la probabilità di avere un falso positivo

$$P(\hat{Y}_i = 1|Y_i = 0)$$

- *TP,* ovvero la porzione di Positivi correttamente classificati come Positivi. In termini probabilistici, si ha la probabilità di avere un vero positivo

$$P(\hat{Y}_i = 1|Y_i = 1)$$

Dai valori ottenuti è possibile calcolare le misure di performance utili per la valutazione del modello *outsample.*

- *Accuracy* è la frazione di istanze correttamente classificate. Varia tra 0 e 1, più è grande più il modello è accurato

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} = \mathbf{0.866}$$

Paradosso dell'accuratezza: si ricorda che c'è un fortissimo sbilanciamento tra le classi, infatti i negativi, ovvero i non diabetici, sono l'86%[32]. Il classificatore sopra, ottiene un'accuratezza anch'essa dell'86%; pertanto, posso assicurarmi la stessa performance assegnando qualunque istanza ai negativi (classe prevalente).

- *Misclassification rate* è la frazione delle istanze misclassificate. È il complemento a 1 dell'Accuracy; pertanto, più grande è, meno il modello è accurato.

$$Err = \frac{FP+FN}{TP+TN+FP+FN} = 1\text{- } Acc = \mathbf{0.134}$$

Importante è specificare che queste due misure trattano le due sorgenti di errore (FN e FP) in modo simmetrico. Cioè classificare un vero diabetico come non diabetico ha lo stesso peso di classificare un non diabetico come tale.

- *FPR (False Positive Rate)* è la frazione dei veri negativi misclassificati. Essendo un falso allarme, più è vicino a 0, più il numero di falsi positivi diminuisce a favore dei veri negativi correttamente classificati

$$FPR = \frac{FP}{TN+FP} = 1\text{- } Specificity = \mathbf{0.021}$$

- *FNR (False Negative Rate)* è la frazione dei veri positivi misclassificati. Anch'esso varia tra 0 e 1, più è vicino a 0, più il numero di falsi negativi diminuisce a favore dei veri positivi correttamente classificati

$$FNR = \frac{FN}{TP+FN} = 1\text{- } Sensitivity = \mathbf{0.838}$$

Esiste un trade-off tra i due rate. Se uno è alto, l'altro è basso. Possono essere ricollegati all'errore di primo e secondo tipo.

Otteniamo un FNR alto a causa dello sbilanciamento delle classi. Il numero di negativi è così alto rispetto ai positivi che il modello classifica l'84% dei veri positivi in negativi. Un

[32] Vedere l'analisi esplorativa univariata

modo per ottenere un FNR più basso è sicuramente impostare un cutoff diverso da quello selezionato.

- *TPR (True Positive Rate)*, detto anche *Sensitivity* o *Recall,* è il complemento a 1 del FNR; pertanto, indica la frazione dei veri positivi correttamente classificati.

$$Sensitivity = \frac{TP}{TP+FN} = \mathbf{0.162}$$

- *TNR (True Negative Rate),* detto anche *Specificity,* è il complemento a 1 del FPR; pertanto, indica la frazione dei veri negativi correttamente classificati.

$$Specificity = \frac{TN}{TN+FP} = \mathbf{0.981}$$

- *PPV (Positive Predictive Value),* detto anche *Precision,* è la frazione dei veri positivi tra le unità classificate come positivi

$$PPV = \frac{TP}{TP+FP} = \mathbf{0.558}$$

  In tal senso, è una misura di esattezza/fedeltà del classificatore. Il 56% circa di tutti i valori classificati come positivi risultano essere stati correttamente classificati.

- *NPV (Negative Predictive Value),* è la frazione dei veri negativi tra le unità classificate come negativi

$$NPV = \frac{TN}{TN+FN} = \mathbf{0.879}$$

  L'89% di tutti i valori classificati come negativi risultano essere stati correttamente classificati.

In generale, si nota che con un $k = 0,5$ il modello tende a fare errori abbastanza evidenti nel classificare correttamente i positivi della variabile target.

Esiste un trade-off tra il *TPR* e il *FPR*: aumentando il cutoff, si riduce il TPR, mentre si riduce anche l'FPR e, di conseguenza, aumenta la *Specificity.* Ciò significa che, aumentando il limite per classificare un'osservazione come positiva, si riducono sia i veri positivi che i falsi positivi.

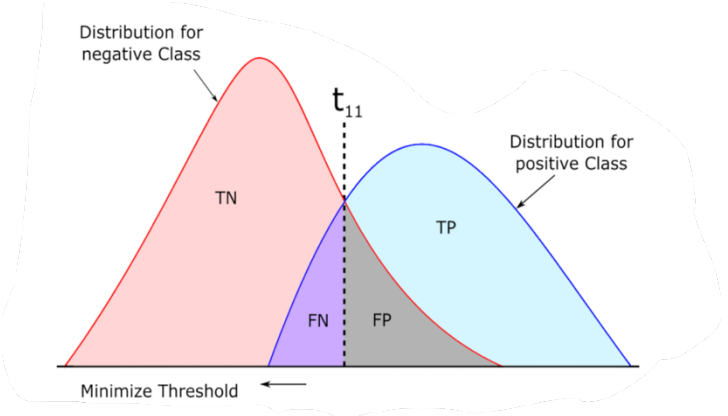

Figura 44: trade-off tra TPR e FPR

Nel nostro caso, sarebbe perfetto andare a diminuire il cutoff per aumentare il limite nel classificare un'osservazione positiva.

Il *trade-off* tra *TPR* e *FPR* è solitamente visualizzato tramite la curva *ROC (Receiver Operating Characteristic)* che, per ogni livello di *k,* rappresenta i valori di *TPR* sull'asse delle ordinate e i valori di *FPR* sull'asse delle ascisse. È possibile selezionare un punto di equilibrio appropriavo sulla curva *ROC,* dove il *TPR* e il *FPR* soddisfano meglio i requisiti del problema. Per un buon modello, la curva *ROC* cresce rapidamente, indicando che il *TPR* sulla asse y cresce più velocemente del *FPR* sull'asse x, al decrescere della soglia.

Quindi, il punto ideale sarebbe in alto a sinistra, nell'angolo del grafico, perché è il punto in cui l'*FPR è 0*, e il *TPR* è 1. Ma è una situazione non realistica.

Lungo la retta a 45 gradi ci sono i classificatori equivalenti al *random guessing*, ovvero classificatori che sparano a caso, con stessa probabilità, osservazioni in una classe o nell'altra. Sopra la retta a 45 gradi ci sono i classificatori migliori del *random guessing,* sotto quelli peggiori.

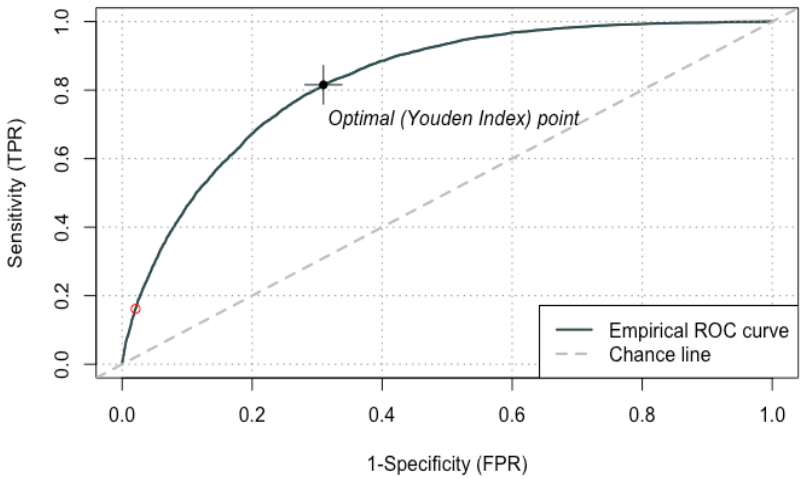

Figura 45: curva di ROC

Il puntino rosso in basso a sinistra rappresenta la combinazione di *TPR* e *FPR* ottenuto con il classificatore Bayesiano (*cutoff = 0.5)*. Per trovare il punto sulla curva *ROC* che massimizza la *Sensitivity* e la *Specificity,* il cosiddetto *Optimal (Youden Index) point*, mi avvalgo del pacchetto *"pRoc"*.

Secondo le etichette reali e le probabilità previste dal modello su dati outsample, la *Tabella 36* riporta le coordinate del punto ottimale e il relativo *cutoff.*

Tabella 36: Optimal (Youden Index) point

Cutoff	*0.12*
Specificity	*0.69*
Sensitivity	*0.82*

Dunque, sarebbe opportuno reimpostare la soglia con un cutoff più ottimale per ricreare la tabella di classificazione.

Tabella 37: Confusion matrix k = 0.12 ottimale

	Classi Vere	
Classi Previste	No	Si
No	**TN** 29689	**FN** 1252
Si	**FP** 13838	**TP** 5777

In seguito si riportano le misure di performance confrontate con quelle del classificatore Bayesiano.

Tabella 38: Confronto delle misure di performance

Cutoff	*0.50*	*0.12*
Accuracy	***0.866***	***0.702***
Misclassification rate	***0.134***	***0.298***
FPR	***0.021***	***0.318***
FNR	***0.838***	***0.178***
Sensitivity	***0.162***	***0.822***
Specificity	***0.981***	***0.682***
PPV	***0.558***	***0.295***
NPV	***0.879***	***0.960***

Dal confronto si nota principalmente che il cutoff ottimale porta un aumento della Sensitivity, che era ciò che volevamo. Sono stati ridotti i *FN* mentre, al contrario, sono aumentati i *FP.* Nel complesso il nuovo classificatore ha un'accuratezza inferiore a quello precedente ma tutto va contestualizzato in base alle esigenze specifiche.

Una metrica comunemente utilizzata per valutare le prestazioni di un modello è *l'AUC (Area Under the Curve).* Essa rappresenta l'area sottesa della curva *ROC* e fornisce una misura complessiva della capacità disciminante del modello. L'*AUC* può variare da 0 a 1, dove:

- Un'AUC pari a 1 indica un modello perfetto che classifica correttamente tutte le osservazioni (poco realistico).
- Un'AUC pari a 0.5 indica un modello che effettua previsioni casuali, essenzialmente equivalente al lancio di una moneta per prendere decisioni.

In pratica, maggiore è l'AUC, maggiore è la capacità del modello di distinguere tra le classi positive e negative. Un valore alto di AUC indica che il modello è in grado di classificare in modo efficace la maggior parte delle osservazioni positive più alte delle osservazioni negative.

L'utilità dell'AUC è particolarmente rilevante quando il dataset è sbilanciato, cioè quando una classe è molto più prevalente dell'altra. In questi casi, l'AUC fornisce un'indicazione più accurata delle prestazioni del modello rispetto all'Accuracy, che potrebbe essere influenzata in modo significativo dalla predominanza di una classe.

Tabella 39: AUC modello logit

AUC	*0.8268*

Infine si calcola l'*error rate* sul test set, un indicatore che è in grado di calcolare la proporzione di errori di classificazione rispetto al totale delle osservazioni presenti nel test set.

$$error_{test} = mean(predicted_{labels} \neq true_{labels}) = \mathbf{0.298}$$

L'espressione restituisce un vettore di valori booleani che indica se le previsioni del modello coincidono o meno con le etichette reali. Questo vettore avrà *TRUE* per ogni errore di classificazione e *FALSE* per ogni previsione corretta.

Poiché in *R*, *TRUE* è rappresentato come 1 e *FALSE* come 0, la media di questo vettore rappresenta la proporzione di errori di classificazione nel test set.

In questo caso si ottiene un $error_{test}$ basso e ciò rispecchia l'accuratezza del modello.

3.5 Analisi del modello probit

Si passa con la stima del modello probit, la cui metodologia è riportata nel *Paragrafo2.1.2*

Il modello probit, così come il modello logit, rientra nella categoria dei modelli lineari generalizzati ed usa, come variabile di *legame* per collegare la variabile di risposta ai predittori, la funzione di ripartizione della distribuzione normale standardizzata.

Utilizzando la stessa fase di pre-processing realizzata sul modello logit, sia per quanto riguarda la suddivisione del dataset, sia per la variable selection con il metodo stepwise di tipo *"both"*, possiamo passare direttamente all'interpretazione dei coefficienti.

3.5.1 Interpretazione dei coefficienti

La *Tabella 40* mostra il vettore β dei coefficienti stimati per ogni predittore del modello. Sappiamo che con $\hat{\beta}$ si vede l'effetto delle variabili sulla probabilità $Y = 1$.

Al contrario del modello logit, in cui si poteva quantificare l'effetto, per il probit possiamo solo dire se l'incremento unitario di una variabile continua, o il cambio di livello per una variabile categoriale, porta un aumento o un decremento della probabilità di successo.

- Se il $\hat{\beta}$ stimato per una determinata variabile è positivo, allora il predittore ha un effetto positivo sulla probabilità di successo.

- Se il $\hat{\beta}$ stimato per una determinata variabile è negativo, allora il predittore ha un effetto negativo sulla probabilità di successo.

Per quanto riguarda il *BMI*, il coefficiente associato è positivo; pertanto, per ogni incremento unitario, la probabilità dell'insorgere del diabete aumenta. Ciò è coerente con quanto detto in fase di analisi esplorativa e nell'interpretazione dei coefficienti per il modello logit.

Le variabili indicatrici associate alla variabile *Age* e *Gen Hlth* hanno un effetto positivo sulla probabilità di successo. Infatti, più ci si allontana dal gruppo di riferimento, equivalente rispettivamente all'essere giovani e ad un'ottima salute percepita, più la probabilità dell'insorgere del diabete aumenta.

Al contrario, le variabili indicatrici associate alla variabile *Income* e *Education* hanno un effetto negativo sulla probabilità di successo. Allo stesso modo, più ci si allontana dal gruppo di riferimento, equivalente rispettivamente a un bassissimo livello di reddito e di istruzione, più la probabilità dell'insorgere del diabete diminuisce.

Le variabili che si riferiscono a problemi legati ad altre patologie come ictus, disturbi cardiaci, alti livelli di colesterolo e di pressione sanguigna hanno, naturalmente, un impatto positivo sull'insorgenza del diabete.

Tabella 40: Sintesi dei risultati del modello probit

	Estimate	*Std.Error*	*Z value*	$Pr(>\|z\|)$	
(Intercept)	-4.4837255	0.0557920	-80.365	< 2e-16	***
BMI	0.0413803	0.0006574	62.950	< 2e-16	***
I(Age = 2) = 1	0.2067266	0.0387256	5.338	9.39e-08	***
I(Age = 3) = 1	0.4480308	0.0361529	12.393	< 2e-16	***
I(Age = 4) = 1	0.6471135	0.0350802	18.447	< 2e-16	***
I(Age = 5) = 1	0.8260432	0.0349308	23.648	< 2e-16	***
I(Age = 6) = 1	0.8839584	0.0354764	24.917	< 2e-16	***
I(Age = 7) = 1	0.7843530	0.0371315	21.124	< 2e-16	***
I(Income = 5) = 1	-0.0724832	0.0136950	-5.293	1.21e-07	***
I(Income = 6) = 1	-0.1187816	0.0128370	-9.253	< 2e-16	***
I(Income = 7) = 1	-0.1299119	0.0129696	-10.017	< 2e-16	***
I(Income = 8) = 1	-0.1990241	0.0124335	-16.007	< 2e-16	***
I(Education = 4) = 1	-0.0345808	0.0098675	-3.504	0.000457	***
I(Education = 6) = 1	-0.0499956	0.0100417	-4.979	6.40e-07	***
Sex = maschio	0.1387776	0.0082407	16.841	< 2e-16	***
High BP = Si	0.3703713	0.0088509	41.846	< 2e-16	***
I(Gen Hlth = 2) = 1	0.3185932	0.0173622	18.350	< 2e-16	***
I(Gen Hlth = 3) = 1	0.6748977	0.0171401	39.375	< 2e-16	***
I(Gen Hlth = 4) = 1	0.9309134	0.0190225	48.937	< 2e-16	***
I(Gen Hlth = 5) = 1	1.0288055	0.0230189	44.694	< 2e-16	***
Diff Walk = Si	0.0699825	0.0105993	6.603	4.04e-11	***
Heart DS = Si	0.1559693	0.0116705	13.364	< 2e-16	***
High Chol = Si	0.2989868	0.0083385	35.856	< 2e-16	***
Hvy AC = Si	-0.3899366	0.0219320	-17.779	< 2e-16	***
I(Ment Hlth = 2) = 1	-0.0709178	0.0103805	-6.832	8.38e-12	***
I(Ment Hlth = 3) = 1	-0.0704117	0.0137969	-5.103	3.34e-07	***
Stroke = Si	0.1222527	0.0165048	7.407	1.29e-13	***

Anche per il modello probit possiamo visualizzare gli effetti marginali medi, ciò che cambia è che si deve considerare la funzione di densità della v.c. normale standardizzata.

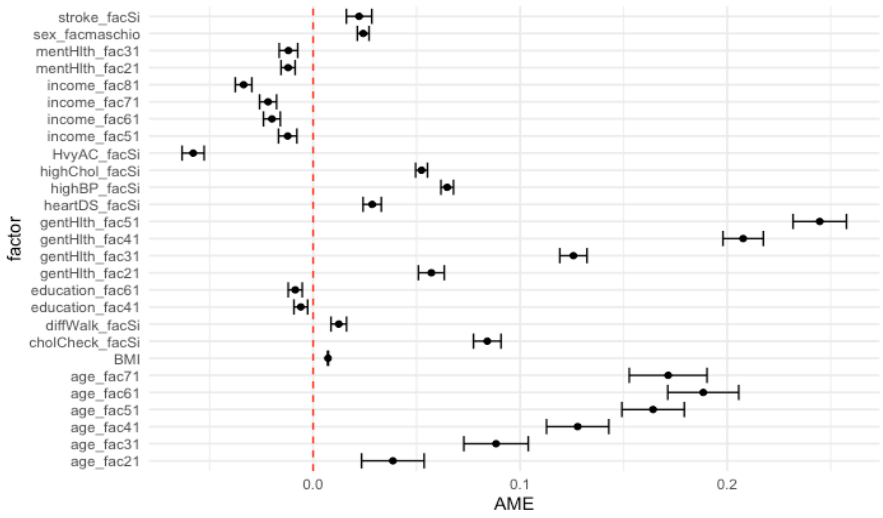

Figura 45: plot degli effetti marginali medi

L’interpretazione dei risultati è sempre la stessa, gli effetti marginali medi rappresentano la variazione media nella variabile di risposta, rispetto a una variazione di una singola unità nelle variabili indipendenti, mantenendo le altre variabili indipendenti costanti. Nel caso di variabili categoriali, gli effetti marginali medi rappresentano la variazione media nella variabile dipendente dovuta a un cambio dalla categoria di riferimento a una specifica categoria nella variabile categorica, mantenendo le altre variabili costanti. Le uniche variabili categoriali, il cui cambio di categoria produce un effetto negativo sulla probabilità di successo sono le variabili indicatrici associati al reddito, all’educazione e alla salute mentale dei candidati.

3.5.2 Validazione del modello

Dopo aver stimato i coefficienti, si procede con la validazione del modello, ugualmente a quello logit. Si parte col valutare la significatività delle variabili prese in considerazione per l’addestramento del modello probit. Otteniamo, attraverso la funzione *"confint( )"* , gli intervalli di confidenza per i parametri stimati del modello probit. Essi forniscono una stima della variabilità dei parametri stimati e aiutano a quantificare l’incertezza associata a tali stime.

Dalla *Figura 46* si evince che nessun intervallo di confidenza contiene il valore 0, ciò significa che le variabili sono statisticamente significative. I risultati sono coerenti con i p-value visti nella *Tabella 40* del paragrafo precedente, riportante la sintesi delle statistiche del modello probit stimato.

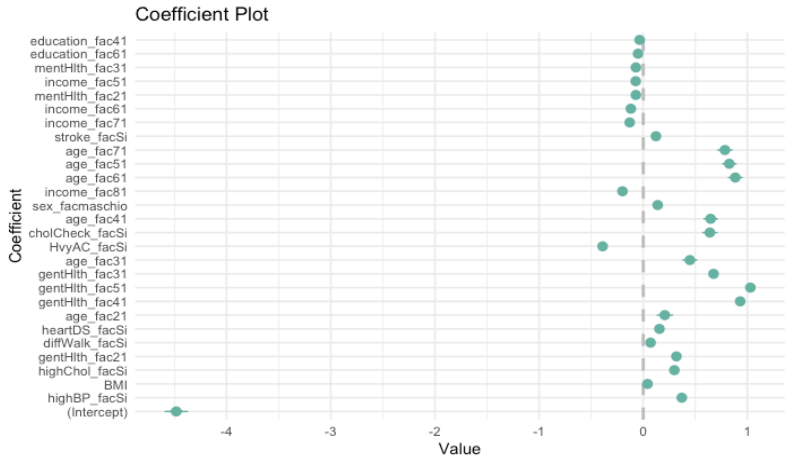

Figura 46: intervalli di confidenza per i coefficienti stimati

Per quanto riguarda la *bontà di adattamento*, è possibile riportare i valori associati a ciascun *Pseudo-*R^2. I valori sono identici a quelli ottenuti per il modello logit.

Tabella 41: Pseudo-R^2

Pseudo-R^2	
Efron	*0.19*
McFadden	*0.22*
Nagelkerke	*0.29*
CoxSnell	*0.16*

Otteniamo adesso i criteri di validazione, utili per il confronto tra i modelli che sarà effettuato nei paragrafi successivi.

Tabella 42: Valori AIC e BIC modello probit

AIC	*127289.4*
BIC	*127575.5*

3.5.3 Valutazione della capacità predittiva

Come per il logit, vengono valutate le prestazioni del modello probit sul test set.

In questo caso, verrà prima rappresentata la curva di *ROC* e, in seguito, la tabella di classificazione usando il *cutoff ottimale,* ovvero quello equivalente allo *Youden index.*

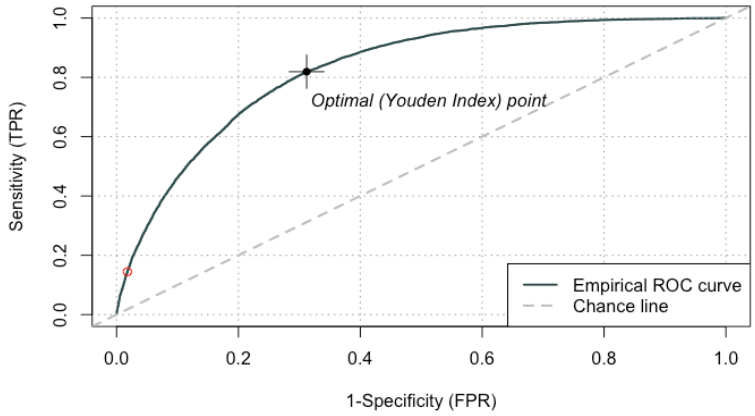

Figura 47: curva ROC

Il puntino rosso in basso a sinistra rappresenta la combinazione di *TPR* e *FPR* ottenuto con il classificatore Bayesiano.

La *Tabella 43* mostra il valore *AUC* calcolato per il modello probit.

Tabella 43: AUC modello probit

AUC	*0.8272*

Come per il logit, si riportano le coordinate del punto ottimale e il relativo *cutoff*

Tabella 44: Optimal (Youden Index) point

Cutoff	*0.13*
Specificity	*0.69*
Sensitivity	*0.82*

Il punto ottimale che massimizza la Specificity e la Sensitivity ha le stesse coordinate di quello ottenuto sul modello logit, mentre il cutoff è cambiato di un centesimo.

Otteniamo adesso la tabella di classificazione utilizzando il cutoff ottimale

Tabella 45: Confusion matrix k = 0.13 ottimale

	Classi Vere	
Classi Previste	No	Si
No	**TN** 30604	**FN** 1403
Si	**FP** 12923	**TP** 5626

In seguito si riportano le misure di performance confrontate con quelle del classificatore Bayesiano.

Tabella 46: Confronto delle misure di performance

Cutoff	*0.50*	*0.13*
Accuracy	***0.866***	***0.717***
Misclassification rate	***0.134***	***0.283***
FPR	***0.018***	***0.297***
FNR	***0.855***	***0.200***
Sensitivity	***0.145***	***0.800***
Specificity	***0.982***	***0.703***
PPV	***0.567***	***0.303***
NPV	***0.877***	***0.956***

Come nel caso logit, anche per il probit ritorna la problematica nel classificare correttamente i positivi della variabile target. Con il cutoff ottimale (k = 0.13), aumenta il limite per classificare un'osservazione come positiva, ma ciò produce un aumento sia dei veri positivi, che dei falsi positivi.

In generale la scelta del cutoff ottimale sembra ridurre l'accuratezza del classificatore che però, come già sappiamo, è un indice influenzato dallo sbilanciamento delle classi, per cui è irrilevante. La *Sensitivity* aumenta di molto a sfavore della *Specificity*, ma era ciò che volevamo.

Infine si calcola l'*error rate* sul test set, un indicatore che è in grado di calcolare la proporzione di errori di classificazione rispetto al totale delle osservazioni presenti nel test set.

$$error_{test} = mean(predicted_{labels} \neq true_{labels}) = \mathbf{0.284}$$

3.6 Analisi del modello C-loglog

Prima di terminare con il confronto tra i modelli, è importante fornire uno sguardo al modello *C-loglog*, la cui metodologia è riportata nel *Paragrafo 2.1.3*. Questo modello è spesso utilizzato quando si affrontano classi sbilanciate, problemi in cui la variabile di risposta è rara o presenta una bassa prevalenza. Siccome sfrutta una funzione di distribuzione asimmetrica come *legame* per collegare la variabile di risposta ai predittori, il modello c-loglog può modellare in modo più flessibile le code della distribuzione delle probabilità. In situazioni in cui la classe di interesse è rara, la funzione c-loglog può aiutare a massimizzare l'informazione disponibile sui casi rari, migliorando la stima dei parametri associati a tali casi.

Si procede con l'addestramento del modello utilizzando lo stesso train set sulle features selezionate attraverso il solito metodo *Stepwise* di tipo *"both"*.

3.6.1 Interpretazione dei coefficienti

Si ricorda che nel caso del modello cloglog, la probabilità è data da

$$p_i = \exp(-\exp(x_i'\beta))$$

quindi i coefficienti sono l'effetto su

$$\hat{\beta} = \log(-\log(\hat{p}_i))$$

Per cui l'interpretazione è un po' più complicata.

- Per quanto riguarda il *BMI*, unica variabile continua del dataset, si evince che, per ogni incremento unitario vi è un aumento nel log(-log(p_i)) della probabilità di successo (diabete = Si) di 0,06, a parità di tutte le altre condizioni.
  Una variazione positiva sebbene minima, confermata ampiamente dall'analisi esplorativa eseguita nel paragrafo precedente, in cui si evidenzia che valori alti di BMI sono considerati fattori di rischio per il diabete.

- Si vedono adesso i coefficienti associati alle variabili binarie. Il coefficiente associato alla variabile binaria indica se la probabilità di $Y_i = 1$ sia più alta per $D_i = 1$ o per $D_i = 0$. Se il coefficiente è positivo, p_i è più alta per $D_i = 1$, se invece è negativo, p_i è più alta per $D_i = 0$.
  Per quanto riguarda la variabile categoriale *Age*, si notano coefficienti crescenti all'aumentare delle classi; tutti i coefficienti sono positivi ma più il candidato è

anziano più log(-log(p_i)) aumenta, ovvero c'è una propensione maggiore allo sviluppo del diabete.

Stesso discorso viene fatto con la variabile *Gen Hlth,* livelli alti sono associati a una salute precaria percepita. All'aumentare delle classi, si notano coefficienti crescenti, ovvero una variazione maggiore del log(-log(p_i)).

Discorso contrario per le variabili *Income* e *Education*. Tutte le variabili indicatrici associate alle due variabili hanno un coefficiente negativo, ovvero la loro presenza porta un decremento nel log(-log(p_i)); tale decremento cresce all'aumentare dei livelli. Livelli alti di reddito e di educazione, infatti, favoriscono allo sviluppo di una vita più sana. La variabile *Sex,* anch'essa statisticamente significativa, ha un coefficiente positivo; ciò viene interpretato come l'essere maschio porta un incremento positivo nel logit della probabilità di successo.

Discorso analogo per le altre variabili che riguardano alcuni de fattori di rischio per lo sviluppo del diabete, come *High BP* e *High Chol.* La presenza delle due variabili porta un aumento del log(-log(p_i)), come si poteva immaginare

Tabella 47: Sintesi dei risultati del modello cloglog

	Estimate	*Std.Error*	*Z value*	$Pr(>\|z\|)$	
(Intercept)	-7.83923	0.11228	-69.817	< 2e-16	***
BMI	0.05676	0.00093	61.028	< 2e-16	***
I(Age = 2) = 1	0.56203	0.08378	6.708	1.97e-11	***
I(Age = 3) = 1	1.05771	0.07850	13.474	< 2e-16	***
I(Age = 4) = 1	1.38708	0.07686	18.047	< 2e-16	***
I(Age = 5) = 1	1.65647	0.07661	21.622	< 2e-16	***
I(Age = 6) = 1	1.73142	0.07713	22.447	< 2e-16	***
I(Age = 7) = 1	1.57524	0.07899	19.941	< 2e-16	***
I(Income = 5) = 1	-0.09005	0.02008	-4.484	7.33e-06	***
I(Income = 6) = 1	-0.16598	0.01928	-8.609	< 2e-16	***
I(Income = 7) = 1	-0.17627	0.01981	-8.899	< 2e-16	***
I(Income = 8) = 1	-0.30008	0.01952	-15.375	< 2e-16	***
I(Education = 4) = 1	-0.05719	0.01465	-3.903	9.49e-05	***
I(Education = 6) = 1	-0.07739	0.01573	-4.921	8.61e-07	***
Sex = maschio	0.21663	0.01267	17.097	< 2e-16	***
High BP = Si	0.62689	0.01501	41.771	< 2e-16	***
I(Gen Hlth = 2) = 1	0.69611	0.03571	19.491	< 2e-16	***
I(Gen Hlth = 3) = 1	1.31749	0.03479	37.866	< 2e-16	***
I(Gen Hlth = 4) = 1	1.66084	0.03648	45.524	< 2e-16	***
I(Gen Hlth = 5) = 1	1.77481	0.04015	44.203	< 2e-16	***
Diff Walk = Si	0.07931	0.01523	5.207	1.92e-07	***
Heart DS = Si	0.17776	0.01600	11.111	< 2e-16	***
High Chol = Si	0.47787	0.01336	35.765	< 2e-16	***
Hvy AC = Si	-0.68558	0.03947	-17.368	< 2e-16	***
I(Ment Hlth = 2) = 1	-0.09932	0.01597	-6.219	4.99e-10	***
I(Ment Hlth = 3) = 1	-0.08512	0.01979	-4.302	1.70e-05	***
Stroke = Si	0.15076	0.02208	6.827	8.64e-12	***

Se consideriamo l'esponenziale delle stime, non otteniamo l'effetto sull'odds, come nel modello logit, ma l'effetto su

$$\exp\{\hat{\beta}\} = (-\log(\hat{p}_i))$$

Se vogliamo, invece, l'effetto sulle probabilità, si considera

$$\exp\{-exp\{\hat{\beta}\}\} = \hat{p}_i$$

Un altro modo sarebbe quello di andare a considerare gli effetti marginali medi

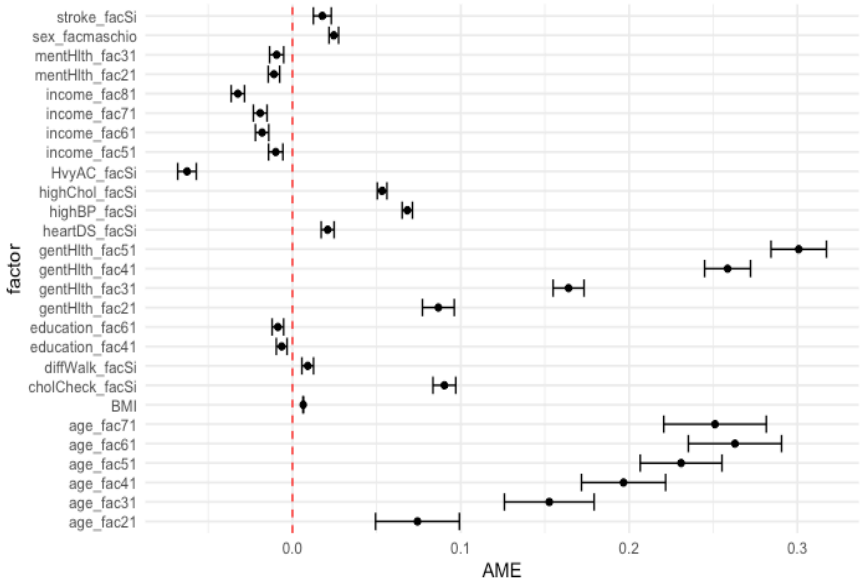

Figura 48: Effetti marginali medi modello c-loglog

La probabilità media di essere diabetico è maggiore per i candidati con un reddito inferiore; ad esempio, i candidati con un reddito pari a 75 mila dollari hanno una probabilità minore del 3,2% rispetto ai candidati con un reddito sotto ai 15 mila (gruppo di riferimento). Ancora, la probabililtà di essere diabetico è maggiore del 26% nei candidati con un età media compresa tra 70 e 79 anni, rispetto ai candidati con un età inferiore ai 30 anni. L'aumento unitario dell'indice di massa corporeo, porta un aumento della probabilità dell'insorgere del diabete del 0,6%, tenendo costanti le altre condizioni. Chi soffre di pressione alta ha un aumento del 7% del rischio di avere il diabete, rispetto a chi non ne soffre; lo stesso discorso va fatto per chi soffre di alti livelli di colesterolo, la cui percentuale scende al 5%.

Dai dati si può evidenziare che, a parità di altre condizioni, l'aumento di colesterolo influisce di più sull'insorgere del diabete rispetto all' aumento della pressione sanguigna.

3.6.2 Validazione del modello

Dalla *Tabella 47* del Paragrafo precedente, in cui sono riportate le sintesi dei risultati del modello addestrato, notiamo bassissimi p-value che comportano il rifiuto dell'ipotesi H_0 del Test Z su singolo parametro; pertanto, tutte le variabili sono statisticamente significative. La *Figura 49* mostra gli intervalli di confidenza per i coefficienti stimati. Il grafico rappresenta un ulteriore conferma della significatività delle features.

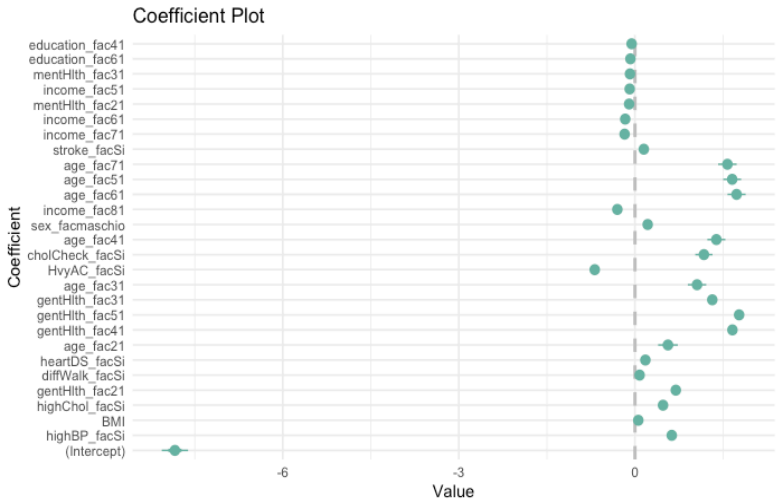

Figura 49: intervalli di confidenza per i coefficienti stimati

Come per il logit e probit, si riportano i valori associati a ciascun Pseudo-R^2. Anche in questo caso i valori sono identici rispetto a quelli calcolati per il modello logit e probit.

Tabella 48: Pseudo-R^2

Pseudo-R^2	
Efron	*0.18*
McFadden	*0.22*
Nagelkerke	*0.29*
CoxSnell	*0.16*

Si ottengono adesso i criteri di validazione, utili per il confronto tra i modelli che sarà effettuato successivamente.

Tabella 49: Valori AIC e BIC modello c-loglog

AIC	*128064.7*
BIC	*128350.8*

3.6.3 Valutazione della capacità predittiva

Attraverso la curva ROC stabiliamo il cutoff ottimale, che nel nostro contesto sarebbe quello massimizza la Sensitivity avendo comunque un valore Specificity accettabile.

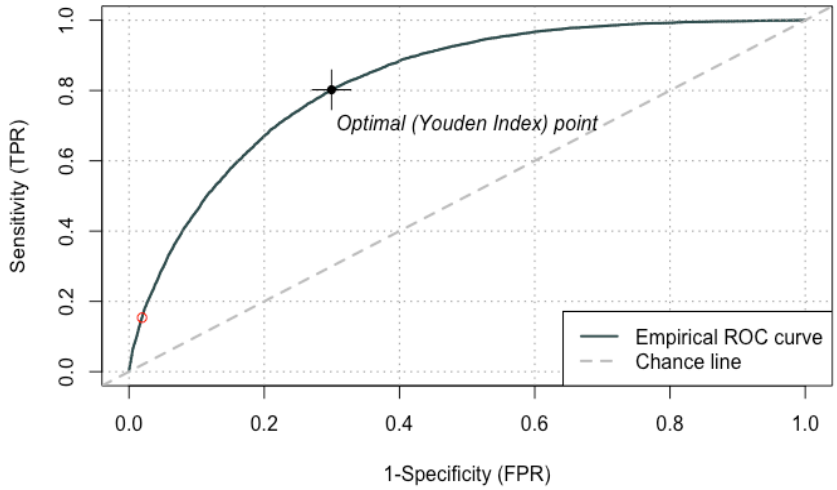

Figura 50: curva ROC

Il punto rosso rappresenta la combinazione di TPR e FPR ottenuto attraverso il classificatore Bayesiano che, come si ricordiamo, stabilisce un cutoff pari a 0.5; ciò significa che tratta le due classi della variabile target in maniera equiprobabile.

La *Tabella 50,* invece, mostra il cutoff ottimale e la combinazione ottima tra Sensitivity e FPR (1-Specificity).

Tabella 50: Optimal (Youden Index) point

Cutoff	*0.13*
Specificity	*0.70*
Sensitivity	*0.80*

In seguito si riporta il valore *AUC* calcolato per il modello c-loglog.

Tabella 51: AUC modello c-loglog

AUC	*0.826*

Calcolato il cutoff ottimale, si passa con la realizzazione della tabella di classificazione.

Tabella 52: Confusion matrix k = 0.13 ottimale

	Classi Vere	
Classi Previste	No	Si
No	**TN** 30725	**FN** 1435
Si	**FP** 12802	**TP** 5594

In seguito si riportano le misure di performance confrontate con quelle del classificatore Bayesiano.

Tabella 53: Confronto delle misure di performance

Cutoff	*0.50*	*0.13*
Accuracy	***0.866***	***0.712***
Misclassification rate	***0.134***	***0.288***
FPR	***0.019***	***0.294***
FNR	***0.846***	***0.204***
Sensitivity	***0.154***	***0.796***
Specificity	***0.981***	***0.706***
PPV	***0.563***	***0.304***
NPV	***0.878***	***0.956***

La Sensitivity, come al solito, è cambiata scegliendo un cutoff inferiore a 0.50; pertanto, il numero di falsi negativi diminuisce.

Infine si calcola l'*error rate* sul test set.

$$error_{test} = mean(predicted_{labels} \neq true_{labels}) = \mathbf{0.282}$$

3.7 Confronto tra i modelli stimati

Nell'analisi dei dati relativi alle abitudini dei rispondenti, è essenziale valutare con precisione l'effetto delle variabili indipendenti sulle probabilità dell'insorgere del diabete. A tal fine, tre modelli di regressione ampiamente utilizzati - Probit, Logit e Cloglog - sono stati applicati per modellare le probabilità dell'evento successo, associata alla presenza della patologia. Questi modelli, benché differenti nella loro funzione di legame e interpretazione, cercano di fornire una rappresentazione accurata delle relazioni sottostanti. Nel presente paragrafo, esamineremo in dettaglio i risultati emersi da ciascuno di questi modelli, confrontandone le prestazioni e l'adattabilità al contesto specifico in esame.

Innanzitutto si parte col confronto dei *Pseudo-R^2,* che aiutano a selezionare il modello che meglio si adatta ai dati. Inoltre, consentono di valutare quale modello spiega in modo più efficace le variazioni nella variabile dipendente.

Tabella 54: Confronto dei Pseudo-R^2 nei tre modelli

Pseudo-R^2	*Logit*	*Probit*	*C-loglog*
Efron	***0.19***	***0.19***	***0.18***
McFadden	***0.22***	***0.22***	***0.22***
Nagelkerke	***0.29***	***0.29***	***0.29***
Cox-Snell	***0.16***	***0.16***	***0.16***

Dalla *Tabella 54* si evince che nella specifica analisi condotta, i tre modelli forniscono un livello simile di adattamento e spiegazione della varianza rispetto alla variabile dipendente. Questa coincidenza potrebbe essere attribuita a diverse ragioni:

- La stessa selezione delle variabili predittive, mediante la stepwise di tipo both, potrebbe aver contribuito a creare un'equivalenza nei valori di *Pseudo-R^2* tra i modelli, indicando che le variabili coinvolte giocano un ruolo significativo in tutti e tre i modelli.

- Nonostante la distribuzione sbilanciata della variabile target, altre variabili nel modello potrebbero bilanciare l'effetto e consentire risultati di *Pseudo-R^2* simili tra i modelli.

Questa convergenza ci invita a considerare ulteriori analisi per determinare altri fattori che possono influenzare l'equivalenza tra i modelli. Inoltre, per validare la bontà di adattamento, bisogna sempre fare prima i conti con il principio di Parsimonia.

Per questo motivo, si confrontano i criteri di validazione.

Tabella 55: Confronto dei criteri di validazione nei tre modelli

	Logit	*Probit*	*C-loglog*
AIC	***127558.4***	***127289.4***	***128064.7***
BIC	***127844.5***	***127575.5***	***128350.8***

Nel processo di selezione del modello più appropriato tra Probit, Logit e Cloglog, abbiamo utilizzato due criteri di informazione diffusi: l'*Akaike Information Criterion* (AIC) e il *Bayesian Information Criterion* (BIC). Questi criteri valutano la qualità del modello bilanciando l'adattamento ai dati con la complessità del modello.

- L'AIC cerca di minimizzare la perdita di informazione nel modello e può scegliere modelli più complessi se ciò migliora l'adattamento. Tuttavia, penalizza la complessità meno pesantemente rispetto al BIC.

- Il BIC, invece, punisce la complessità più severamente rispetto all'AIC, promuovendo la selezione di modelli più parsimoniosi.

In questo caso, si sceglie il modello con valori di AIC e BIC più bassi.

Considerando AIC e BIC in modo congiunto, possiamo dedurre che il modello ***probit*** è il modello che meglio bilancia adattamento e complessità, rappresentando quindi la scelta più appropriata nel nostro contesto.

Terminato il confronto sulla bontà di adattamento, adesso si cerca di individuare il modello con una maggiore capacità predittiva. Confrontando i modelli in base alla loro capacità predittiva, possiamo identificare quale modello riesce a minimizzare l'error rate sul test set, ottenendo così previsioni più accurate e affidabili.

Si parte col confronto delle misure di performance, prendendo in considerazione, per tutti e tre i modelli, i risultati ottenuti mediante il *cutoff ottimale*.

Tabella 56: Confronto delle misure di performance nei tre modelli

Cutoff	*Logit*	*Probit*	*C-loglog*
Accuracy	***0.702***	***0.717***	***0.712***
Misclassification rate	***0.298***	***0.283***	***0.288***
FPR	***0.318***	***0.297***	***0.294***
FNR	***0.178***	***0.200***	***0.204***
Sensitivity	***0.822***	***0.800***	***0.796***
Specificity	***0.682***	***0.703***	***0.706***
PPV	***0.295***	***0.303***	***0.304***
NPV	***0.960***	***0.956***	***0.956***

Il cutoff ottimale ha permesso di massimizzare la *Sensitivity* avendo comunque una *Specificity* accettabile.

"Ma perché è così importante?"

La misura di Sensitivity è di fondamentale importanza nell'ambito dei dati medici e in molti altri contesti, specialmente quando si tratta di diagnosi e valutazione delle prestazioni di un modello predittivo. Essa misura la capacità di un modello di identificare correttamente i veri casi positivi di una determinata condizione medica.

Un falso negativo si verifica quando il modello indica che un paziente non ha la malattia quando in realtà ce l'ha. La sensitivity aiuta a ridurre il numero di falsi negativi, garantendo che i pazienti con la condizione medica non vengano erroneamente trascurati o

sottovalutati. Per questo motivo è importante assicurarsi che il modello abbia un valore elevato di *TPR.* Spesso però, bisogna fare i conti con l'altra "faccia" del trade-off; una bassa Specificity significa che il modello è più incline a classificare erroneamente come positivi (malati) individui che in realtà non sono affetti dalla malattia. Ciò comporta un aumento dei falsi positivi, che possono portare a un eccessivo allarme e preoccupazione per i pazienti. L'aumento dei falsi positivi può portare a un fenomeno noto come *"overdiagnosis"*[33]*,* in cui vengono diagnosticati e trattati pazienti che in realtà non hanno la condizione. Questo può portare a trattamenti non necessari (*overtreatment),* con conseguente spreco di risorse mediche. La faccia oscura della luna è popolata da tutte le conseguenze negative di essere "etichettati" come malati (*labeling effect*), dai rischi legati a test diagnostici e trattamenti non necessari, dallo spreco di risorse economiche che potrebbero essere utilizzate in maniera più appropriata.

Il progresso tecnologico ha determinato un progressivo aumento della Sensibility analitica sia dei test di laboratorio (in grado di rilevare concentrazioni sieriche sempre più basse), sia di imaging (capaci di identificare lesioni sempre più piccole). Questa evoluzione, se da un lato ha portato a valori prossimi al 100% la sensibilità clinica dei test diagnostici (capacità di identificare i veri malati), dall'altro ne ha enormemente diminuito la specificità (capacità di escludere i soggetti sani). Di conseguenza, se è sempre meno probabile che un test diagnostico risulti falsamente negativo in soggetti malati, il numero di falsi positivi cresce parallelamente all'evoluzione tecnologica. Di conseguenza vengono diagnosticate patologie sempre più lievi che continuano ad essere trattate con gli stessi approcci terapeutici delle forme moderate-severe, contribuendo a sovrastimare l'efficacia dei trattamenti.

Nel nostro contesto, il modello che ci assicura un livello di *Sensitivity* più accurato è quello ***logit***. Per quanto riguarda la misura di accuratezza, data la classe target così sbilanciata, è opportuno guardare all'*AUC.*

Tabella 57: Confronto dei valori AUC

	Logit	*Probit*	*C-loglog*
AUC	***0.8268***	***0.8272***	***0.8260***

Il modello con un *AUC* leggermente maggiore è il modello ***probit***; pertanto, è il modello con una maggiore capacità di distinguere tra le classi positive e negative.

Non ci resta che confrontare gli error rate sul test set per decretare il modello che presenta un accuratezza predittiva migliore.

Tabella 58: Confronto dell' error rate sul test set

	Logit	*Probit*	*C-loglog*
$error_{test}$	***0.298***	***0.284***	***0.282***

[33] Riferimenti website https://www.evidence.it/articolodettaglio/209/it/359/overdiagnosis-la-faccia-oscura-del-progresso-tecnologico/articolo

Il confronto tra gli error rate consente di identificare quale modello ha la performance migliore e più accurata nell'ambito specifico di interesse. L'error rate aiuta, inoltre, ad individuare modelli che mantengono una buona performance su diversi set di dati, indicando robustezza e stabilità rispetto alle variazioni nei dati di input.

Per questo motivo il modello ***c-loglog*** presenta un error rate leggermente minore rispetto agli altri due; ricordiamo che il modello c-loglog è spesso utilizzato quando si affrontano classi sbilanciate.

3.8 Cross-Validation

La *Cross-Validation* è una tecnica fondamentale utilizzata per valutare le prestazioni di un modello predittivo in modo robusto ed efficiente. In questo paragrafo, in seguito a una dettagliata spiegazione su cosa rappresenta la cross-validation, si desidera ottenere ulteriori conferme circa la validità dei risultati ottenuti attraverso questa metodologia.

La cross-validation è un metodo di ricampionamento dei dati utile per valutare la capacità di generalizzazione dei modelli predittivi e prevenire l'overfitting. Esistono molte varianti di Cross-validation e per comprenderne l'utilità è necessario introdurre il concetto di Bias-variance tradeoff. Tale aspetto si ritrova nella stima dell'errore, infatti, si può dimostrare che l'errore è costituito da una parte irriducibile che è rappresentata dalla varianza della popolazione e da una parte riducibile che dipende dal bias e dalla varianza. Dunque, l'obiettivo è individuare il giusto compromesso tra bias e variance tale da garantire il minimo errore.

Per avere una maggiore chiarezza sui passi da seguire è possibile descrivere un algoritmo per tale strumento:

Algoritmo di base per la Cross-Validation

Algorithm 1: Cross Validation

input: data set $\mathbb{X}_n = \{(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)\}$

for $s = 1, \ldots, S$ **do**

Split: partiziona casualmente i dati in $\mathbb{X}_{(s)}^{train}$ e $\mathbb{X}_{(s)}^{test}$

Train: stima $\hat{f}_s(\bullet)$ su $\mathbb{X}_{(s)}^{train}$

Predict: prevedi l'outcome y in $\mathbb{X}_{(s)}^{test}$ usando $\hat{f}_s(\bullet)$

Test: calcola $\bar{L}_s \leftarrow$ media sul $\mathbb{X}_{(s)}^{test}$

end

return

$$\text{err}_{\text{CV}} \leftarrow \sum_{s=1}^{S} \bar{L}_s$$

Le versioni di cross-validation più utilizzati sono la *K-fold cross-validation (KFCV)* e la *leave one out cross validation (LOOCV),* in quanto permettono di ottenere test set non *overlapping*. Tali versioni permettono di gestire in modo differente il bias-variance tradeoff.

- Nella *K-folds Cross Validation* il dataset iniziale viene suddiviso in *K* parti o "fold" di dimensioni approssimativamente uguali tali che

$$Fold_1 \cap Fold_2 \cap \ldots \cap Fold_K = 0$$

  Il numero di iterazioni è *S* che coincide con *K*, il numero dei fold. Per ogni iterazione, uno dei fold viene utilizzato come test set e gli altri *K-1* fold vengono usati come set di addestramento. Generalmente, se ci troviamo all'interazione *s,* il fold utilizzato per il test set è il $Fold_s$, per cui ad ogni step si avranno sempre test set diversi. La grandezza del train e del test set è determinata da *k,* infatti, se ad ogni step *k-1* fold fungono da train set, significa che il numero di dati assegnati al set di addestramento è pari a $1 - \frac{n}{k}$, mentre la frazione di dati assegnata al test set è pari a $\frac{n}{k}$.

- La *Leave One Out Cross Validation*, invece, è il caso estremo della *K-folds* quando il numero di split e, di conseguenza, il numero di iterazioni, coincide con *n,* la numerosità campionaria. In questo caso all'iterazione *s,* il $Fold_s$ è rappresentato da una sola osservazione del dataset, mentre il train set sarà composto dalle restanti *n-1* osservazioni del dataset. Quindi, ad ogni step si avranno sempre test set diversi.

La questione dell'*overlapping* è fondamentale; si preferisce avere train set simili piuttosto che test set simili perché se i test set si somigliassero la correlazione tra gli errori sarebbe inevitabile e quindi comporterebbe un aumento della varianza. Invece, se i train set si somigliassero non necessariamente accadrebbe ciò, in quanto la covarianza è dovuta dalla somiglianza tra i previsori e questi potrebbero non essere legati linearmente. Questo comporterebbe assenza di legame lineare e quindi varianza più piccola, il quale si traduce in un errore più piccolo.

Per bilanciare il trade-off tra bias e varianza, si utilizza il parametro K.

- Per valori di *K* elevati, si ottengono train set di grandi dimensioni e quindi questo migliorerebbe la stima di $\hat{Y}(\bullet)$, ovvero diminuirebbe il bias, ma allo stesso tempo si ridurrebbe la dimensione del test set e quindi si ridurrebbe la qualità delle stime, ovvero aumenterebbe la varianza;

- Per valori di *K* piccoli si avrà esattamente l'andamento opposto, ovvero aumenterebbe il bias, ma allo stesso tempo diminuirebbe la variabilità.

Dalle fonti scientifiche emerge che i valori ottimali per K sono 5,10 e *n*. La scelta di un determinato valore di K dipenderà dal risultato che si desidera ottenere. Se si vuole usare la Cross-validation per validazione, ovvero scegliere il modello migliore si usa la *LOOCV*, mentre se si vuole selezionare il valore dell'iperparametro migliore si usa la *KFCV*.

3.8.1 Risultati della *Leave-one-out C.V.*

Siccome si desidera ottenere ulteriori conferme per decretare quale è il modello con una capacità predittiva maggiore, si usa la Leave One Out Cross-Validation. Essa ci permette di valutare le prestazioni di un modello predittivo in modo robusto ed efficiente essenzialmente perché le prestazioni medie ottenute attraverso *K* iterazioni sono più affidabili rispetto a una singola suddivisione train-test, riducendo l'impatto della casualità nella selezione dei dati. Date le elevate dimensioni del dataset, eseguire la LOOCV su tutte le osservazioni potrebbe richiedere molto tempo computazionale. Pertanto, spesso si opta per una strategia più efficiente, eseguendo la cross-validation su un subset rappresentativo del dataset.

Tabella 59: Confronto dei risultati della LOOCV sui tre modelli

	Logit	*Probit*	*C-loglog*
Accuracy	***0.865***	***0.864***	***0.867***
K	***0.243***	***0.216***	***0.242***

L'Accuracy è la proporzione di previsioni corrette rispetto al totale delle previsioni fatte dal modello. Indica la proporzione di osservazioni correttamente classificate, ovvero il rapporto tra il numero di osservazioni correttamente previste e il totale delle osservazioni.

Il modello da preferire è, quindi, il ***cloglog*** anche se ci conferisce una *Sensitivity* minore del modello ***logit***.

CONCLUSIONI

In conclusione, il presente studio ha fornito una visione approfondita e dettagliata sul diabete mellito, concentrandosi sull'analisi esauriente delle abitudini dei pazienti affetti da questa patologia. Durante l'esplorazione dei dati, è emerso che fattori come l'indice di massa corporea, l'età e la presenza di altre patologie cardiache giocano un ruolo significativo nell'insorgenza del diabete. Questi risultati enfatizzano l'importanza di considerare attentamente tali variabili per una gestione efficace della malattia.

Nel valutare i modelli statistici, la sensitivity si è dimostrata un indicatore cruciale, specialmente nell'ambito medico. La sensitivity, rappresentando la capacità di identificare correttamente i veri positivi, risulta fondamentale per garantire una diagnosi accurata e tempestiva tra i pazienti affetti da diabete. La regressione logistica ha mostrato la maggiore sensitivity tra i modelli analizzati, sottolineando la sua importanza nel riconoscimento preciso dei pazienti diabetici.

D'altra parte, il modello probit ha dimostrato una migliore adattabilità ai dati, evidenziata dai valori più elevati di pseudo R^2 e dai criteri di validazione. D'altra parte, il modello cloglog, essendo appositamente sviluppato per affrontare classi sbilanciate, ha minimizzato l'error rate sul test set attraverso la LOOCV, confermando la sua efficacia predittiva in un contesto in cui la variabile target presenta uno sbilanciamento, come previsto.

In sintesi, questa ricerca non solo ha contribuito a una comprensione più approfondita del diabete e delle abitudini dei pazienti diabetici, ma ha anche sottolineato l'importanza cruciale di selezionare il modello statistico più adatto per analizzare e prevedere accuratamente in questo contesto medico. Come abbiamo visto, l'impiego di diversi modelli statistici nella valutazione di un fenomeno complesso come il diabete e le abitudini dei pazienti diabetici è essenziale per ottenere una comprensione completa e dettagliata del problema in esame. Ogni modello ha i suoi punti di forza e di debolezza, che li rendono più adatti a specifici aspetti dell'analisi. Utilizzando una pluralità di modelli e, quindi, un approccio poliedrico, si può ottenere una visione più completa e bilanciata dell'argomento di studio. Integrando i risultati, è possibile ottenere un quadro complessivo delle relazioni tra le variabili e delle previsioni accurate. I risultati ottenuti costituiscono una solida base per ulteriori indagini e per lo sviluppo di strategie personalizzate per la gestione e la prevenzione del diabete.

RIFERIMENTI

[1] Baron AD, Schaeffer L, Shragg P, Kolterman OG. Role of hyperglucagonemia in maintainance of increased rates of hepatic glucose output in type II diabetics. Diabetes 1987; 36:274-83.

[2] Hamaguchi T, Fukushima H, Uehara M, Wada S, Shirotani T, Kishikawa H, et al. Abnormal glucagon response to arginine and its normalization in obese hyperinsulinemic patients with glucose intolerance: importance of insulin action onpancreatic alpha cells. Diabetologia 1991; 34:801-6.

[3] Rhodes CJ. Type 2 diabetes - a matter of beta-cell life and death? Science 2005; 307:380-4.

[4] Poitout V, Robertson RP. Minireview: secondary beta-cell failure in type 2 diabetes - a convergence of glucotoxicity and lipotoxicity. Endcorinology 2002; 143:339-42.

[5] Drucker DJ. Glucagon-like peptides: regulators of cell proliferation, differentiation, and apoptosis. Mol Endocrinol. 2003 Feb;17(2):161-71. Review

[6] Gareth E. Lim and Patricia L. Brubaker. Glucagon-Like Peptide 1 Secretion by the L-Cell. The View from Within. Doi: 10.2337/db06-S020 Diabetes December 2006 vol. 55 no. Supplement 2 S70-S77.

[10] Center for disease control and prevention, website https://www.cdc.gov/about/

[11] Behavioral Risk Factor Surveillance System, website: https://www.cdc.gov/brfss/index.html

[12] Bolasco, S., (2002). Analisi multidimensionale dei dati: metodi, strategie e criteri d'interpretazione. Carocci Editore.

[13] James, G., Witten, D., Hastie, T. & Tibshirani, R., (2013). An introduction to statistical learning (Vol. 112). Springer, second edition.

[14] Fabbris, L., (1997). Statistica multivariata. Analisi esplorativa dei dati. McGraw-Hill.

[15]Riferimenti website
https://www.evidence.it/articolodettaglio/209/it/359/overdiagnosis-la-faccia-oscura-del-progresso-tecnologico/articolo

[17] https://www.msdmanuals.com/it-it/casa/disturbi-ormonali-e-metabolici/diabete-mellito-dm-e-disturbi-del-metabolismo-degli-zuccheri-nel-sangue/diabete-mellito-dm

[18] https://www.salute.gov.it/imgs/C_17_pubblicazioni_1885_allegato.pdf

[19] https://www.santagostino.it/it/santagostinopedia/diabete

[20] https://www.simg.it/Riviste/rivista_simg/2008/02_2008/8.pdf