UNIVERSITÀ DEGLI STUDI
DI SALERNO

Università degli Studi di Salerno

Dipartimento di Informatica

Statistica e Analisi dei Dati

Progetto del corso

Analisi del Dataset UCI "Estimation of Obesity Levels Based on Eating Habits and Physical Condition"

Studente: Salvatore Di Martino
Matricola: NF22500114

Anno Accademico 2025-2026

Indice

1 Introduzione

In questo lavoro utilizziamo il dataset *Estimation of Obesity Levels Based On Eating Habits and Physical Condition*, reso disponibile tramite l'archivio UCI Machine Learning Repository [1]. Il dataset raccoglie informazioni su individui provenienti da alcuni Paesi dell'America Latina (in particolare Messico, Perù e Colombia), con l'obiettivo di stimare il livello di obesità a partire da abitudini alimentari e condizioni fisiche.

Le osservazioni comprendono variabili relative a caratteristiche anagrafiche e fisiche (ad esempio età, altezza, peso), abitudini alimentari (frequenza di consumo di determinati cibi, numero di pasti giornalieri, spuntini), stili di vita (attività fisica, uso di dispositivi tecnologici, consumo di alcolici, abitudine al fumo) e contesto familiare. A partire da queste informazioni viene definita una variabile di risposta binaria, che distingue due gruppi di soggetti sulla base del loro livello di peso corporeo.

Il dataset è sbilanciato rispetto alla variabile target, in quanto uno dei due gruppi di interesse è rappresentato da una quota sensibilmente inferiore di osservazioni rispetto all'altro. Inoltre, una parte dei dati è stata generata sinteticamente mediante tecniche di oversampling (con l'ausilio di strumenti di data mining), mentre la restante porzione è costituita da risposte raccolte direttamente dagli utenti tramite una piattaforma web.

Nel complesso, il dataset non presenta valori mancanti ed è stato pensato per supportare diversi tipi di compiti di apprendimento automatico, quali classificazione, regressione e clustering. Le analisi statistiche ed esplorative presentate in questo elaborato verranno condotte utilizzando il software R, che offre strumenti avanzati per la gestione dei dati, la visualizzazione e l'applicazione di metodi di modellazione.

2 Obiettivi

3 Introduzione al Dataset

Il dataset utilizzato nel presente studio è composto da 2111 osservazioni e 16 variabili esplicative, alle quali si aggiunge la variabile di risposta binaria `target`. Tale variabile distingue due gruppi di individui: il valore `1` identifica i soggetti classificati come *Overweight Level I*, mentre il valore `0` aggrega tutte le rimanenti categorie di peso corporeo definite nella versione estesa del dataset originale (Insufficient Weight, Normal Weight, Overweight Level II, Obesity Type I, Obesity Type II e Obesity Type III).

Il dataset nella versione originale contiene sette classi distinte di livello di obesità, etichettate sulla base dell'*Indice di Massa Corporea* (BMI), calcolato tramite la formula:

$$\text{BMI} = \frac{\text{Peso (kg)}}{\text{Altezza (m)}^2}$$

e successivamente categorizzato secondo le linee guida dell'Organizzazione Mondiale della Sanità (OMS).

La distribuzione delle etichette mostra una forte asimmetria: soltanto 290 osservazioni appartengono alla classe minoritaria (`1`), pari a circa il 13.7% del totale, rendendo il dataset sensibilmente sbilanciato.
Il dataset è pensato per essere utilizzato in molteplici compiti di data mining, tra cui classificazione, predizione, clustering e analisi delle associazioni, grazie alla presenza di attributi sia numerici sia categorici.

3.1 Features

Le variabili incluse nel dataset coprono aspetti anagrafici, comportamentali e relativi alle abitudini alimentari e allo stile di vita degli individui. La selezione delle feature deriva da un'analisi di letteratura sui fattori riconosciuti come associati all'insorgenza dell'obesità e ai rischi cardiovascolari. Di seguito si riporta un riepilogo delle feature, coerente con la struttura del questionario utilizzato per la raccolta dati:

- **Gender** (Categorica): genere dell'individuo; valori: *Female*, *Male*.
- **Age** (Continua): età in anni.
- **Height** (Continua): altezza in metri.
- **Weight** (Continua): peso in chilogrammi.
- **family_history_with_overweight** (Binaria): presenza di casi di sovrappeso in famiglia; valori: *yes*, *no*.
- **FAVC** (Binaria): consumo frequente di cibi ad alto contenuto calorico; valori: *yes*, *no*.

- **FCVC** (Continua): frequenza di consumo di verdure durante i pasti; valori: *never - 0*, *sometimes - 1*, *always - 2*.
- **NCP** (Continua): numero di pasti principali giornalieri; valori: *One - 1*, *Two - 2*, *Three - 3*, *More than three - 4*.
- **CAEC** (Categorica): consumo di cibo tra un pasto e l'altro; valori: *no*, *Sometimes*, *Frequently*, *Always*.
- **SMOKE** (Binaria): abitudine al fumo; valori: *yes*, *no*.
- **CH2O** (Continua): quantità d'acqua consumata quotidianamente; valori: *Less than a liter - 1*, *Between 1 and 2L - 2*, *More than 2L - 3*.
- **SCC** (Binaria): monitoraggio dell'assunzione calorica giornaliera; valori: *yes*, *no*.
- **FAF** (Continua): frequenza dell'attività fisica settimanale; valori: *I do not have - 0*, *1 or 2 days - 1*, *2 or 4 days - 2*, *4 or 5 days - 3*.
- **TUE** (Continua): tempo di utilizzo quotidiano di dispositivi tecnologici; valori: *0-2 hours - 0*, *3-5 hours - 1*, *More than 5 hours - 2*.
- **CALC** (Categorica): frequenza del consumo di alcolici; valori: *no*, *Sometimes*, *Frequently*, *Always*.
- **MTRANS** (Categorica): mezzo di trasporto maggiormente utilizzato; valori: *Automobile*, *Bike*, *Motorbike*, *Public_Transportation*, *Walking*.

3.2 Processo di acquisizione dei dati

La raccolta iniziale dei dati è avvenuta tramite una piattaforma web, attraverso la quale utenti anonimi provenienti da Colombia, Perù e Messico hanno compilato un questionario sulle proprie abitudini alimentari e condizioni fisiche. In questa fase sono stati acquisiti 485 record, sui quali è stato condotto un processo di pulizia che ha comportato la rimozione di dati mancanti, anomali o non validi, oltre a una fase di normalizzazione necessaria per la successiva applicazione di tecniche di data mining.
Poiché la distribuzione iniziale delle classi risultava fortemente sbilanciata (con alcune categorie dell'obesità scarsamente rappresentate), gli autori hanno applicato il metodo *SMOTE* (Synthetic Minority Over-sampling Technique) tramite la piattaforma *Weka* al fine di generare nuove osservazioni sintetiche nelle classi minoritarie.
Dopo il bilanciamento, il dataset ha raggiunto l'attuale dimensione di 2111 record. La presenza di un'ampia componente sintetica (pari al 77% del totale) è un elemento dichiarato dagli autori e va considerata nell'interpretazione dei risultati, poiché consente un miglior addestramento dei modelli ma introduce una struttura meno rappresentativa rispetto a un campionamento interamente naturale. Il dataset finale non presenta valori mancanti.

3.3 Preprocessing

Dopo l'importazione del dataset in R viene effettuata una fase preliminare di preprocessing per verificarne l'adeguatezza rispetto alle analisi successive. Il dataset non presenta valori mancanti, caratteristica confermata anche dopo il caricamento, e le variabili categoriche vengono automaticamente convertite in oggetti di tipo `factor`. Particolare attenzione è rivolta alle variabili `FAF`, `TUE`, `CH2O`, `FCVC` e `NCP`: pur essendo descritte come continue nella documentazione originale, esse derivano da scale a risposta chiusa con pochi livelli interi e vanno quindi interpretate come misure ordinali. Per preservare questa natura e mantenerne l'utilizzabilità nelle analisi descrittive, vengono trattate come variabili numeriche ordinali. Le variabili `Weight` e `Height`, invece, mantengono la loro forma continua originale senza necessità di trasformazioni.

4 Statistica descrittiva univariata

4.1 Indici di sintesi

Gli indici di sintesi costituiscono strumenti fondamentali per riassumere in modo compatto l'informazione contenuta in un insieme di dati numerici. Il loro scopo è quello di facilitare la comprensione delle principali caratteristiche della distribuzione, riducendo la complessità del dataset a pochi valori significativi che ne rappresentano gli aspetti più rilevanti.
Porremo maggiore attenzione sulle feature `Age`, `Height` e `Weight`, siccome le altre feature numeriche derivano da scale a risposta chiusa con pochi livelli interi.

4.1.1 Indici di centralità

Le misure di centralità descrivono il punto attorno al quale i dati tendono a concentrarsi e forniscono quindi una prima indicazione sulla posizione della distribuzione. Di seguito si riporta la tabella 1 relativa agli indici principali — valore minimo, massimo, media e mediana — calcolati sulle variabili numeriche del dataset.

Variabile	**Min**	**Max**	**Media**	**Mediana**
Age	14.000	61.000	24.316	23.000
Height	1.450	1.980	1.702	1.700
Weight	39.000	173.000	86.586	83.000
FCVC	1.000	3.000	2.423	2.000
NCP	1.000	4.000	2.688	3.000
CH2O	1.000	3.000	2.015	2.000
FAF	0.000	3.000	1.007	1.000
TUE	0.000	2.000	0.665	1.000

Tabella 1: Statistiche descrittive univariate delle variabili numeriche.

La distribuzione della variabile `Age` mostra una coda verso destra, evidenziando una forte asimmetria positiva: la media risulta infatti maggiore della mediana. Una dinamica analoga, seppur meno marcata, è riscontrabile anche per la variabile `Weight`. Al contrario, la variabile `Height` presenta media e mediana molto simili, suggerendo una distribuzione più simmetrica. Le Figure 1, 2 e 3 riportano gli istogrammi delle tre variabili, consentendo una visualizzazione immediata della forma delle loro distribuzioni. Oltre alla semplice ispezione grafica, è stata calcolata anche la *skewness* come indice quantitativo di asimmetria, utile per valutare in modo oggettivo la presenza di code più o meno pronunciate rispetto alla media.

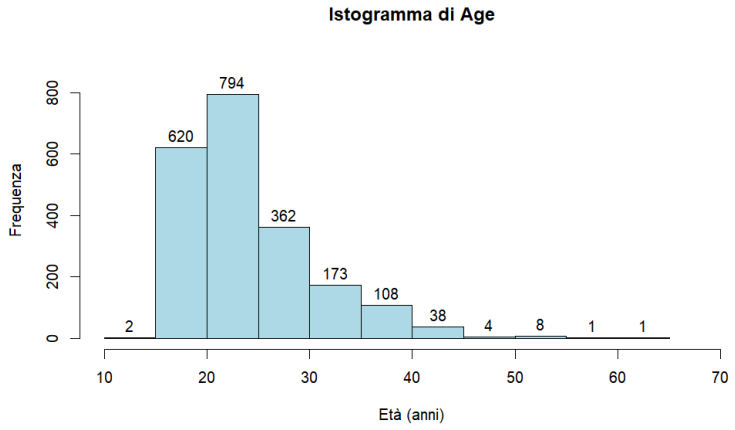

Figura 1: Istogramma della variabile `Age` - **Skewness 1,52**

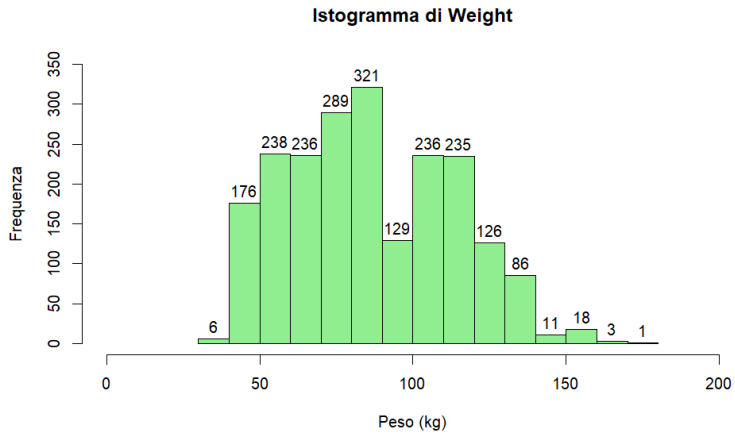

Figura 2: Istogramma della variabile `Weight` - **Skewness 0,26**

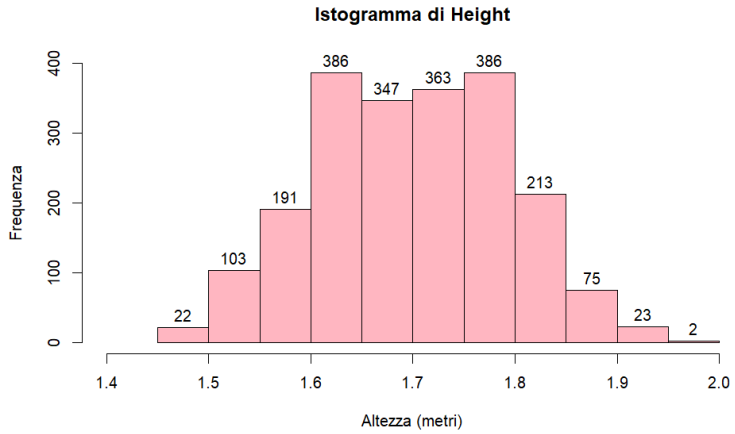

Figura 3: Istogramma della variabile `Height` - **Skewness -0,01**

4.1.2 Indici di dispersione

Gli indici di dispersione descrivono quanto i valori di una variabile si distanziano dal loro centro, integrando le misure di posizione con informazioni sulla variabilità interna dei dati. Essi permettono di valutare quanto una distribuzione sia concentrata o, al contrario, quanto presenti valori eterogenei. Nel presente studio sono stati considerati quattro indicatori principali: la varianza, che misura la dispersione quadratica rispetto alla media; la deviazione standard, sua radice quadrata e indice di più immediata interpretazione; il coefficiente di variazione (CV), che rapporta la deviazione standard alla media consentendo un confronto tra variabili con scale diverse; e infine lo scarto interquartile (IQR), che quantifica l'ampiezza della fascia centrale del 50% dei dati, risultando meno sensibile alla presenza di valori anomali. La Tabella 2 riporta tali misure per le variabili `Age`, `Height` e `Weight`.

Variabile	**Varianza**	**Deviazione standard**	**CV**	**Scarto interquartile**
Age	40.412	6.357	0.261	6.000
Height	0.009	0.093	0.055	0.138
Weight	685.978	26.191	0.302	41.957
FCVC	0.341	0.584	0.241	1.000
NCP	0.656	0.810	0.301	0.000
CH2O	0.474	0.689	0.342	0.000
FAF	0.802	0.895	0.890	2.000
TUE	0.454	0.674	1.014	1.000

Tabella 2: Indici di dispersione per le variabili numeriche del dataset.

La variabilità risulta particolarmente elevata per la variabile `Weight`, che presenta una deviazione standard pari a 26,19 kg e un coefficiente di variazione vicino al 30%, indicando una dispersione relativamente ampia dei valori attorno alla media. Anche la variabile `Age` mostra una variabilità moderata, con una deviazione standard di 6,36 anni e un CV di circa il 26%, segnalando una distribuzione dei valori meno concentrata rispetto all'altezza ma comunque non eccessivamente dispersa. Al contrario, la variabile `Height` evidenzia una variabilità molto ridotta: la deviazione standard è pari a 0,09 metri (circa 9 cm) e il coefficiente di variazione si attesta intorno al 5,5%. Anche lo scarto interquartile, pari a 0,14 metri (circa 14 cm), conferma la maggiore omogeneità dei valori. Tali valori numericamente contenuti derivano dal fatto che l'altezza è espressa in metri; tuttavia, il coefficiente di variazione, essendo un indice adimensionale che mette in relazione deviazione standard e media, fornisce comunque una misura coerente e comparabile della reale variabilità della distribuzione.

4.2 Barplot

Per le restanti caratteristiche, costituite da variabili categoriche e da variabili numeriche che derivano da scale ordinali a pochi livelli, risulta particolarmente utile analizzare la distribuzione delle frequenze mediante rappresentazioni grafiche a barre.

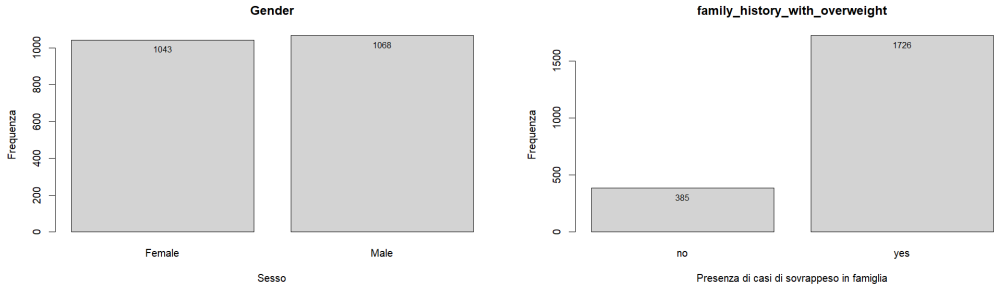

Figura 4: Barplot delle variabili `Gender` e `family_history_with_overweight`.

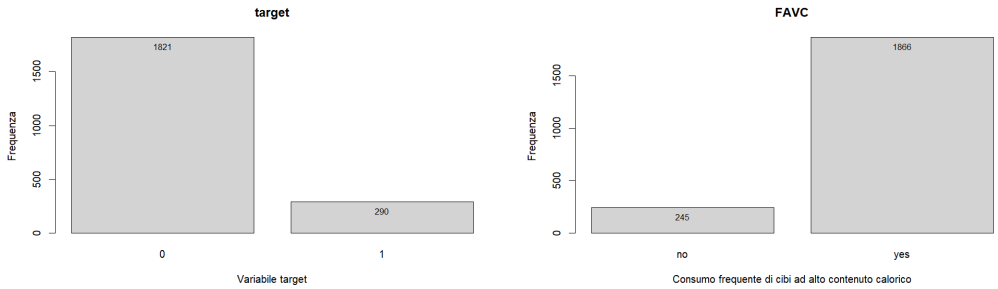

Figura 5: Barplot delle variabili `target` e `FAVC`.

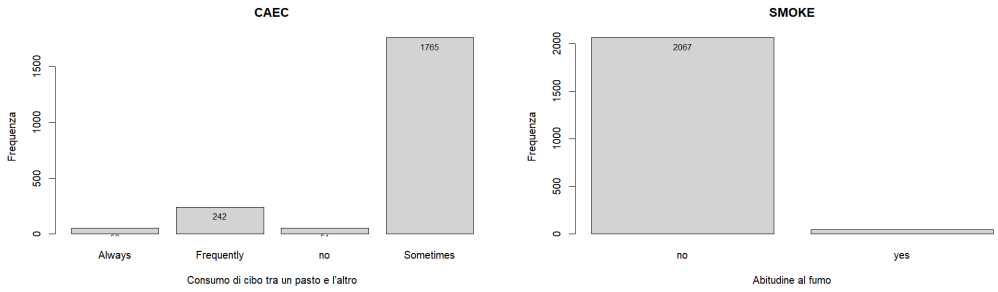

Figura 6: Barplot delle variabili `CAEC` e `SMOKE`.

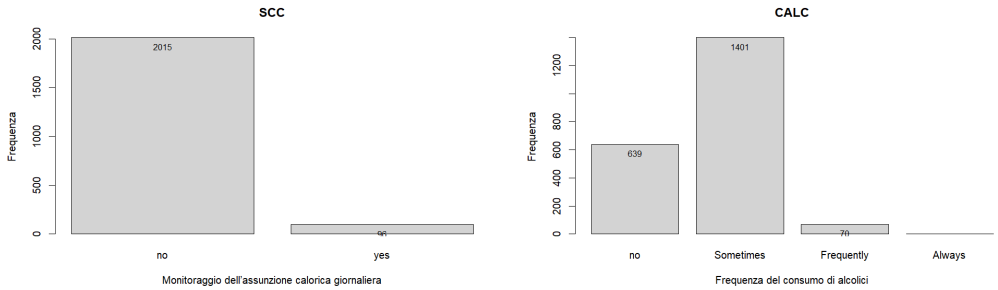

Figura 7: Barplot delle variabili `SCC` e `CALC`.

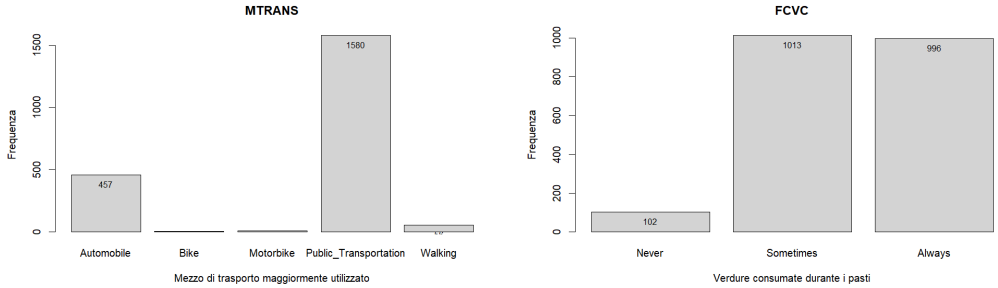

Figura 8: Barplot delle variabili `MTRANS` e `FCVC`.

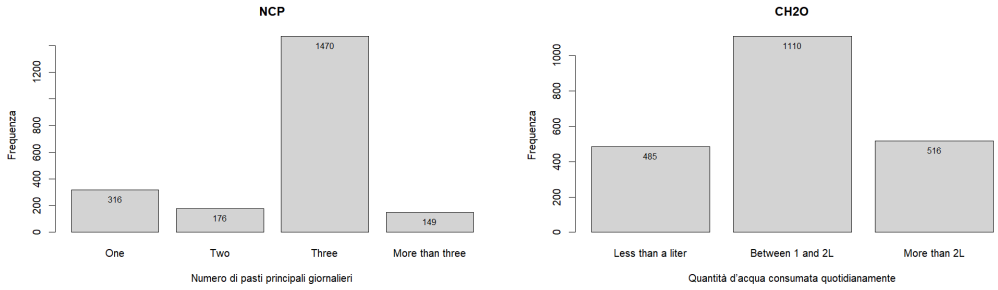

Figura 9: Barplot delle variabili `NCP` e `CH2O`.

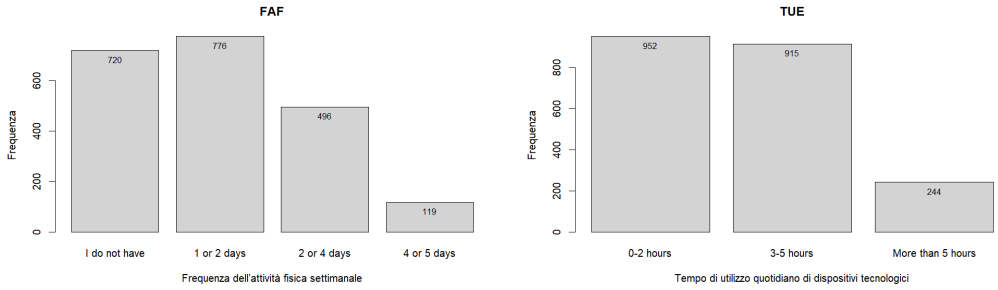

Figura 10: Barplot delle variabili `FAF` e `TUE`.

Un aspetto preliminare di rilievo riguarda la variabile `Gender`: nel dataset il numero di uomini e donne risulta pressoché bilanciato, garantendo una buona rappresentatività rispetto

al genere. Per quanto concerne le restanti caratteristiche categoriche e ordinali, emergono pattern comportamentali piuttosto marcati nel campione analizzato. In particolare:

- l'88% degli individui dichiara di consumare frequentemente cibi ad alto contenuto calorico;
- l'84% consuma regolarmente spuntini tra un pasto e l'altro;
- il 98% non fuma;
- il 95% non monitora l'assunzione calorica giornaliera;
- il 97% riferisce di non consumare alcol o di consumarlo solo saltuariamente;
- il 96% utilizza come principale mezzo di trasporto l'automobile o i trasporti pubblici;
- il 5% non consuma verdure durante i pasti;
- il 70% afferma di effettuare tre pasti principali al giorno;
- il 53% beve quotidianamente tra 1 e 2 litri d'acqua;
- il 34% non pratica attività fisica settimanale;
- il 12% utilizza dispositivi tecnologici per più di cinque ore al giorno.

Queste osservazioni, supportate dai barplot riportati nelle figure, offrono una panoramica immediata delle abitudini alimentari e comportamentali del campione e costituiscono la base per possibili interpretazioni successive riguardo ai legami con il livello di obesità.

4.3 Quartili e boxplot

I quartili rappresentano valori di soglia che suddividono una distribuzione ordinata in quattro parti di uguale numerosità. In particolare, il primo quartile (Q1) individua il punto sotto il quale si colloca il 25% delle osservazioni, la mediana coincide con il secondo quartile (Q2) e delimita il 50% dei valori inferiori, mentre il terzo quartile (Q3) separa il 75% dei dati dal restante 25%. La differenza tra Q3 e Q1 definisce lo *scarto interquartile* (Interquartile Range, IQR), misura particolarmente robusta della dispersione, poiché poco influenzata dai valori anomali.
Il boxplot, o diagramma a scatola, costituisce una rappresentazione sintetica ed efficace di questi elementi: la scatola racchiude l'intervallo interquartile compreso tra Q1 e Q3, la linea interna rappresenta la mediana, mentre le estremità dei cosiddetti *baffi* si estendono fino ai valori più estremi non considerati outlier (tipicamente definiti come osservazioni entro 1.5 volte l'IQR dai quartili). I punti al di fuori di tale intervallo vengono riportati graficamente come outlier e segnalano la presenza di valori particolarmente distanti dalla struttura centrale dei dati.

Di seguito analizziamo i boxplot delle variabili `Age`, `Height` e `Weight`, al fine di evidenziare eventuali outlier e possibili forme di asimmetria.

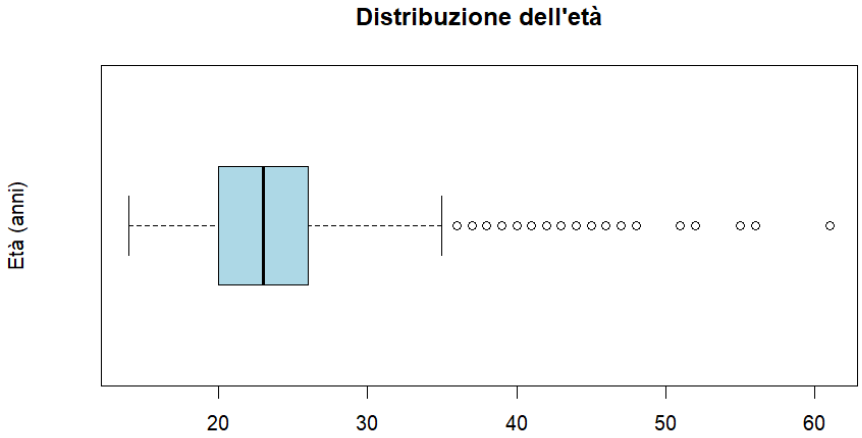

Figura 11: Boxplot della variabile `Age`

	Age
Estremo baffo inferiore	14
Q1	20
Mediana	23
Q3	26
Estremo baffo superiore	35

Tabella 3: Statistiche di posizione per l'età

Il boxplot della variabile `Age` evidenzia una marcata asimmetria positiva, già osservata tramite gli istogrammi: la presenza di numerosi outlier sulla coda destra segnala una concentrazione di valori estremamente elevati rispetto al resto della distribuzione.

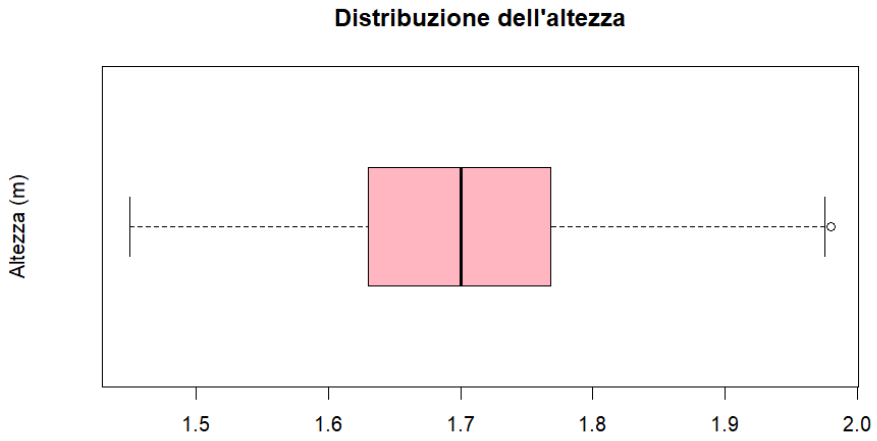

Figura 12: Boxplot della variabile `Height`

	Height
Estremo baffo inferiore	1.45
Q1	1.63
Mediana	1.70
Q3	1.77
Estremo baffo superiore	1.98

Tabella 4: Statistiche di posizione per l'altezza

La variabile `Height` mostra una distribuzione sostanzialmente simmetrica: i baffi presentano lunghezze molto simili e si osserva un solo outlier, indicativo di un valore isolato ma non influente sulla struttura complessiva dei dati.

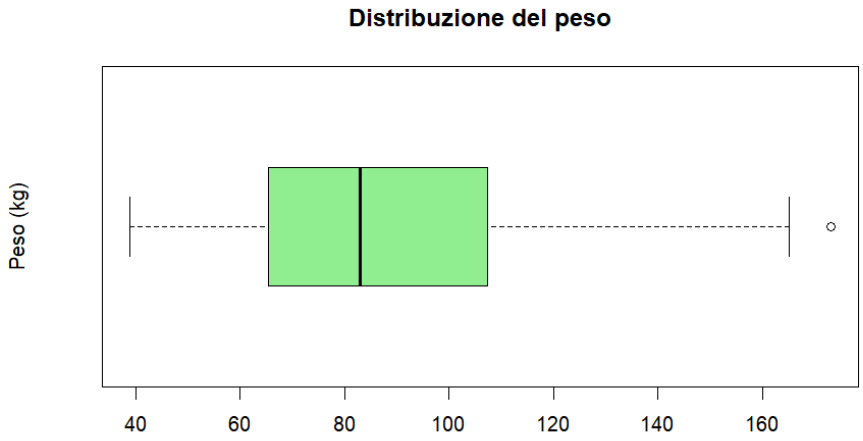

Figura 13: Boxplot della variabile `Weight`

	Weight
Estremo baffo inferiore	39
Q1	65.47
Mediana	83
Q3	107.43
Estremo baffo superiore	165.06

Tabella 5: Statistiche di posizione per il peso

Il boxplot della variabile `Weight` mette in evidenza una leggera asimmetria positiva: il baffo superiore risulta più lungo di quello inferiore e l'IQR è più esteso nella parte superiore della distribuzione (Q3–Mediana > Mediana–Q1). È inoltre presente un singolo outlier, che non altera sostanzialmente l'interpretazione generale della variabile.

5 Statistica descrittiva bivariata

Riferimenti bibliografici

[1] Fabio Mendoza Palechor and Alexis De la Hoz Manotas. Estimation of obesity levels based on eating habits and physical condition. `https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition`, 2019. UCI Machine Learning Repository.