# PROGETTO BIG DATA AND MACHINE LEARNING A.A. 2025-26

## INTRODUZIONE

Il nostro progetto si propone di analizzare un dataset dedicato alla rilevazione dello stress. In particolare, ha l'obiettivo di studiare: caratteristiche individuali, condizioni lavorative, aspetti relazionali, abitudini quotidiane e indicatori fisiologici degli individui di riferimento. Inoltre analizzeremo se la combinazione di questi fattori possa determinare il livello di stress degli individui, caratterizzando dunque il benessere psicofisico.
Il tema dello stress è stato più volte discusso negli ultimi anni, evidenziando ricadute negative sulla salute, sulla produttività e sulla qualità della vita. Risulta, pertanto, di grande interesse statistico analizzare e comprendere tale fenomeno, specialmente in un contesto moderno caratterizzato da ritmi accelerati, carichi lavorativi intensi e crescente sedentarietà.
L'analisi si basa su un dataset contenente 773 osservazioni rappresentanti ognuna un individuo e un totale di 22 variabili. La variabile principale del dataset è quella relativa alla rilevazione dello stress (Stress_Detection), ossia una variabile qualitativa ordinale, che classifica lo stress in tre categorie: low, medium e high.
Al fine di focalizzare l'indagine sui fattori più rilevanti per la determinazione del livello di stress, abbiamo deciso di analizzare 14 variabili, suddivise in 5 variabili qualitative e 9 variabili quantitative.

Per quanto riguarda le variabili qualitative abbiamo deciso di analizzare:

- **Genere (Gender)**: variabile qualitativa dicotomica nominale con modalità *Female* e *Male*, utile per analizzare, tramite la sua correlazione con i livelli di stress degli individui, possibili differenze tra uomini e donne.
- **Occupazione (Occupation)**: variabile qualitativa nominale, contenente 169 lavori diversi, che permette di distinguere condizioni lavorative eterogenee associate a diversi carichi di lavoro e che di conseguenza potrebbe incidere sui differenti livelli di stress.
- **Stato civile (Marital_Status)**: variabile qualitativa nominale, con modalità *Divorced, Married* e *Single*, che fornisce informazioni sul contesto familiare e relazionale, il quale potrebbe influire sul supporto sociale percepito e quindi sul livello di stress.
- **Abitudine al fumo (Smoking_Habit)**: variabile qualitativa nominale dicotomica, con modalità *No* e *Yes*, che indica la presenza di un comportamento a rischio per la salute spesso legato alla gestione dello stress.
- **Livello di stress (Stress_Detection)**: variabile qualitativa ordinale con modalità *Low*, *Medium* e *High*, che rappresenta il livello di stress percepito dall'individuo. Essa è la variabile più importante nel nostro dataset e sarà fondamentale nella parte relativa ai cluster, nella quale verificheremo, tramite un'analisi di tipo confermativo-previsivo, se la classificazione attuale in tre livelli risulta adeguata oppure se questa suggerisce una diversa suddivisione.

In riferimento alle variabili quantitative, abbiamo selezionato 9 variabili oggetto di analisi:

- **Età (Age)**: variabile quantitativa discreta, rilevata in anni interi. Controlleremo se l'età dei soggetti differisce in modo significativo tra i tre livelli di stress (Low, Medium, High).
- **Durata del sonno (Sleep_Duration)**: variabile quantitativa continua, misura il numero medio di ore di sonno per notte. Una durata di ore di sonno insufficiente o molto variabile potrebbe essere associata ad un maggior livello di stress.
- **Qualità del sonno (Sleep_Quality)**: variabile quantitativa continua, rappresenta la valutazione soggettiva del livello di qualità del sonno, rilevata tramite una scala numerica da 2 a 5.
- **Attività fisica (Physical_Activity)**: variabile quantitativa continua, rappresenta la frequenza dell'attività fisica e permette di valutare quanto uno stile di vita più attivo possa essere associato ad un livello di stress inferiore o superiore. Rilevata tramite una scala numerica da 1 a 5.
- **Tempo davanti allo schermo (Screen_Time)**: variabile quantitativa continua, rappresenta il numero medio di ore trascorse al giorno davanti a schermi elettronici (PC, Smartphone, TV…). È un indicatore indiretto di sedentarietà potenzialmente collegato ad una peggiore qualità del sonno e di conseguenza ad un maggior livello di stress.
- **Assunzione di caffeina (Caffeine_Intake)**: variabile quantitativa discreta, rappresenta la quantità media di caffeina consumata al giorno ( tazze di caffè). Un'assunzione di caffeina elevata è spesso associata a maggiore nervosismo, agitazione e difficoltà ad addormentarsi, fattori che possono contribuire ad un maggior livello di stress.
- **Ore di lavoro (Work_Hours)**: variabile quantitativa discreta, indica il numero di ore lavorate al giorno. Un numero elevato di ore può tradursi in meno tempo per il riposo, le relazioni sociali e le attività di recupero, fattori che possono incrementare lo stress percepito.
- **Livello di pressione sanguigna (Blood_Pressure)**: variabile quantitativa discreta, indica il livello di pressione sanguigna dell'individuo (da 110 a 170).
  Valori elevati possono essere relazionati a condizioni di stress cronico, stili di vita poco salutari o altre condizioni mediche.
- **Livello di glicemia (Blood_Sugar_Level)**: variabile quantitativa discreta, indica il livello di glicemia dell'individuo (misurata in mg/dL). Alterazioni del controllo glicemico possono dipendere dallo stile di vita, dall'attività fisica e da condizioni di stress elevato.

Nella prima parte del progetto svolgeremo un'analisi descrittiva del dataset per analizzare e raffigurare graficamente la distribuzione e le caratteristiche delle variabili da noi selezionate, mettendo anche queste in relazione tra loro. Successivamente, applicheremo tecniche di clustering, utilizzando metodi non supervisionati, per confrontare i risultati ottenuti con la classificazione attuale in tre livelli di stress ed integrando anche un approccio su Spark per effettuare una classificazione predittiva, scomponendo il campione tramite training e test.

# ANALISI DEI DATI IN R

Innanzitutto importiamo il file “stress_detection_data” all’interno del codice tramite il comando *read.csv* e, per capire se ci sono eventuali valori mancanti, costruiamo un sottoinsieme del dataset contenente soltanto le colonne prive di NA. Notiamo che il file non presenta valori mancanti, possiamo quindi procedere senza ulteriori pulizie.

Successivamente creiamo un nuovo dataframe, conservando solo le variabili che abbiamo indicato precedentemente nell’introduzione. Questo al fine di semplificare i dati sui quali lavorare e concentrare la nostra attenzione su quelle variabili che possono influenzare maggiormente il livello di stress. Visualizziamo i nomi delle variabili e la struttura complessiva del dataset, la quale evidenzia il numero di osservazioni,il numero di variabili e il tipo di variabili.

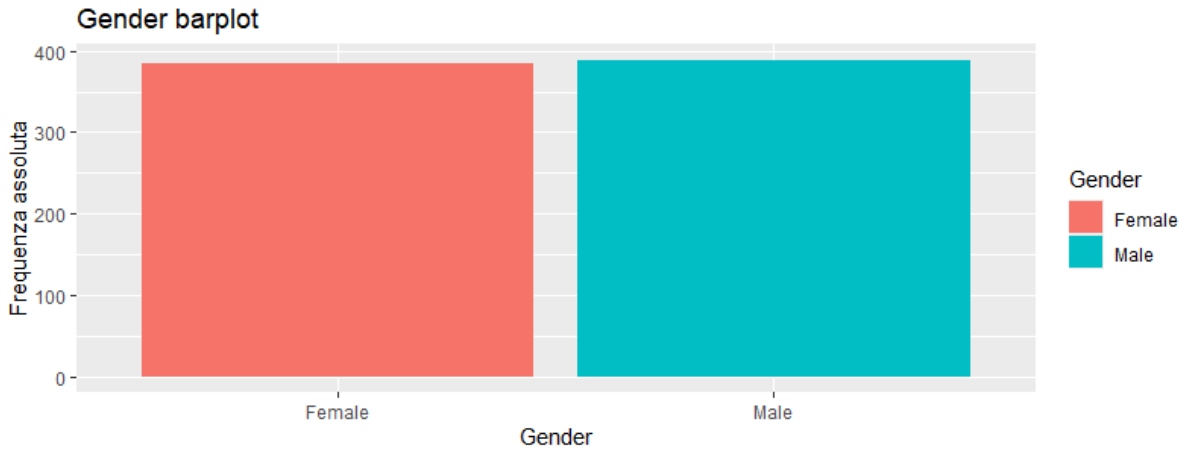
# ANALISI DESCRITTIVA

### VARIABILI QUALITATIVE

Iniziamo la nostra analisi descrittiva con lo studio di 5 variabili qualitative (Genere, Occupazione, Stato Civile, Abitudine al fumo, Livello di stress).

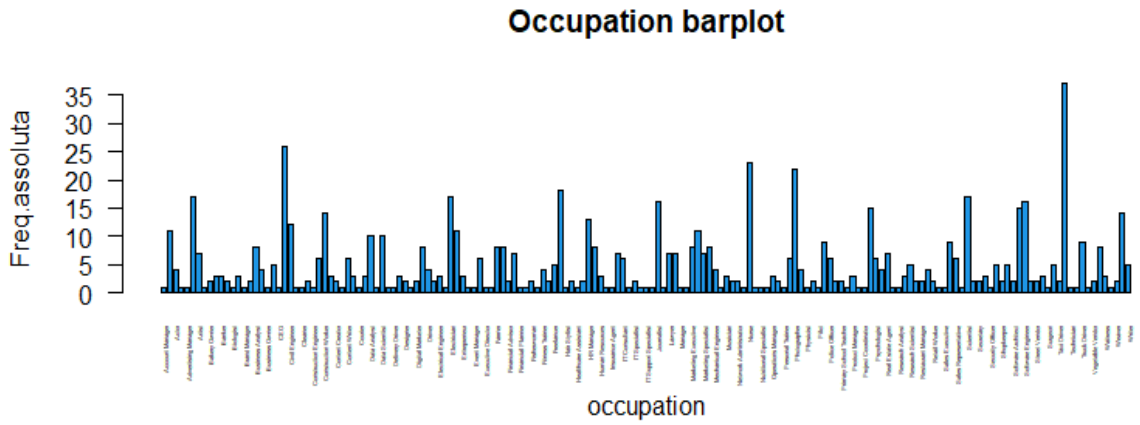
Per ciascuna variabile calcoliamo tre diverse tipologie di frequenze: assolute, relative e percentuali. Le frequenze assolute permettono di osservare quante unità statistiche ricadono in ciascuna modalità; le frequenze relative esprimono la quota di osservazioni associata a ogni modalità rispetto alla numerosità campionaria; le frequenze percentuali corrispondono alle frequenze relative espresse in percentuale, rendendo il confronto tra modalità più immediato. Inoltre per ogni variabile viene individuata la modalità più frequente, ovvero la moda. Infine realizziamo dei grafici a barre per mostrare visivamente la distribuzione, utilizzando dove possibile il pacchetto *ggplot2*.

Per la variabile **Gender** il grafico riporta sull’asse delle x le categorie e sull’asse delle y le frequenze assolute, mentre i colori ci consentono di distinguere al meglio le modalità.



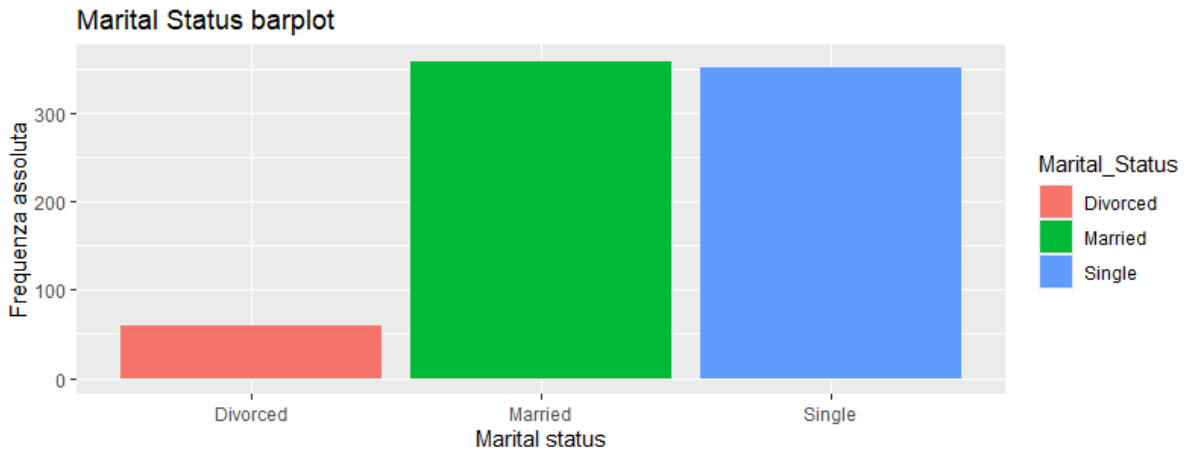

Vi sono 384 (49,68%) individui di genere Female e 389 (50,32%) individui di genere Male, quindi il campione è quasi perfettamente bilanciato tra i due generi. Concludiamo quindi che non esiste una prevalenza numerica significativa di un genere rispetto all'altro.

Usiamo lo stesso procedimento per la variabile **Occupation**:




Dall'analisi di questa variabile capiamo che la professione più frequente nel campione è l'insegnante con ben 37 individui. Il campione è estremamente frammentato, con 169 lavori diversi, la maggior parte dei quali presenta frequenze molto basse (1 o 2 individui).

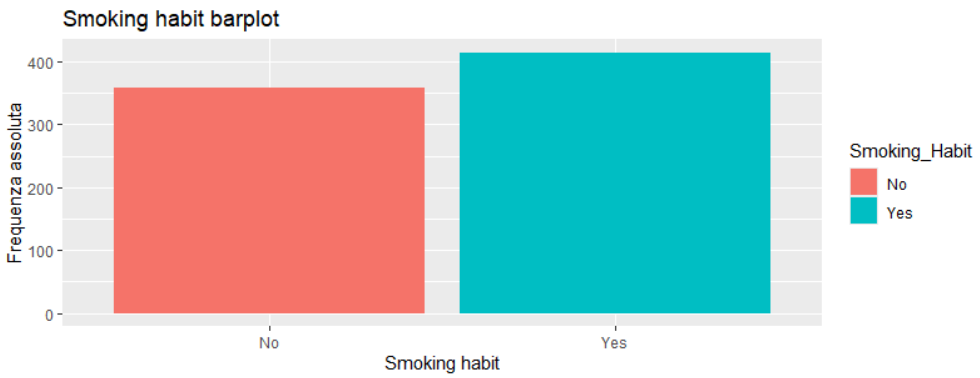
La variabile **Marital_Status** viene analizzata con le stesse modalità; l'obiettivo è comprendere la composizione del campione in termini di stato civile e verificare l'eventuale presenza di categorie predominanti.




| **Marital Status** | **Frequenza assoluta** | **Frequenza relativa** | **Frequenza Percentuale** |
|---|---|---|---|
| **Divorced** | 60 | 0.08 | 7.7% |
| **Married** | 360 | 0.47 | 46.5% |
| **Single** | 353 | 0.45 | 45.6% |

Notiamo che 360 individui (46,57 %) sono sposati, i single sono 353 (45,67%), mentre i divorziati sono 60 (7,76%).

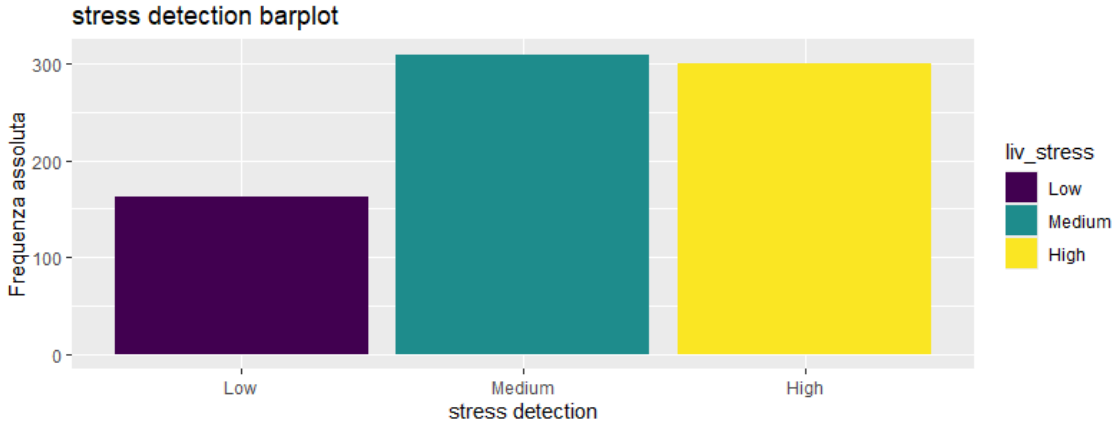
Anche per la variabile **Smoking_Habit**, dopo aver calcolato le frequenze, produciamo un grafico a barre per visualizzare la distribuzione degli individui in base alla loro abitudine al fumo.




Il fumatori rappresentano il 53,69% del campione totale, ovvero di 415 individui. I non fumatori sono 358, ossia il 46,31% del campione.

Infine dedichiamo una particolare attenzione alla variabile **Stress_Detection**, che rappresenta il livello di stress percepito. Come prima cosa, tramite la funzione *factor* definiamo le modalità della variabile e ne specifichiamo l'ordine crescente (livelli: Low, Medium, High). Trattandosi di una variabile qualitativa ordinale, calcoliamo il quartili subito dopo aver codificato i livelli rispettivamente come 1,2,3:  non più del 25% del campione presenta un livello di stress Low, mentre il  50% ha un livello pari o inferiore a Medium . Individuiamo la moda, che risulta essere il livello Medium e costruiamo la tabella delle frequenze:

| **Livello di stress** | **Frequenza assoluta** | **Frequenza relativa** | **Frequenza Percentuale** |
|---|---|---|---|
| **Low** | 162 | 0.21 | 20.96% |
| **Medium** | 310 | 0.40 | 40.10% |
| **High** | 301 | 0.39 | 38.94% |




Concludiamo che 162 individui (20,96%) hanno un livello basso di stress, 310 (40,10%) sono mediamente stressati e infine 301 (38,94%) sono molto stressati. Notiamo quindi che circa l'80% del campione percepisce uno stress medio-alto.

**VARIABILI QUANTITATIVE**

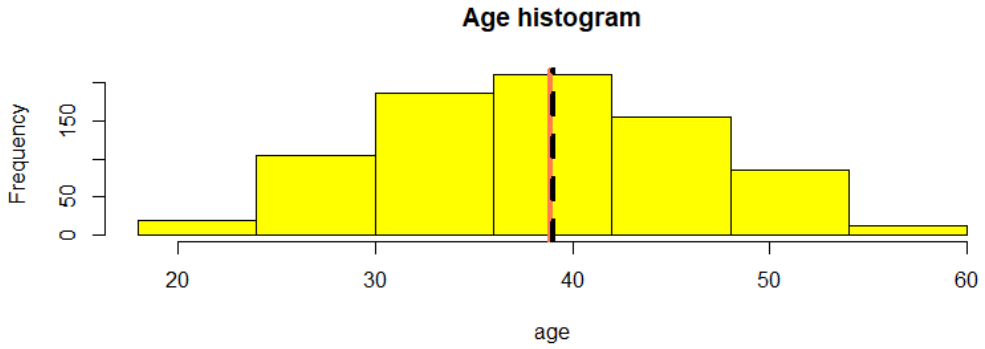
Passando alle variabili quantitative, consideriamo innanzitutto la variabile **Age**, che rappresenta l'età degli individui.

Partiamo da un'analisi descrittiva: tramite la funzione *summary* (che ci permette di ottenere valore minimo, primo quartile, mediana, media, terzo quartile e valore massimo delle nostre osservazioni) otteniamo i seguenti risultati:

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 18.00 | 33.00 | 39.00 | 38.89 | 45.00 | 60.00 |

Osserviamo che l'età degli individui è compresa tra un minimo di 18 e un massimo di 60 anni; si tratta quindi esclusivamente di popolazione in età adulta/lavorativa, senza la presenza né di minori né di individui in età avanzata.

Tramite il primo quartile notiamo che il 25% degli individui possiede un'età pari o inferiore a 33 anni, definendo il limite superiore della fascia dei giovani all'interno del dataset. Il terzo quartile indica che il 75% degli individui non supera i 45 anni, evidenziando che solo il 25% del campione si colloca al di sopra di tale età in una fascia più anziana.



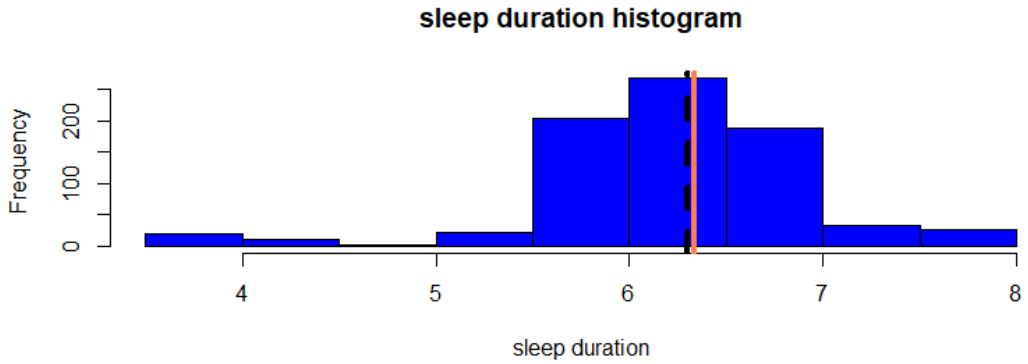

La mediana indica che il 50% degli individui ha un'età non superiore a 39 anni.
La media è pari a 38,89 anni (con deviazione standard pari a 7.68 anni), un valore molto vicino alla mediana: questa prossimità suggerisce una distribuzione dell'età centrata intorno ai 39 anni, senza asimmetrie significative verso fasce più giovani o più anziane, come confermato dall'istogramma.

Analizzando la variabile continua **Sleep_Duration**, che rappresenta il numero medio di ore di sonno per notte, otteniamo i seguenti risultati:

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 3.50 | 6.00 | 6.30 | 6.34 | 7.00 | 8.00 |

In questo caso osserviamo che il numero medio di ore di sonno degli individui è compreso tra un minimo di 3,5 e un massimo di 8 ore. Il primo quartile indica che il 25% degli individui dorme al massimo 6 ore, mentre il terzo quartile indica che il 75% degli individui non supera le 7 ore di sonno. Di conseguenza, notiamo che la maggior parte degli individui dorme tra 6 e 7 ore per notte.

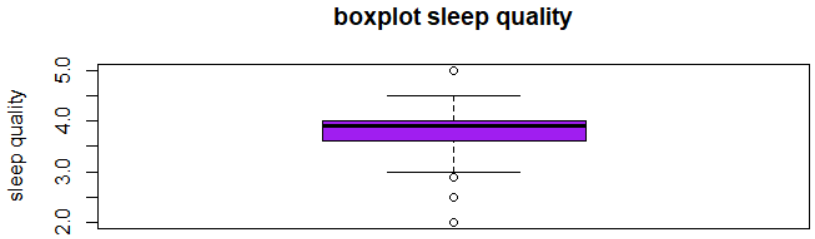
La media precisamente è pari a 6,33 ore (con deviazione standard pari a 0,73 ore), un valore molto vicino alla mediana, la quale indica che il 50% degli individui non dorme più di 6,30 ore per notte.



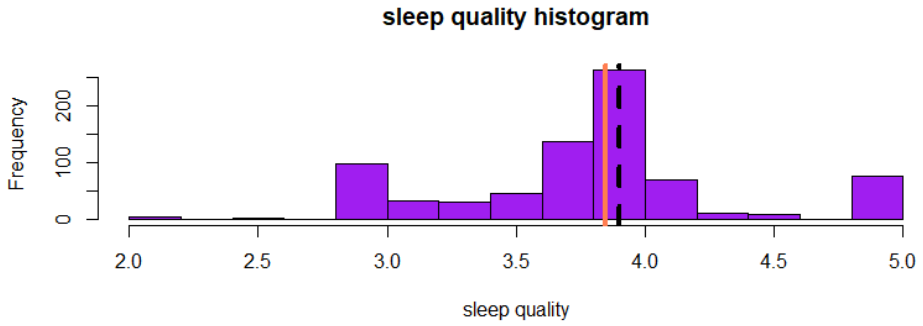

Dall'istogramma la distribuzione risulta essere approssimativamente asimmetrica negativa con coda a sinistra.
Infine il comando *t.test()* consente di ricavare l'intervallo di confidenza al 95% della media. Il confronto con μ=0 non ha reale significato in quanto non è plausibile una durata media del sonno pari a 0 ore (p-value < 2.2e-16). Tuttavia l'aspetto rilevante del t-test è l'intervallo di confidenza che il comando riporta e che risulta tra 6.29 e 6.39 ore, indicando che,con un grado di fiducia pari al 95%, la popolazione da cui è stato estratto il nostro dataset ha una durata del sonno media tra le 6,3 e 6,4 ore circa.

Riguardo la variabile continua **Sleep_Quality**, che rappresenta la valutazione soggettiva della qualità del sonno su scala numerica (da 2 a 5) ricaviamo:

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 2.00 | 3.60 | 3.90 | 3.84 | 4.00 | 5.00 |



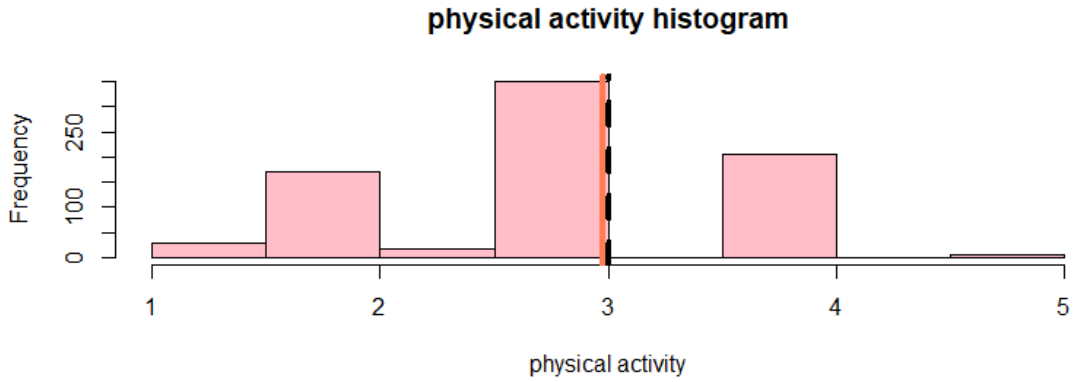

Dall'analisi emerge che la qualità del sonno degli individui è compresa tra un minimo di 2 e un massimo di 5.
Il primo quartile indica che il 25% degli individui ha una qualità del sonno non superiore a 3,6, mentre il terzo quartile indica che il 75% degli individui non supera il valore 4. La media è pari a 3,84 (con deviazione standard pari a 0,54), un valore molto vicino alla mediana, la quale indica che il 50% degli individui ha una qualità del sonno non superiore a 3.9.




La distribuzione è asimmetrica negativa con coda a sinistra: infatti la maggior parte degli individui valuta la propria qualità del sonno tra 3,6 e 4, mentre sono presenti poche osservazioni con valori inferiori che causano l'allungamento della coda verso sinistra.
Infine il comando *t.test()* consente di ricavare l'intervallo di confidenza al 95% della media generale, che è compresa tra 3.80 e 3.88.

Proseguendo con la variabile continua **Physical_Activity**, che rappresenta la frequenza dell'attività fisica, la funzione *summary* riporta:

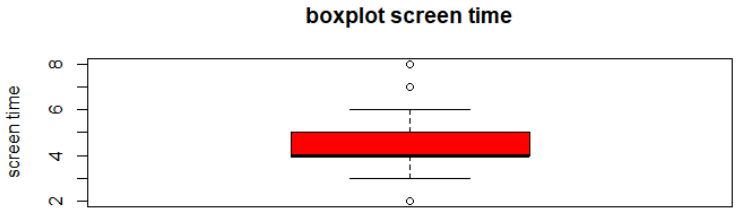
| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 1.00 | 2.00 | 3.00 | 2.97 | 4.00 | 5.00 |



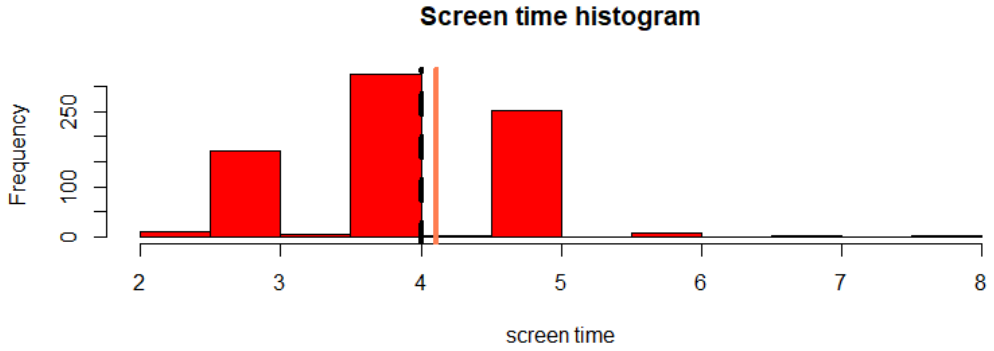

Osserviamo che l'attività fisica degli individui è compresa tra un minimo di 1 e un massimo di 5. Il primo quartile indica che il 25% degli individui ha una frequenza di attività fisica non superiore a 2, mentre il terzo quartile indica che il 75% degli individui non supera il valore 4.
La media è pari a 2,97 (con deviazione standard pari a 0,79), un valore molto vicino alla mediana, la quale indica che il 50% degli individui ha una frequenza di attività fisica non superiore a 3.
La distribuzione è approssimativamente simmetrica: la maggior parte delle osservazioni sono concentrate sul valore centrale 3, mentre i valori estremi sono meno frequenti. Ciò indica che, la maggior parte degli individui svolge attività fisica con una frequenza media, mentre sono relativamente pochi sia gli individui molto sedentari sia quelli molto attivi. Infine possiamo affermare, con un grado di fiducia del 95%, che la media generale è compresa tra 2,92 e 3,03.

Per ciò che concerne la variabile continua **Screen_Time**:

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 2.00 | 4.00 | 4.00 | 4.10 | 5.00 | 8.00 |




In questo caso osserviamo, anche tramite il grafico del boxplot, che il numero medio di ore trascorse al giorno davanti a schermi è compreso tra un minimo di 2 e un massimo di 8 ore. Il primo quartile indica che il 25% degli individui trascorre al massimo 4 ore davanti allo schermo, mentre il terzo quartile indica che il 75% degli individui non supera le 5 ore. La media è pari a 4,10 ore (con deviazione standard pari a 0,81) ore mentre la mediana mostra che il 50% degli individui non trascorre più di 4 ore al giorno davanti allo schermo.

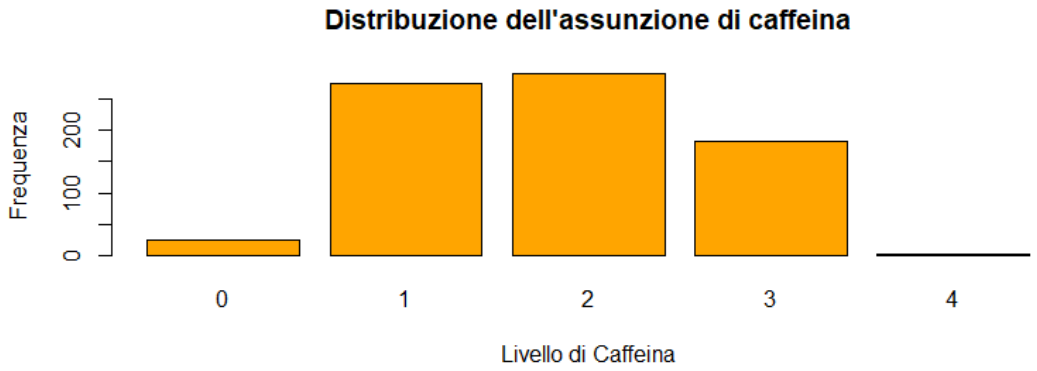



La distribuzione è asimmetrica positiva con coda a destra: infatti la maggior parte degli individui trascorre circa tra 4 e 5 ore al giorno davanti allo schermo, mentre solo una quota ridotta di individui trascorre più ore davanti allo schermo, determinando l'allungamento della distribuzione verso valori più alti.
L'intervallo di confidenza al 95% della media risulta essere tra 4.04 e 4.16 ore.

I risultati che otteniamo dallo studio della variabile discreta **Caffeine_Intake** sono:

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.00 | 1.00 | 2.00 | 1.81 | 2.00 | 4.00 |

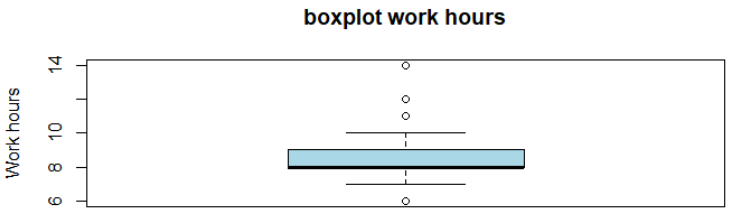
Osserviamo che le tazze di caffè consumate al giorno dagli individui sono comprese tra un minimo di 0 e un massimo di 4 tazze. Il primo quartile indica che il 25% degli individui assume una quantità di caffeina non superiore a 1 tazza, mentre il terzo quartile indica che il 75% degli individui non supera il valore 2.
La media è pari a 1.81 (con deviazione standard pari a 0.83) e la mediana ci permette di dedurre che il 50% degli individui assume una quantità di caffeina non superiore a 2 tazze.



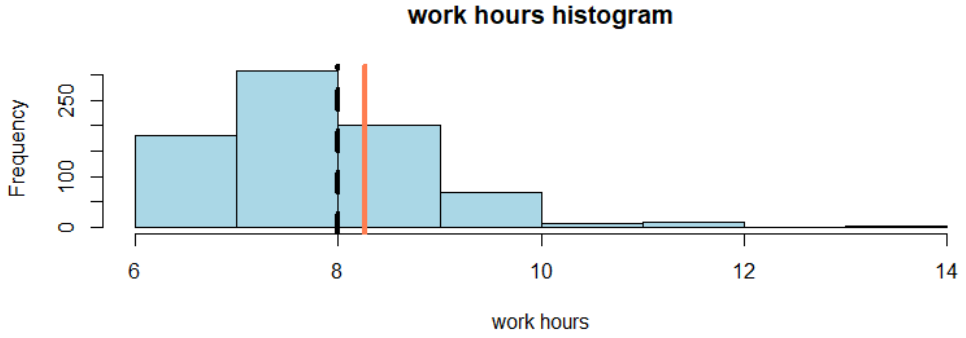

La variabile è stata rappresentata tramite un barplot in quanto presenta unicamente 5 modalità distinte (da 0 a 4); dall'analisi delle frequenze assolute si osserva che la maggior parte delle osservazioni si concentra nelle modalità 1,2,3 con rispettivamente 274, 289 e 182 individui, mentre pochissimi individui non assumono caffeina o ne abusano(modalità 4).

Dalla variabile discreta **Work_Hours** notiamo:

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 6.00 | 8.00 | 8.00 | 8.25 | 9.00 | 14.00 |



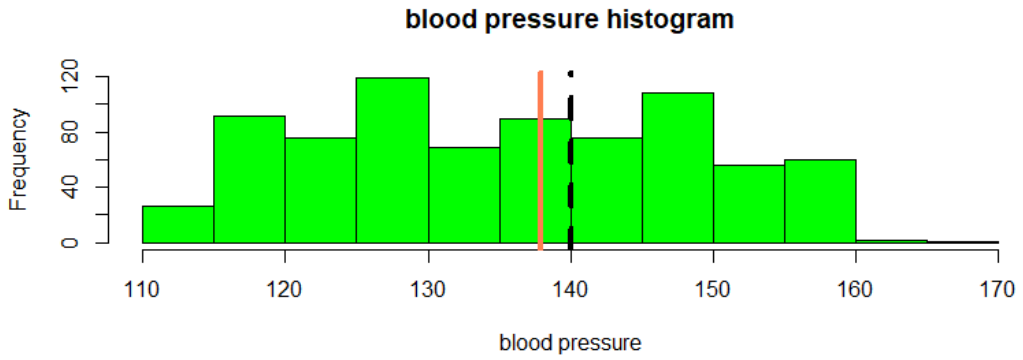

Deduciamo che il numero di ore lavorate al giorno è compreso tra un minimo di 6 e un massimo di 14 ore.
Il primo quartile indica che il 25% degli individui lavora al massimo 8 ore al giorno, mentre il terzo quartile indica che il 75% degli individui non supera le 9 ore.
La media è pari a 8.25 ore (con deviazione standard pari a 1.06 ore). La mediana ci mostra che il 50% degli individui non lavora più di 8 ore.




La distribuzione è asimmetrica positiva con coda a destra, infatti solo una quota ridotta di individui lavora più di 10 ore.
La variabile, per essere rappresentata tramite un istogramma, è stata raggruppata in classi di ampiezza 1 ora e dall'analisi delle frequenze assolute emerge una marcata concentrazione nelle prime classi, mentre nelle ore più elevate la numerosità diminuisce progressivamente.

L’analisi della variabile discreta **Blood_Pressure** riporta:

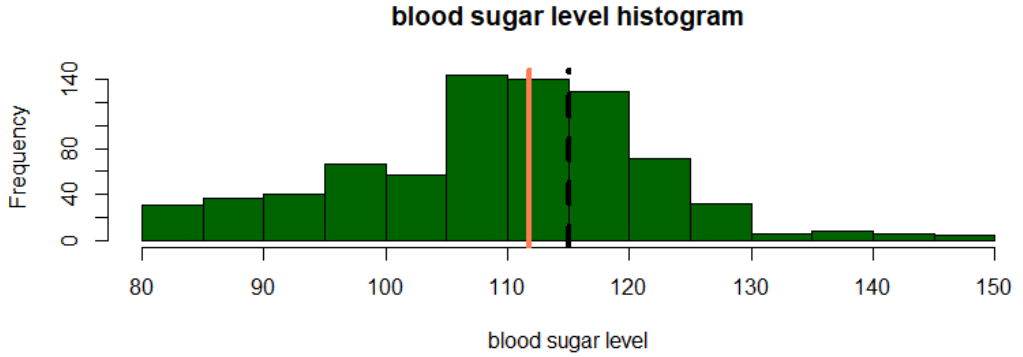
| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 110.0 | 130.0 | 140.0 | 137.9 | 150.0 | 170.0 |




Notiamo che il livello di pressione sanguigna degli individui è compreso tra un minimo di 110 e un massimo di 170 mmHg. Il primo quartile indica che il 25% degli individui possiede un livello non superiore a 130, mentre il terzo quartile indica che il 75% degli individui non supera il valore 150.
La media è pari a 137.9 (con deviazione standard pari a 13.12), mentre la mediana è pari a 140. La distribuzione è approssimativamente simmetrica: infatti la maggior parte delle osservazioni sono concentrate tra i valori 130 e 140.

In conclusione analizziamo la variabile discreta **Blood_Sugar_Level**:

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 80.0 | 105.0 | 115.0 | 111.8 | 120.0 | 150.0 |




Osserviamo che il livello di glicemia degli individui è compreso tra un minimo di 80 mg/dL e un massimo di 150 mg/dL. Il primo quartile indica che il 25% degli individui possiede un livello non superiore a 105, mentre il terzo quartile indica che il 75% degli individui non supera il valore 120.
La media è pari a 111.8 (con deviazione standard pari a 12.5) e il valore della mediana equivale a 115. La distribuzione è asimmetrica positiva con coda a destra.

**RELAZIONE TRA 2 VARIABILI QUALITATIVE**

Analizziamo ora come il livello di stress sia influenzato dalle variabili qualitative Gender, Marital_Status e Smoking_Habit.

Per quanto riguarda la variabile **Gender** creiamo la tabella di contingenza( o distribuzione doppia di frequenze assolute), che riporta il numero di individui appartenenti ad ogni combinazione delle due variabili.

| Livello di stress | Gender**: Female** | Gender**: Male** |
|---|---|---|
| **Low** | 91 | 71 |
| **Medium** | 171 | 139 |
| **High** | 122 | 179 |

Trasformiamo i conteggi in frequenze percentuali, in modo tale da confrontare meglio i gruppi.

| Livello di stress | **Female** | **Male** |
|---|---|---|
| **Low** | 11,8% | 9,2% |
| **Medium** | 22,1% | 18% |
| **High** | 15,8% | 23,2% |

Successivamente calcoliamo il test di indipendenza Chi-quadro, per capire se l'associazione tra le due variabili, Livello di stress e Genere, sia dovuta al caso o se esiste una relazione reale. Notiamo che il p-value è 0.0002568, quindi un valore inferiore ad un livello di significatività pari a 0,05: di conseguenza il risultato del test è statisticamente significativo e ci porta a rifiutare l'ipotesi nulla di indipendenza. Esiste quindi una relazione significativa tra il livello di stress e il genere: gli uomini risultano essere maggiormente stressati rispetto alle donne, le quali tendono più frequentemente a presentare un livello di stress Medium rispetto agli uomini.

Calcoliamo le frequenze teoriche d'indipendenza, ovvero i valori che ci aspetteremmo di osservare se tra le due variabili non esistesse alcuna relazione.

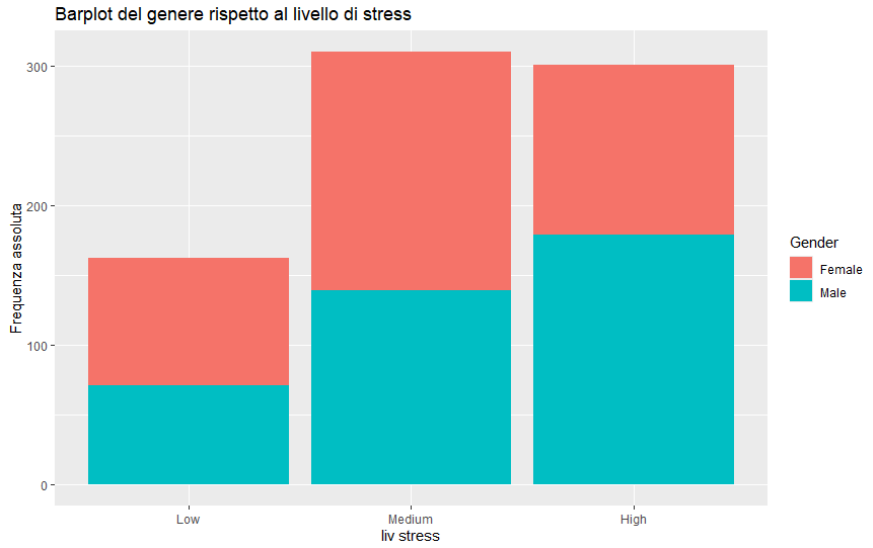
| Livello di stress | **Female** | **Male** |
|---|---|---|
| **Low** | 80.47 | 81.52 |
| **Medium** | 153.99 | 156.00 |
| **High** | 149.52 | 151.47 |

In seguito calcoliamo le contingenze quadratiche, ovvero i residui al quadrato, per identificare le celle che si discostano maggiormente dall'ipotesi di indipendenza. Più questi valori sono elevati, maggiore è il contributo di quella specifica cella al rifiuto dell'ipotesi nulla.

| Livello di stress | **Female** | **Male** |
|---|---|---|
| **Low** | 1.37 | 1.35 |
| **Medium** | 1.87 | 1.85 |
| **High** | 5.06 | 5.00 |

Dalle contingenze quadratiche emerge che le categorie Low e Medium mostrano residui molto bassi, ciò significa che non vi è nessuna deviazione importante rispetto alle frequenze teoriche. Mentre il livello High presenta dei residui al quadrato più elevati, per entrambi i generi, a prova del fatto che esiste una differenza tra uomini e donne nei livelli alti di stress.

In seguito generiamo un diagramma a barre sovrapposte che ci aiuta a visualizzare graficamente come lo stress cambia tra le varie categorie.

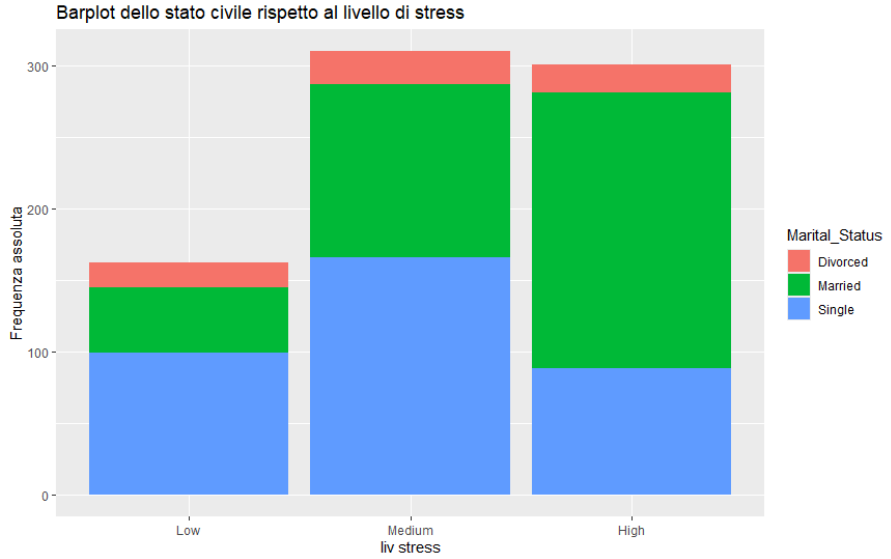



Usiamo la stessa identica procedura per le altre due variabili qualitative: Marital_Status e Smoking_Habit.

Osserviamo le frequenze percentuali della variabile **Marital_Status**:

| Livello di Stress | **Divorced** | **Married** | **Single** |
|---|---|---|---|
| **Low** | 2,2% | 5,9% | 12,8% |
| **Medium** | 3,0% | 15,6% | 21,5% |
| **High** | 2,6% | 25% | 11,4% |

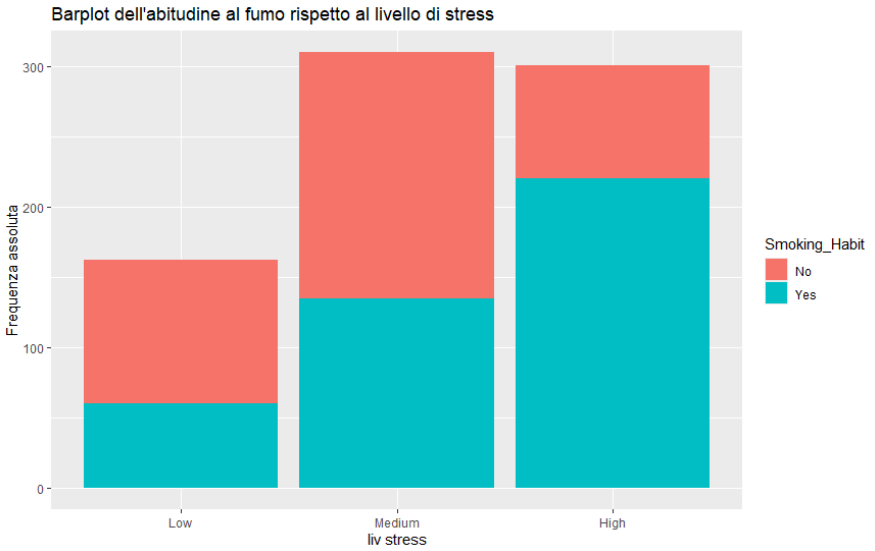
Notiamo che esiste una dipendenza statisticamente significativa con la variabile stress, in particolare gli individui sposati sono associati a un livello di stress più elevato rispetto ai Single. Con il calcolo delle frequenze teoriche di indipendenza rappresentiamo i valori che otterremmo se lo Stato Civile non fosse influenzato minimamente dallo stress. Per i Married con stress High, il valore atteso era 140, ma quello osservato è 193. Poiché il valore reale è molto più alto di quello teorico, c'è una forte dipendenza tra l'essere sposati e avere uno stress alto.




Passiamo ora alla variabile **Smoking_Habit**:

| Livello di stress | **No** | **Yes** |
|---|---|---|
| **Low** | 13,2% | 7,8% |
| **Medium** | 22,6% | 17,5% |
| **High** | 10,5% | 28,5% |

Esiste una relazione statisticamente significativa con la variabile Stress, come ci mostra il valore del p-value minore di 0.05. I fumatori tendono quindi a mostrare livelli di stress decisamente più elevati rispetto ai non fumatori.

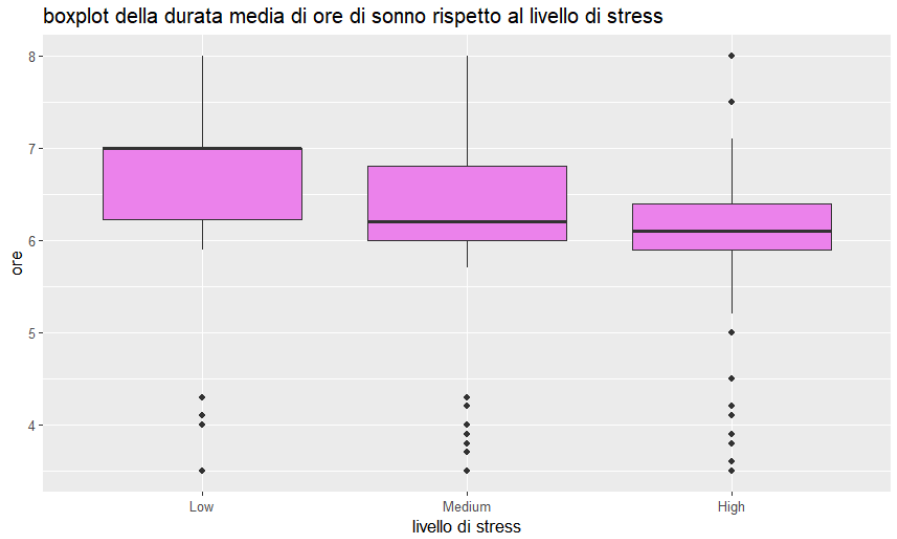


**RELAZIONE TRA UNA VARIABILE QUALITATIVA ED UNA QUANTITATIVA**

Successivamente, abbiamo messo in relazione una variabile qualitativa con una quantitativa; in primo luogo, osserviamo la variabile quantitativa relativa alla durata del sonno rispetto ai livelli di stress: possiamo notare come all'aumentare del livello di stress la durata media del sonno diminuisca.

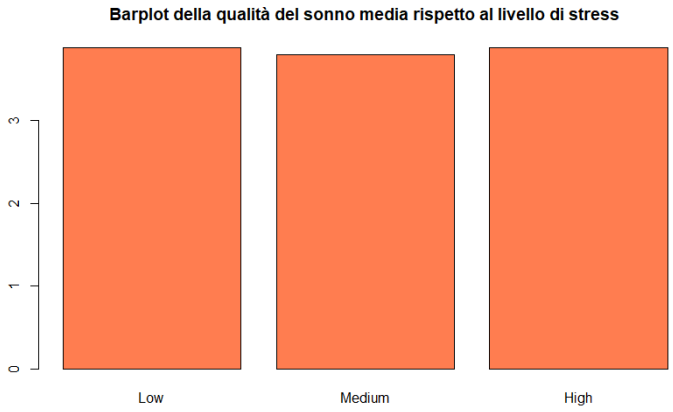
| Livello di stress | Durata media del sonno in ore |
|---|---|
| **Low** | 6.67 ore |
| **Medium** | 6.33 |
| **High** | 6.16 |

Confermiamo questa progressiva riduzione analizzando il boxplot, il quale ci mostra una mediana più alta per il gruppo di stress Low, che si distingue in modo netto rispetto agli altri gruppi, ed una concentrazione elevata di valori fortemente al di sotto del primo quartile per i gruppi Medium ed High, dimostrando che molti individui dei gruppi più stressati dormono pochissimo.

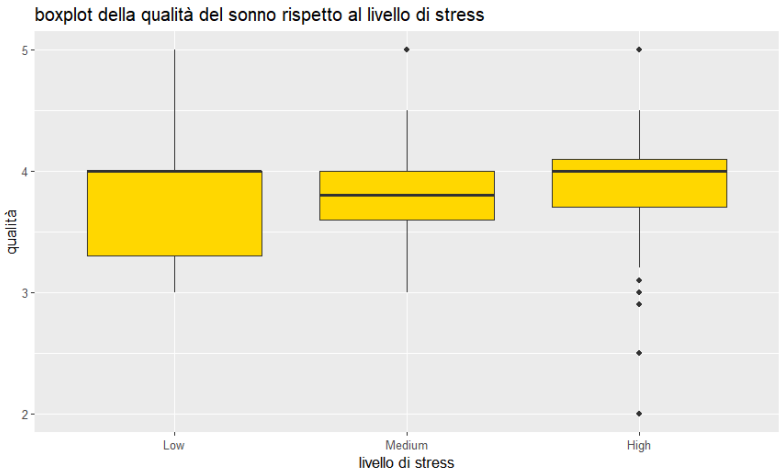



Tramite il test Anova, che riporta un p-value estremamente vicino a 0, possiamo rifiutare l'ipotesi nulla dell'uguaglianza delle medie dei gruppi, dimostrando l'esistenza di una differenza statisticamente significativa nella durata media del sonno per almeno due gruppi e di una relazione negativa tra le ore di sonno e il livello di stress.

Abbiamo analizzato rispetto al livello di stress anche le variabili relative alla qualità del sonno, alle ore di lavoro e all'attività fisica;




Per ciò che concerne la qualità del sonno le medie dei tre gruppi risultano essere molto simili tra loro con un rating medio che si attesta intorno a 3.8/5.



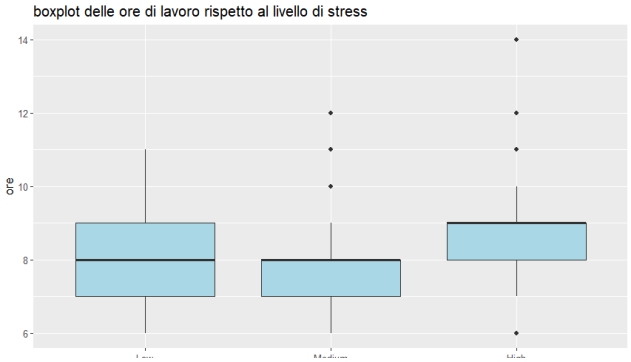

Dal boxplot possiamo notare che non ci sono particolari differenze tra gruppi, salvo la mediana del livello Medium che risulta leggermente inferiore rispetto agli altri 2, di conseguenza è probabile che non vi sia alcuna differenza statisticamente significativa tra i 3 gruppi. Ciò è dimostrato subito dopo nel test Anova che riporta un p-value maggiore di 0,05.

Rappresentando le variabili quantitative relative alle ore di lavoro e al livello di attività fisica rispetto ai livelli di stress, osserviamo che queste presentano caratteristiche molto simili;

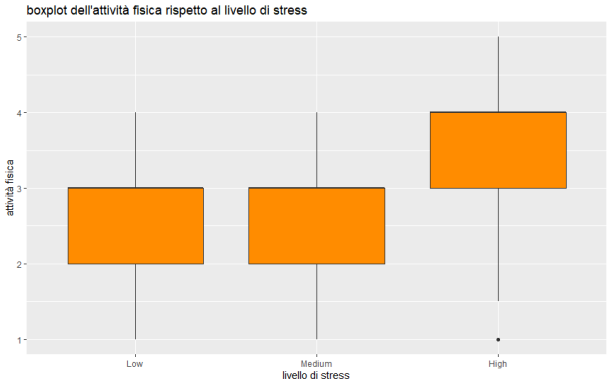
| Livelli di stress | **Ore di lavoro medie** rispetto allo stress |
|---|---|
| **Low** | 8.08 |
| **Medium** | 7.97 |
| **High** | **8.65** |

| Livelli di stress | Valore medio del **livello di attività fisica** rispetto allo stress |
|---|---|
| **Low** | 2.57 |
| **Medium** | 2.75 |
| **High** | **3.43** |

Per entrambe, infatti, i livelli di stress low e medium presentano medie praticamente identiche, mentre è pronunciata la differenza con il livello High, la cui media in entrambi i casi risulta più elevata.

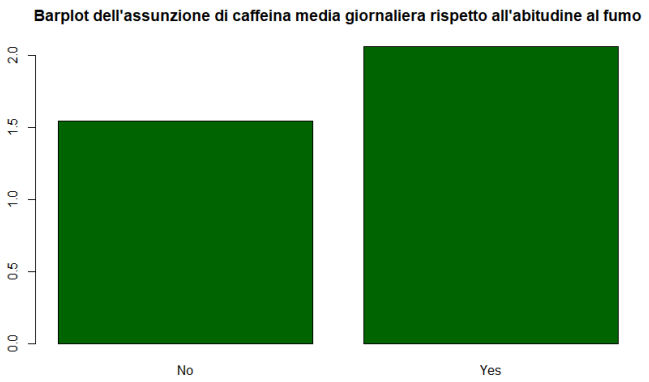



Dalla visione del boxplot delle ore di lavoro è interessante notare che nel livello High coesistono gli estremi inferiore e superiore del campione (sia 6 ore di lavoro che 14); questo ci fa capire che l'orario di lavoro non influenza in modo totale, o comunque con uno schema ben preciso, lo stress.




Particolare è il dato relativo all'attività fisica; si osserva che in linea di massima gli individui più stressati sono anche quelli che si allenano di più, il che risulta in controtendenza rispetto agli studi secondo cui un fattore legato allo stress possa essere la scarsa attività fisica.

Il test Anova riporta una differenza statisticamente significativa nelle ore di lavoro e nel livello di attività fisica tra i diversi gradi di stress, non specificando tuttavia quali e quanti gruppi si distinguono dagli altri. L'osservazione dei dati suggerisce che le differenze principali riguardano il gruppo ad alto livello di stress, dato che i livelli di basso e medio stress risultano molto simili, anche graficamente, tra loro.

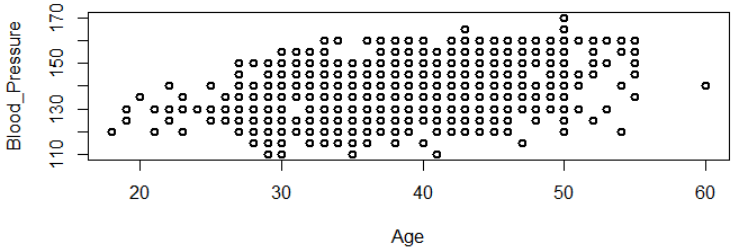



Allontanandoci per un attimo dall'analisi relativa ai diversi livelli di stress abbiamo voluto studiare se esiste una possibile correlazione positiva tra l'abitudine al fumo e il livello di caffeina assunta quotidianamente per gli individui del nostro campione. Osservando il barplot emerge chiaramente che la media del livello di caffeina assunto tra i fumatori è sensibilmente più elevata rispetto a quella dei non fumatori, suggerendo una stretta interdipendenza tra queste due abitudini. Questa differenza tra fumatori e non fumatori risulta anche statisticamente significativa, dato un p-value con un valore prossimo allo 0 che permette di rifiutare l'ipotesi nulla di uguaglianza delle medie.

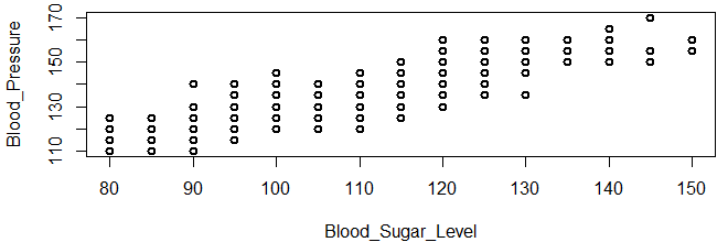
**RELAZIONE TRA 2 VARIABILI QUANTITATIVE**

Infine, mettendo in relazione coppie di variabili quantitative è possibile valutare la correlazione tra esse.

A questo fine vengono utilizzati il diagramma di dispersione e il coefficiente di correlazione di Pearson, che consentono un'analisi grafica e la quantificazione numerica della correlazione. Sono state selezionate specifiche coppie di variabili con lo scopo di mettere in relazione variabili anagrafiche (età), variabili fisiologiche (pressione sanguigna e livello di zucchero nel sangue) e variabili comportamentali (sonno, lavoro, consumo di caffeina e tempo di esposizione agli schermi).

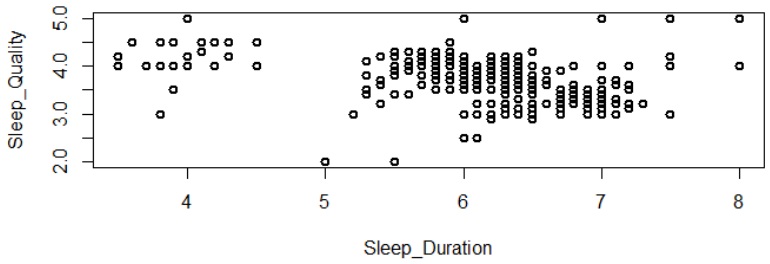



La prima relazione analizzata è tra l'età e la pressione sanguigna, scelta in quanto l'età è un fattore anagrafico tipicamente associato a modificazioni fisiologiche dell'apparato cardiovascolare.
Nel diagramma di dispersione osserviamo un andamento tendenzialmente crescente: all'aumentare dell'età, i valori di pressione sanguigna tendono in media ad aumentare. Il coefficiente di Pearson è pari a 0.39 e indica una correlazione positiva di entità moderata tra le due variabili.



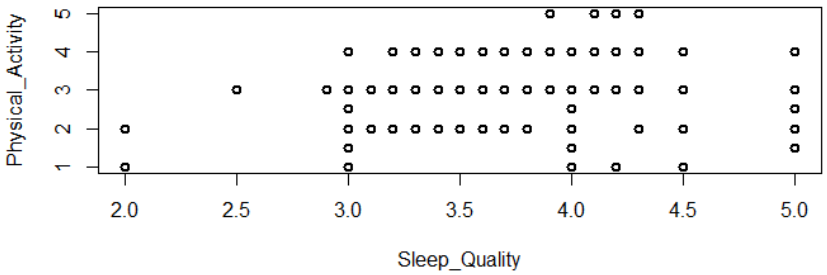

Successivamente abbiamo analizzato la relazione tra il livello di zucchero del sangue e la pressione sanguigna, in quanto sono due variabili fisiologiche strettamente connesse allo stato cardio-metabolico dell'individuo.
Nel diagramma di dispersione notiamo un andamento crescente: all'aumentare del livello di zucchero nel sangue aumenta il livello di pressione sanguigna, i punti risultano piuttosto allineati tra loro evidenziando una relazione più marcata rispetto a quella precedente.
Tale andamento viene confermato da un coefficiente di Pearson pari a 0.82 che indica una correlazione positiva forte tra le due variabili.

E' stata poi considerata la relazione tra le variabili durata del sonno e qualità del sonno, in quanto la durata del sonno è tipicamente considerata un fattore che può influenzarne la qualità.

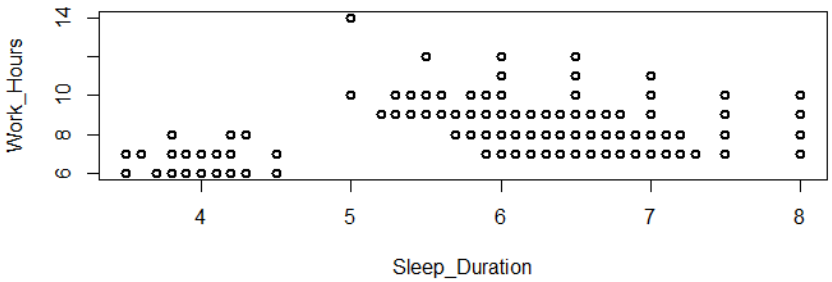



Osservando il diagramma notiamo un'elevata dispersione delle osservazioni, indicando una correlazione molto debole tra la durata e la qualità del sonno.
Questa correlazione molto debole viene confermata da un coefficiente di Pearson pari a 0.21.

Successivamente abbiamo analizzato la relazione tra qualità del sonno e attività fisica, scelta per verificare se diversi livelli di attività fisica si riflettono sulla qualità del riposo. Dal diagramma di dispersione non emerge un andamento definito in quanto le osservazioni risultano caratterizzate da un'elevata dispersione.

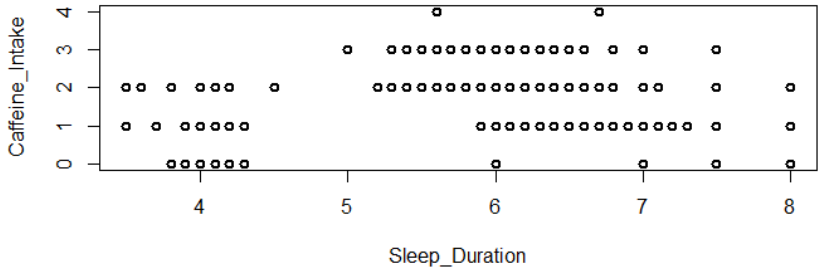



Il coefficiente di Pearson è pari a 0.09, valore molto vicino allo zero, e conferma la presenza di una relazione positiva ma estremamente debole.



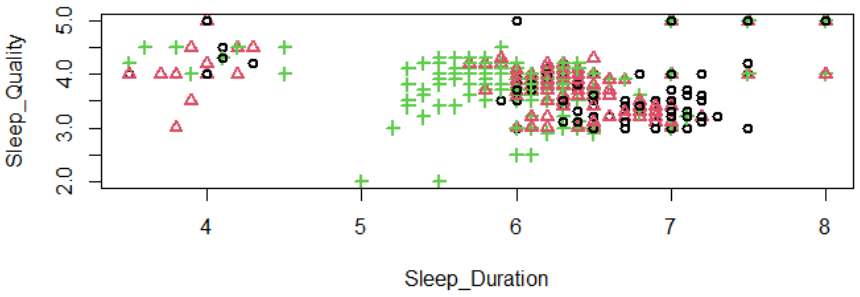

Allo stesso modo, abbiamo analizzato la relazione tra durata del sonno e ore di lavoro. Il diagramma di dispersione mostra una correlazione positiva molto debole, confermata da un coefficiente di Pearson pari a 0.24.

E' stata inoltre analizzata la relazione tra l'assunzione di caffeina e la durata del sonno, poiché la caffeina è una sostanza stimolante che può influenzare i ritmi del sonno.

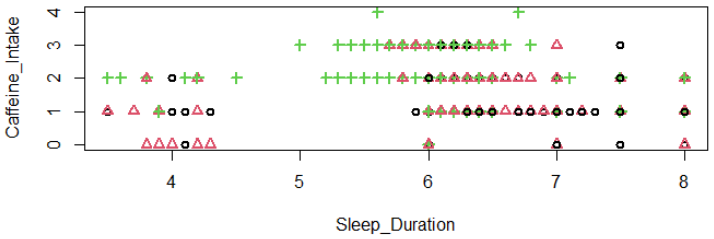



Dal diagramma di dispersione emerge un andamento tendenzialmente decrescente: all'aumentare del consumo di caffeina tendono a diminuire in media le ore di sonno. Il coefficiente di Pearson è pari a -0.24 e conferma una correlazione negativa debole tra le due variabili.

Successivamente abbiamo rappresentato le unità statistiche all'interno del diagramma di dispersione distinguendole in base ad una variabile qualitativa.
In particolare, abbiamo considerato le variabili qualitative genere e il livello di stress, in modo da verificare se esse incidono sulla distribuzione delle osservazioni nelle relazioni tra le variabili quantitative. Le variabili qualitative sono state inizialmente fattorizzate e poi convertite in forma numerica, in modo tale da poter distinguere graficamente le osservazioni tramite colore e forma.
I livelli del genere Female e Male sono stati codificati rispettivamente come 1 e 2, mentre i livelli di stress Low, Medium e High sono stati codificati come 1, 2 e 3. Ai codici numerici 1, 2 e 3 sono stati associati rispettivamente il colore nero, rosso e verde.

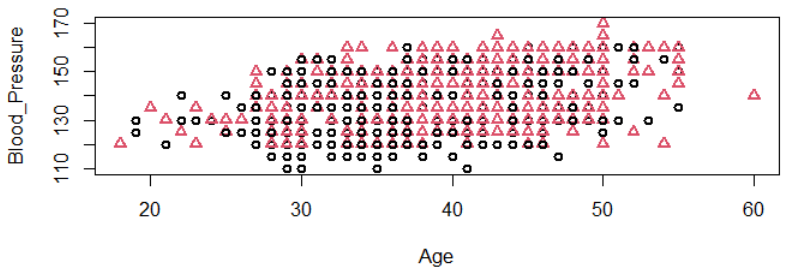



Nel diagramma di dispersione tra sleep duration e sleep quality, notiamo che le unità statistiche caratterizzate da un livello di stress basso (nero) tendono ad assumere valori più elevati sia di durata sia di qualità del sonno, quelle con livello di stress medio (rosso) tendono ad assumere valori intermedi, mentre quelle con livello di stress alto (verde) valori più bassi di entrambe le variabili. Inoltre ci sono anche alcune osservazioni che presentano una durata media del sonno molto bassa (inferiore a 5 ore a notte), ma questa loro condizione non dipende dal livello di stress di appartenenza.




Nel diagramma di dispersione tra durata del sonno e assunzione di caffeina emerge una relazione tendenzialmente decrescente tra le due variabili. Notiamo che gli individui con un livello di stress alto si collocano con valori elevati di assunzione caffeina e valori inferiori di durata del sonno (sono tendenzialmente concentrati in alto e nella parte centrale).

Infine, nel diagramma di dispersione tra l'età e la pressione sanguigna, con osservazioni classificate in base al genere, i due gruppi female e male risultano tendenzialmente sovrapposti, evidenziando che non sono presenti particolare distinzioni in base al genere, per cui il genere non è un fattore caratterizzante.



# FASE DI CLUSTERING CONFERMATIVO NON SUPERVISIONATO

### METODI GERARCHICI

Iniziamo la nostra analisi confermativa preparando i dati da standardizzare.
Innanzitutto vediamo come sono distribuiti gli individui in base al loro livello di stress:

LIVELLO DI STRESS

| **Low** | **Medium** | **High** |
|---|---|---|
| 162 | 310 | 301 |

Proseguiamo creando un nuovo dataset denominandolo Stress2, eliminando le variabili qualitative, in quanto i metodi di clustering che useremo si basano su distanze numeriche.
Successivamente tramite la funzione aggregate osserviamo la media di ogni variabile quantitativa per ciascun livello di stress:

| **Livello stress** | **Età media** | **Durata sonno** | **Qualità sonno** | **Attività fisica** | **Screen time** | **Caffeina** | **Ore lavoro** | **Pressione sanguigna** | **Livello di zucchero nel sangue** |
|---|---|---|---|---|---|---|---|---|---|
| **Low** | 36.12 | 6.67 | 3.88 | 2.57 | 3.58 | 1.30 | 8.08 | 128.02 | 103.40 |
| **Medium** | 37.81 | 6.33 | 3.80 | 2.75 | 3.90 | 1.63 | 7.97 | 135.94 | 108.79 |
| **High** | 41.49 | 6.16 | 3.88 | 3.44 | 4.60 | 2.29 | 8.65 | 145.35 | 119.34 |

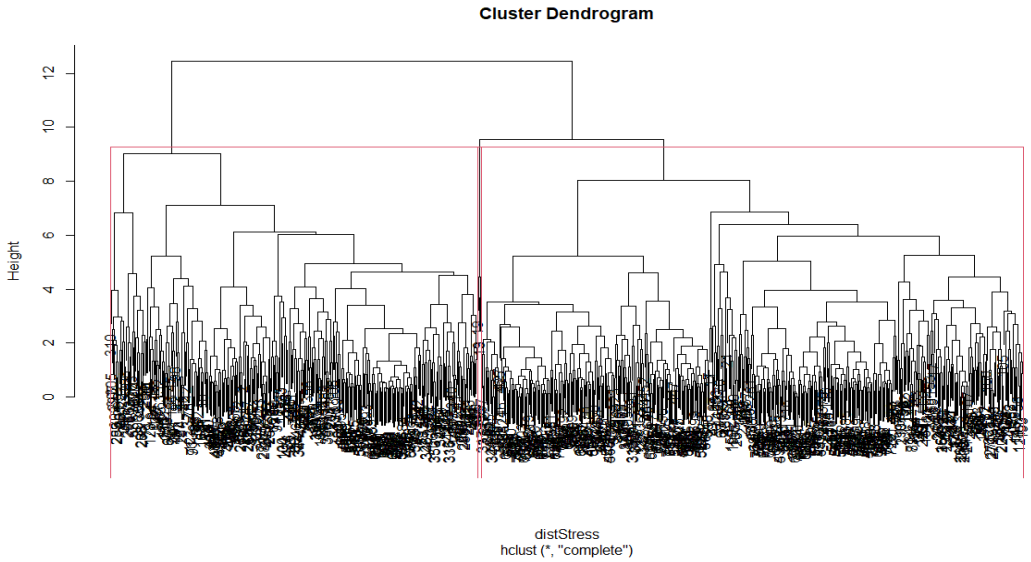
Standardizziamo il dataset affinché le variabili possano essere confrontabili tra loro eliminando l'effetto delle diverse unità di misura.
Calcoliamo la matrice delle distanze utilizzando la distanza euclidea, necessaria per utilizzare le tecniche di clustering gerarchico, che ci permetteranno di ricostruire in modo progressivo la formazione dei gruppi tramite dei dendrogrammi. Successivamente tagliamo i dendrogrammi in modo tale da avere un numero di gruppi pari a tre, come i nostri livelli di stress originali.

Ci serviamo di tutti i metodi affrontati a lezione: il metodo del legame singolo, medio, completo, centroide e ward.
Osserviamo che, forzando un'agglomerazione in 3 gruppi, in modo tale da ottenere un risultato il più simile possibile alla distribuzione degli individui in base al loro livello di stress, i metodi del legame singolo, medio e centroide non forniscono dei risultati soddisfacenti, in quanto concentrano praticamente tutte le osservazioni in un unico gruppo.

| Metodo | Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|---|
| **Legame singolo** | 771 | 1 | 1 |
| **Legame medio** | 753 | 3 | 17 |
| **Legame Centroide** | 769 | 1 | 3 |

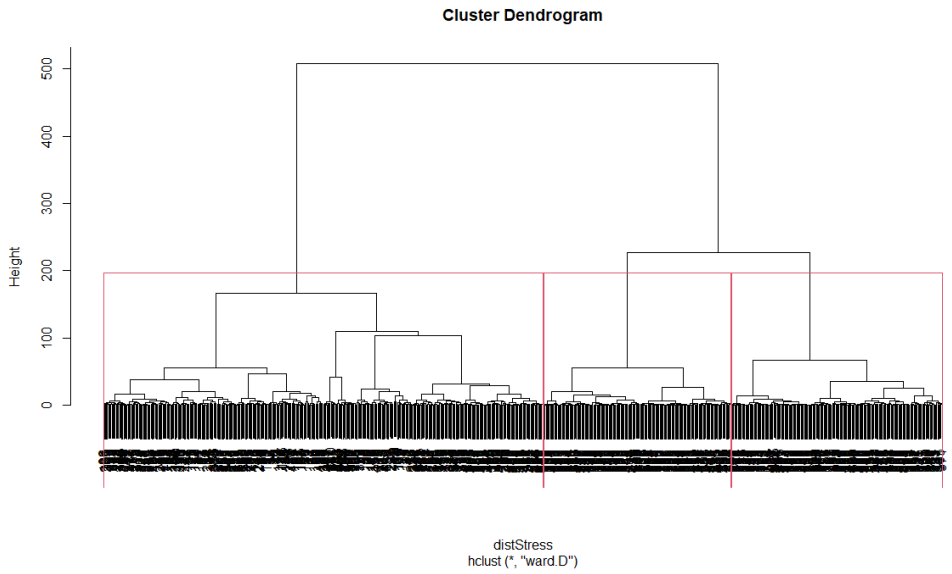
Invece con il metodo del legame completo otteniamo un miglioramento, in quanto risultano due gruppi nettamente separati a differenza dei risultati precedenti, avvicinandoci in parte alla distribuzione originale. Tuttavia non è ancora un risultato ottimale…

| | | | |
|---|---|---|---|
| **Legame completo** | 311 | 459 | 3 |



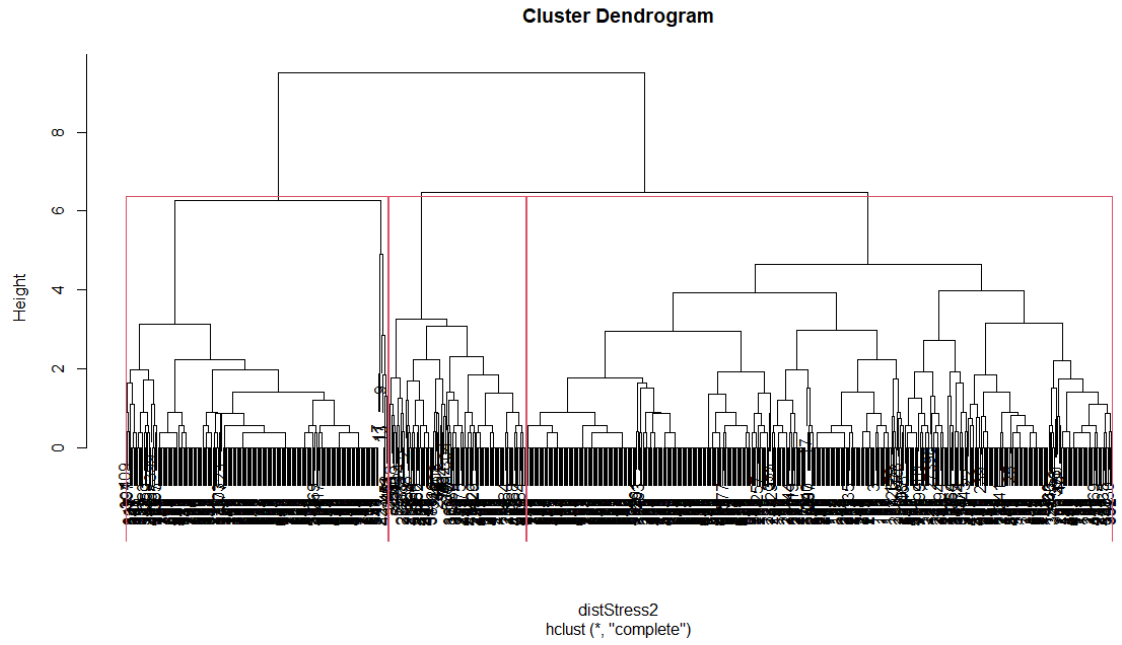

Con il metodo del legame Ward, invece, individuando 3 gruppi ben separati e di dimensioni più bilanciate, riusciamo ad ottenere il risultato migliore, coerente con la classificazione originale. Tenendo conto che i metodi gerarchici non sono supervisionati, quindi privi di variabile target, e si basano esclusivamente su distanze tra variabili quantitative, dunque come risultato può essere ritenuto complessivamente soddisfacente.

| | | | |
|---|---|---|---|
| **Legame Ward** | 405 | 195 | 173 |



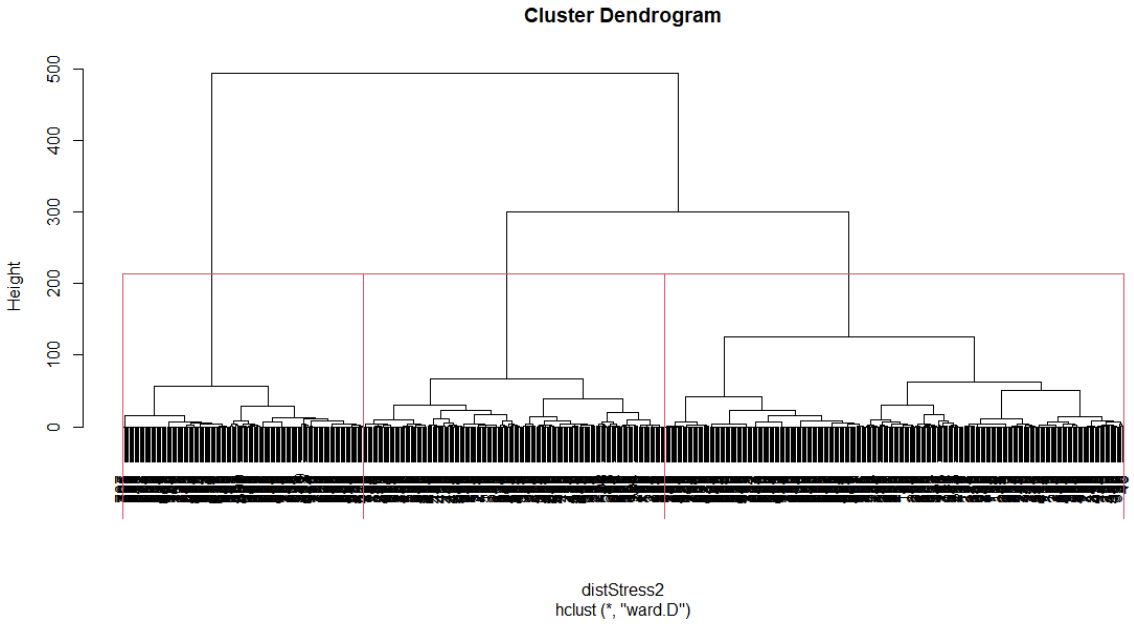

Per ottenere un risultato migliore, decidiamo di utilizzare solamente le variabili quantitative che hanno differenze significative in termini di media per i diversi livelli di stress. Per cui riduciamo il nostro dataset mantenendo come variabili: il livello di caffeina assunta, il livello di pressione sanguigna, i livelli di zucchero nel sangue e lo screen time. Questa volta utilizziamo solamente i metodi che ci avevano dato un risultato migliore con il dataset completo: il metodo del legame completo e del legame ward. Notiamo che con il legame completo vi è un netto miglioramento, poiché ora riusciamo ad identificare tre gruppi meglio distribuiti.

| | | | |
|---|---|---|---|
| **Legame completo** | 108 | 459 | 206 |



Il risultato ottimale anche in questo caso riusciamo ad ottenerlo con il metodo del legame ward, che ci permette di migliorare ulteriormente la chiarezza della struttura dei cluster rendendoli più omogenei al loro interno. Di conseguenza, il legame Ward rappresenta la soluzione migliore per la nostra analisi, perché fornisce una partizione più bilanciata e interpretabile rispetto agli altri metodi considerati.

| | | | |
|---|---|---|---|
| **Legame Ward** | 232 | 355 | 186 |



**SCELTA DEL NUMERO OTTIMALE DI CLUSTER TRAMITE NBCLUST**

**Dataset completo**

Precedentemente, utilizzando i metodi di clustering gerarchico abbiamo determinato, a posteriori, il numero di gruppi ottenibili effettuando un taglio del dendrogramma al livello in cui k=3, per rimanere coerenti con la classificazione originale dei livelli di stress e abbiamo ottenuto il numero di osservazioni che riempivano ogni gruppo.

Cambiamo ora prospettiva e concentriamoci sulla scelta del numero ottimale di cluster servendoci del pacchetto *NbClust,* al cui interno troviamo diversi indici. Ognuno di questi indici evidenzierà il miglior numero di gruppi per organizzare le nostre osservazioni. Per ogni indice fissiamo un intervallo che va da 2 a 5 cluster, applicando il metodo del legame completo, ward e l'algoritmo k-means.

Per il legame completo la maggioranza degli indici ci suggerisce come numero ottimale di cluster per classificare i nostri individui k=2.

| Numero di Cluster ottimale | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Numero di Indici | 2 | 1 | 9 | 3 | 6 | 5 |

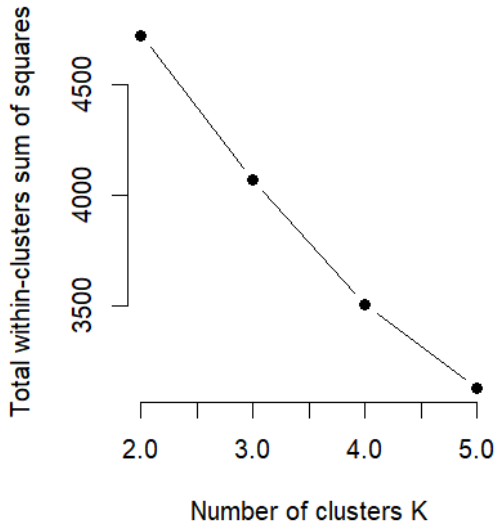
K=0 e K=1 come cluster ottimali sono forniti da indici non attendibili o che non si possono utilizzare nel nostro caso.

Anche utilizzando il metodo del legame Ward, nel caso del dataset completo, il numero ottimale di cluster risulta essere 2.

| Numero di Cluster ottimale | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Numero di Indici | 2 | 1 | 8 | 4 | 4 | 5 |

Estendiamo l'analisi servendoci dell'algoritmo k-means, inizialmente tramite il metodo del gomito. Per prima cosa l'algoritmo, per ogni possibile numero di k cluster compreso tra 2 e 5, calcola la WSS, cioè la somma delle distanze quadratiche tra le osservazioni e il centroide del cluster di appartenenza. Più questa somma sarà minore, più i nostri cluster saranno compatti e ben suddivisi. L'algoritmo k-means viene eseguito più volte (10) per rendere i nostri risultati più attendibili.

Con il metodo del gomito troviamo il numero di cluster ottimale, oltre il quale un gruppo aggiuntivo determina una riduzione del valore della WSS solo marginale e quindi trascurabile.




Dal grafico troviamo il gomito in corrispondenza di k=4, ovvero il numero di gruppi ottimale che l'algoritmo ci suggerisce.

Successivamente applichiamo il pacchetto *NbClust* all'algoritmo k-means ed osserviamo che il numero ottimale di gruppi proposto dalla maggioranza degli indici risulta essere 2, esattamente come per i metodi gerarchici ward e completo.

| Numero di Cluster ottimale | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Numero di Indici | 2 | 1 | 9 | 6 | 7 | 1 |

Ricapitolando, per il dataset completo, la maggior parte degli indici individua 2 cluster come soluzione ottimale ad eccezione del metodo del gomito, l'unico che si differenzia particolarmente proponendo 4 gruppi. Questo ci suggerisce che, lavorando su tutte le variabili quantitative e le loro distanze matematiche, si dovrebbe prediligere una suddivisione degli individui in due gruppi a differenza della classificazione originale in tre categorie.

**Dataset ridotto**

Ripetiamo l'analisi servendoci ora del dataset ridotto, che contiene solo le variabili quantitative che hanno differenze significative in termini di media per i diversi livelli di stress, cioè il livello di caffeina assunta, il livello della pressione del sangue, i livelli di zucchero nel sangue e lo screen time. Utilizzando il metodo gerarchico del legame ward, la maggioranza degli indici propone come k ottimale 3, esattamente come il raggruppamento originale per i livelli di stress.

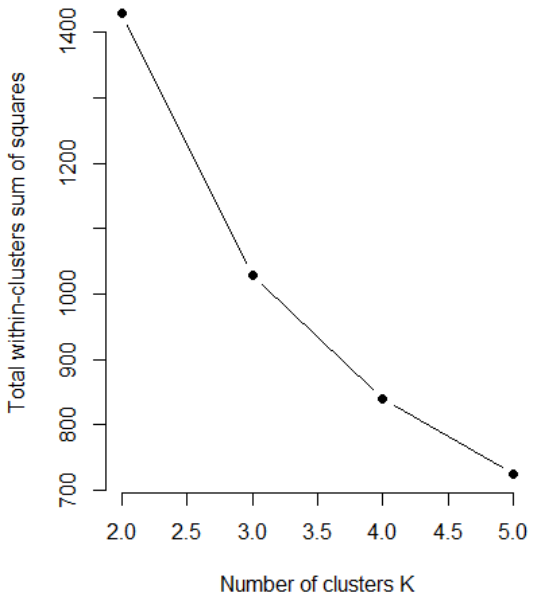
| Numero di Cluster ottimale | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Numero di Indici | 2 | 1 | 6 | 12 | 4 | 1 |

Anche con il metodo del legame completo, il numero ottimale di cluster che gli indici ci suggeriscono rimane 3 sebbene 9 indici convergano nuovamente su 2 cluster.

| Numero di Cluster ottimale | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Numero di Indici | 2 | 1 | 9 | 10 | 1 | 3 |

Anche in questo caso applichiamo l'algoritmo k-means tramite il metodo del gomito e il pacchetto *NbClust*.

.



Nel primo caso si osserva chiaramente dal grafico che la WSS diminuisce in maniera elevata passando da 2 a 3 cluster, mentre con l'aggiunta di un quarto gruppo, diminuisce in misura minore; di conseguenza riteniamo che il gomito in questo caso sia in k=3.

Nel secondo caso *NbClust,* che non tiene conto solamente del valore della WSS ma si serve di diversi indici con criteri differenti, conferma un numero di k ottimale pari a 2 anche con il dataset ridotto.

| Numero di Cluster ottimale | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Numero di Indici | 2 | 1 | 11 | 2 | 8 | 2 |

Nel complesso, utilizzando il dataset ridotto, la maggioranza dei metodi ha individuato come numero di cluster ottimale k=3, risultato conforme alla classificazione originale. Le 4 variabili utilizzate in questo caso possono essere prese come riferimento per attuare una suddivisione in gruppi ben distribuita degli individui, risultando sufficientemente discriminanti e portando alla luce un risultato migliore a quello ottenuto con l'analisi del dataset completo.

## ANALISI PREVISIVA CON K-MEANS SU SPARK

Per concludere la nostra analisi, consideriamo un approccio confermativo-previsivo applicando nuovamente il metodo delle k-means, servendoci, questa volta, del pacchetto *sparklyr* che permette di connetterci all'interfaccia del software *Spark*, consentendoci di applicare tecniche di machine learning.

A differenza dell'approccio precedente tramite metodo del gomito e pacchetto *NbClust,* il cui scopo era determinare il numero ottimale di cluster in cui classificare i nostri individui, in questo caso, l'obiettivo è valutare, studiando i valori delle variabili quantitative disponibili, la capacità del modello di prevedere per ogni individuo il gruppo di appartenenza (livello di stress).

Pur suddividendo il dataset in training set e test set, non si tratta però di un modello supervisionato vero e proprio, in quanto l'algoritmo k-means non utilizza la variabile target relativa al livello di stress, bensì raggruppa le osservazioni in k cluster omogenei sulla base della somiglianza delle loro caratteristiche. Di conseguenza determina i 3 gruppi con i relativi centroidi sul training set e successivamente inserisce gli individui del test set nei cluster creati nella fase di training, ogni individuo rispettivamente nel cluster il cui centroide ha una distanza minore da lui. Alla fine confronteremo questi 3 cluster con la classificazione originale per livello di stress.

Innanzitutto, affinchè Spark possa lavorare in maniera corretta sui dati, abbiamo convertito la matrice standardizzata StressZ in un dataframe e successivamente a questo dataframe abbiamo incluso la nostra variabile qualitativa di riferimento Stress Detection per confrontare le previsioni con la distribuzione originale.

Per prima cosa abbiamo caricato il dataframe su *Spark* tramite il pacchetto *sparklyr* e lo abbiamo denominato “stressdata”; il dataset è stato poi suddiviso casualmente in 2 gruppi, uno di training composto dal 70% delle osservazioni, ed uno di test, composto dal restante 30%, e abbiamo fissato un seed tale da poter riprodurre i risultati in modo tale da ottenere sempre la stessa ripartizione di dati. Il gruppo di training viene utilizzato per costruire il modello, utilizzando l’algoritmo k-means, che individua i cluster ed i relativi centroidi, mentre il gruppo di test comprende le osservazioni sulle quali verrà testato il modello ottenuto nella fase precedente. Le osservazioni classificate per livello di stress sono partizionate in questo modo:

**TRAINING SET**

| Livello di Stress | Numero di osservazioni |
|---|---|
| **Low** | 113 |
| **Medium** | 200 |
| **High** | 203 |
| | totale = 516 osservazioni |

**TEST SET**

| Livello di Stress | Numero di osservazioni |
|---|---|
| **Low** | 49 |
| **Medium** | 110 |
| **High** | 98 |
| | totale = 257 osservazioni |

Applichiamo l’algoritmo k means al gruppo di training definendo k=3, lavorando su tutte le 9 variabili quantitative del dataframe mentre la variabile qualitativa target Stress Detection la manteniamo come riferimento per la valutazione finale, sebbene questa non venga utilizzata da k means direttamente.
Tutte le osservazioni vengono così ripartite in 3 cluster numerati come 0, 1 e 2, ma questi valori non hanno un significato intrinseco e di conseguenza, per interpretarli, devono essere confrontati con la ripartizione per livello di stress originale in modo tale da essere prima di tutto definiti. Confrontando le colonne “label” e “stress detection” verifichiamo che al cluster 0 è associato il livello di stress high, al cluster 1 medium, mentre al cluster 2 low. Le previsioni del modello sul gruppo di appartenenza di ogni individuo sono invece riportate come “prediction”.

Successivamente estraiamo i centroidi dei gruppi per ciascuna variabile quantitativa, i cui valori sono ovviamente standardizzati, e la numerosità di ogni cluster, che riportiamo in una tabella:

| Livello di stress a cui fa riferimento | Numero del cluster | Size del gruppo |
|---|---|---|
| **High** | 0 | 211 |
| **Medium** | 1 | 189 |
| **Low** | 2 | 116 |

Passiamo ora alla fase di previsione vera e propria, in cui il modello che ha costruito una strategia per definire i gruppi, trovato nella fase di training, viene applicato alle osservazioni del gruppo di test, in modo tale da classificare ogni osservazione nel cluster il cui centroide risulta meno distante.

Riportiamo le sue previsioni in una tabella affiancando anche la reale distribuzione degli individui rispetto a stress detection:

| Livello di stress a cui fa riferimento | Distribuzione reale | Previsione |
|---|---|---|
| **High** | 98 | 115 |
| **Medium** | 110 | 86 |
| **Low** | 49 | 56 |

Per valutare i risultati ottenuti ci serviamo di una tabella di confusione, la quale ci permette di confrontare le classi reali (riportate sulle righe) con i cluster predetti( riportati sulle colonne).

| | **Medium** | **High** | **Low** |
|---|---|---|---|
| **Medium** | 64 | 16 | 18 |
| **High** | 44 | 48 | 18 |
| **Low** | 7 | 22 | 20 |

Osserviamo che i valori sulla diagonale principale rappresentano le classificazioni corrette (il cluster reale coincide con il cluster predetto), mentre i valori fuori la diagonale indicano gli errori di classificazione (cluster reale diverso dal cluster predetto).

A partire da questa tabella, calcoliamo l'error rate, ovvero il numero totale di errori di classificazione, come differenza tra il totale delle osservazioni nella tabella e la somma dei valori sulla diagonale, che rappresentano le classificazioni corrette.

Nel nostro caso:

error rate=totale delle osservazioni−somma della osservazioni sulla diagonale=125

Questo risultato indica che 125 osservazioni su 257 non ricadono sulla diagonale principale e risultano quindi assegnate a un cluster diverso da quello reale, evidenziando un numero di errori complessivo non trascurabile.

Infine calcoliamo l'accuratezza come misura complementare all'error rate: rapportiamo il numero di errori al totale delle osservazioni e sottraiamo tale quota da 1, in maniera tale da ottenere la percentuale di osservazioni correttamente classificate dal modello.

Applicando la formula otteniamo:
accuracy = 1 - (error rate / totale osservazioni)= 0,513

Questo valore indica che circa il 51,3% delle osservazioni viene classificato correttamente, evidenziando una capacità predittiva del modello complessivamente discreta.

**Dataset ridotto**

Ripetiamo l'analisi servendoci ora del dataset ridotto, che contiene solo le variabili quantitative che hanno differenze significative in termini di media per i diversi livelli di stress, cioè il livello di caffeina assunta, il livello della pressione del sangue, i livelli di zucchero nel sangue e lo screen time, per vedere se riusciamo a migliorare i risultati ottenuti.
Denominiamo il dataset sul quale lavoreremo in StressZ_ML_ridotto e suddividiamolo, analogamente a come abbiamo fatto precedentemente, in un gruppo di training e in uno di test .

**TRAINING SET**

| Livello di Stress | Numero di osservazioni |
|---|---|
| **Low** | 104 |
| **Medium** | 217 |
| **High** | 195 |
| | totale = 516 osservazioni |

**TEST SET**

| Livello di Stress | Numero di osservazioni |
|---|---|
| **Low** | 58 |
| **Medium** | 93 |
| **High** | 106 |
| | totale = 257 osservazioni |

Applichiamo l'algoritmo k-means al gruppo di training definendo k=3, questa volta solo sulle 4 variabili quantitative del dataframe.
Confrontando le colonne "label" e "stress detection" verifichiamo che al cluster 0 è ora associato il livello di stress medium, al cluster 1 high, mentre al cluster 2 low. Le previsioni del modello sul gruppo di appartenenza di ogni individuo sono invece riportate come "prediction".
Successivamente estraiamo i centroidi dei gruppi per ciascuna variabile quantitativa e la numerosità di ogni cluster, che riportiamo in una tabella:

| Livello di stress a cui fa riferimento | Numero del cluster | Size del gruppo |
|---|---|---|
| **Medium** | 0 | 193 |
| **High** | 1 | 189 |
| **Low** | 2 | 134 |

Riportiamo le previsioni del modello applicato agli individui del test set in una tabella, affiancando anche la reale distribuzione degli individui rispetto a stress detection.

| Livello di stress a cui fa riferimento | Distribuzione reale | Previsione |
|---|---|---|
| **Medium** | 93 | 84 |
| **High** | 106 | 107 |
| **Low** | 58 | 66 |

Per valutare i risultati ottenuti ci serviamo anche in questo caso di una tabella di confusione:

| | **Medium** | **High** | **Low** |
|---|---|---|---|
| **Medium** | 32 | 55 | 6 |
| **High** | 17 | 32 | 57 |
| **Low** | 35 | 20 | 3 |

Ci rendiamo conto immediatamente di aver ottenuto risultati peggiori di quelli precedenti, dato che nella diagonale principale si trovano solo 67 individui. In particolare Il gruppo low risulta quasi completamente formato da osservazioni che non fanno parte del cluster reale con unicamente 3 individui stimati correttamente.

In questo caso otteniamo un error rate pari a 190 su 257 osservazioni che determina un tasso di accuratezza pari al 26%.

**CONSIDERAZIONI FINALI**

In conclusione possiamo affermare che il modello previsivo dell'algoritmo k-means non ci ha fornito dei risultati propriamente accurati nella classificazione per livello di stress.

Se nella fase confermativa avevamo notato che l'utilizzo di un dataset ridotto ci permetteva di ottenere un numero di cluster coerente con la classificazione reale, nella fase previsiva, invece, la riduzione delle variabili non migliora il modello, in quanto l'accuratezza diminuisce notevolmente passando dal 51% al 26%.

Dunque riteniamo che il modello utilizzato non è in grado di prevedere in modo efficace quanto sia realmente stressato ogni individuo. Probabilmente le variabili da noi scelte non sono sufficienti per una corretta stima, la quale potrebbe dipendere da altre caratteristiche personali riguardanti aspetti psicologici ed emotivi.