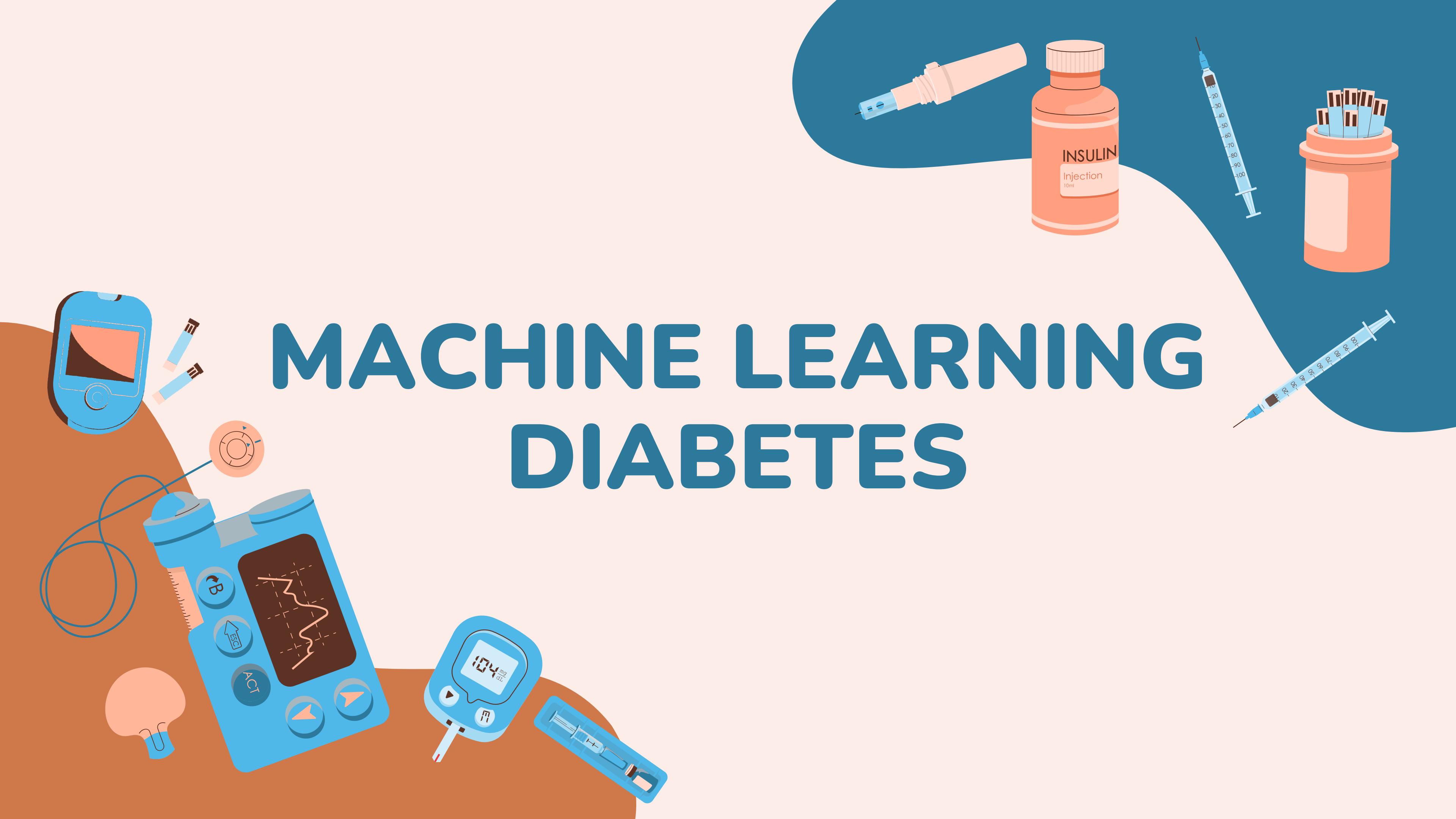


# MACHINE LEARNING DIABETES



# INTRODUZIONE

Questo progetto prevede la collaborazione con un'organizzazione sanitaria dedicata alla prevenzione e al monitoraggio delle malattie croniche. Il mio lavoro sosterrà il team medico nel prevedere l'evoluzione del diabete nei pazienti, consentendo interventi più tempestivi e mirati.

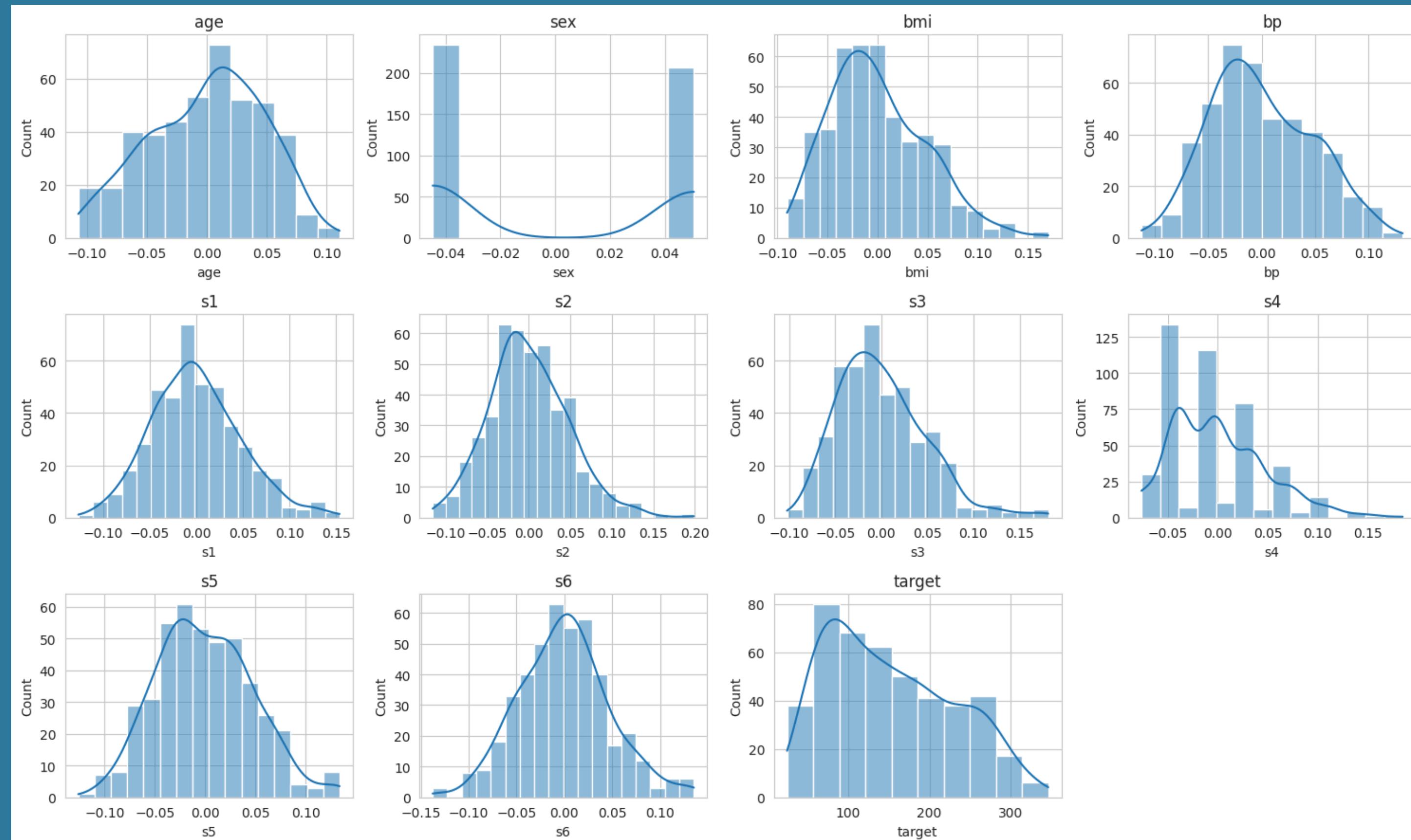
Utilizzerò il database Diabetes di scikit-learn per sviluppare un modello predittivo.

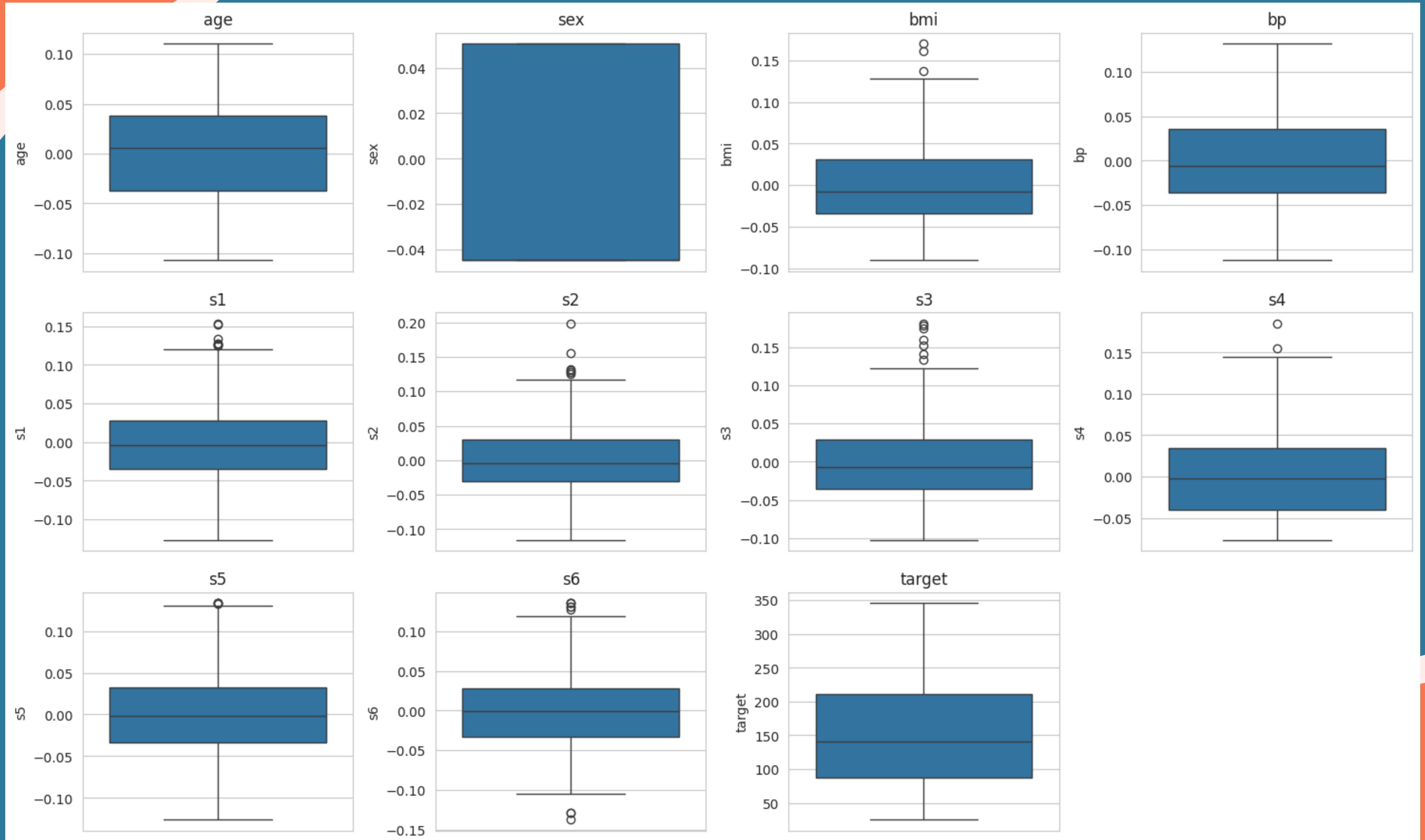


# FASI DEL PROGETTO

1. Setup Iniziale e Caricamento Dati
2. Analisi Esplorativa dei Dati (EDA)
3. Considerazioni e Approcci per la Modellazione basati sull'EDA
4. Preparazione dei Dati per la Modellazione (Split Dataset)
5. Test Modelli Scelti
6. Analisi e Spiegazione dei Risultati dei Modelli e Scelta per la Proseguimento
7. Tecniche per Migliorare le Performance del Lasso Regressor
8. Ottimizzazione Estesa degli Iperparametri (Lasso)
9. Analisi Comparativa delle Ottimizzazioni del Lasso Regressor
10. Feature Engineering Mirato
11. Feature Engineering Mirato (Test nuove feature con Iperparametri Estesi)
12. Analisi Comparativa delle Ottimizzazioni del Lasso Regressor con e senza Feature Engineering
13. Valutazione Finale di Lasso Regressor sul Set di Test
14. Risultati Finali e Analisi delle Performance del Lasso Regressor sul Set di Test

# ANALISI ESPLORATIVA





# ANALISI GRAFICI DI DISTRIBUZIONE

## Istogrammi

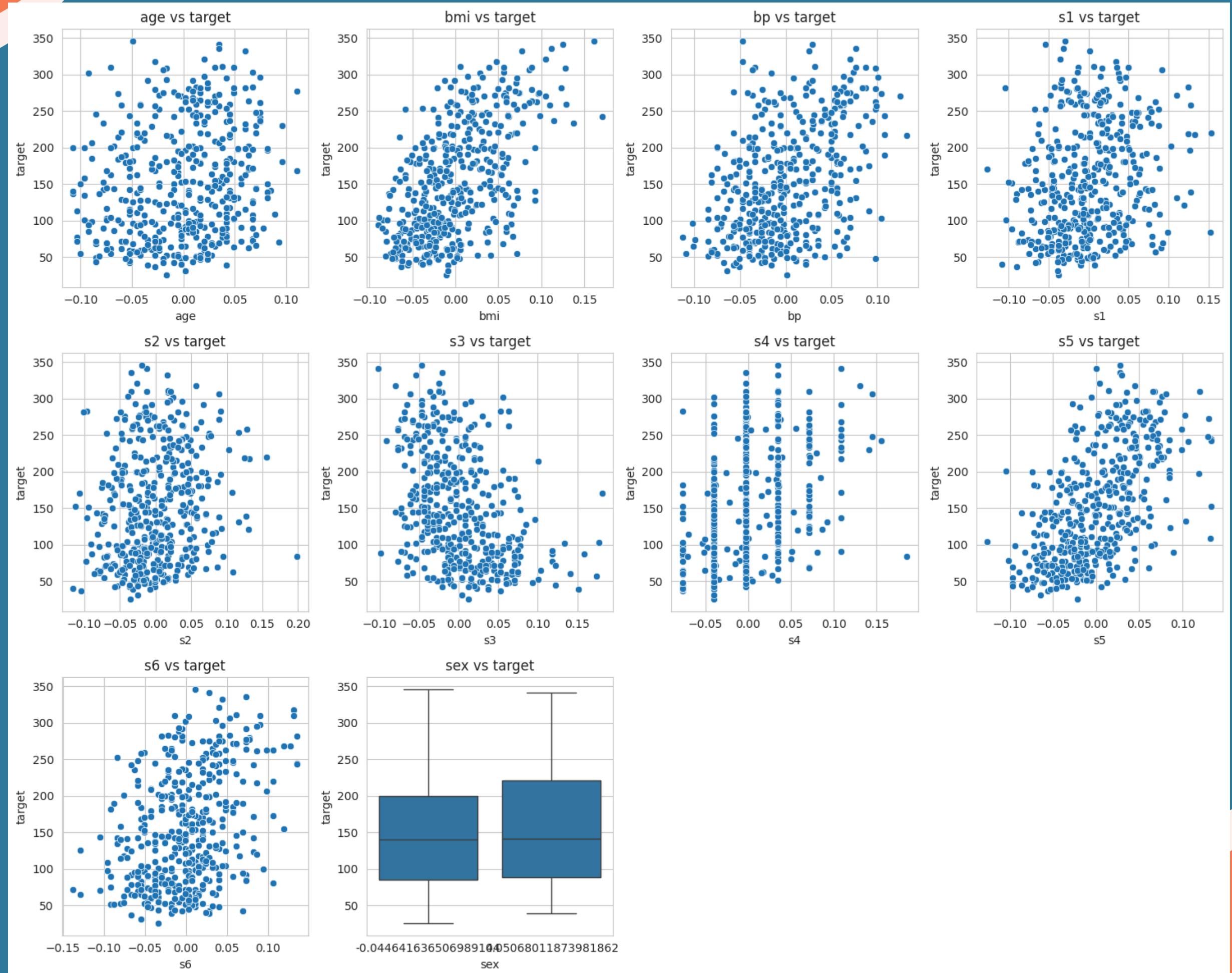
- Le variabili predittive ('age', 'bmi', 'bp', 's1', 's2', 's3', 's5', 's6') mostrano distribuzioni vicine alla normale, con qualche asimmetria.
- La variabile 'sex' è binaria e bilanciata tra i due gruppi.
- La variabile 's4' presenta una distribuzione irregolare e multimodale.
- La variabile target è asimmetrica: la maggior parte dei pazienti mostra una progressione bassa del diabete, ma esistono casi con valori molto elevati.

## Box Plot

- Confermano centratura e scalatura delle variabili (mediane vicine a zero, range simile).
- Evidenziano diversi outlier in 'bmi', 'bp', 's1', 's2', 's3', 's4', 's5', 's6'.

## Sintesi

- L'analisi univariata ha permesso di:
  - Valutare la forma delle distribuzioni.
  - Confermare il preprocessing (scalatura/centratura).
  - Identificare la presenza di outlier.



# ANALISI DEI GRAFICI A DISPERSIONE VS TARGET

## Correlazioni Positive Evidenti

- 'bmi', 'bp', 's5', 's6' mostrano una chiara relazione lineare positiva con la target.
- Queste variabili sono probabili predittori chiave della progressione del diabete.

## Correlazioni Meno Evidenti / Non Lineari

- 'age': dispersione uniforme, non forte predittore lineare.
- 's1', 's2', 's3', 's4': pattern più complessi.
  - 's3' con leggera tendenza negativa (biologicamente plausibile).
  - 's4' mostra bande verticali → possibile interazione con altre variabili.

## Variabile Categorica ('sex')

- Differenze lievi tra i due gruppi: mediana leggermente più alta per un sesso.
- Ampia sovrapposizione: non forte predittore da solo.

## Outlier

- Presenti in quasi tutte le variabili ('bmi', 'bp', 's1', 's2', 's3', 's4', 's5', 's6').
- Possono influenzare fortemente i modelli lineari.

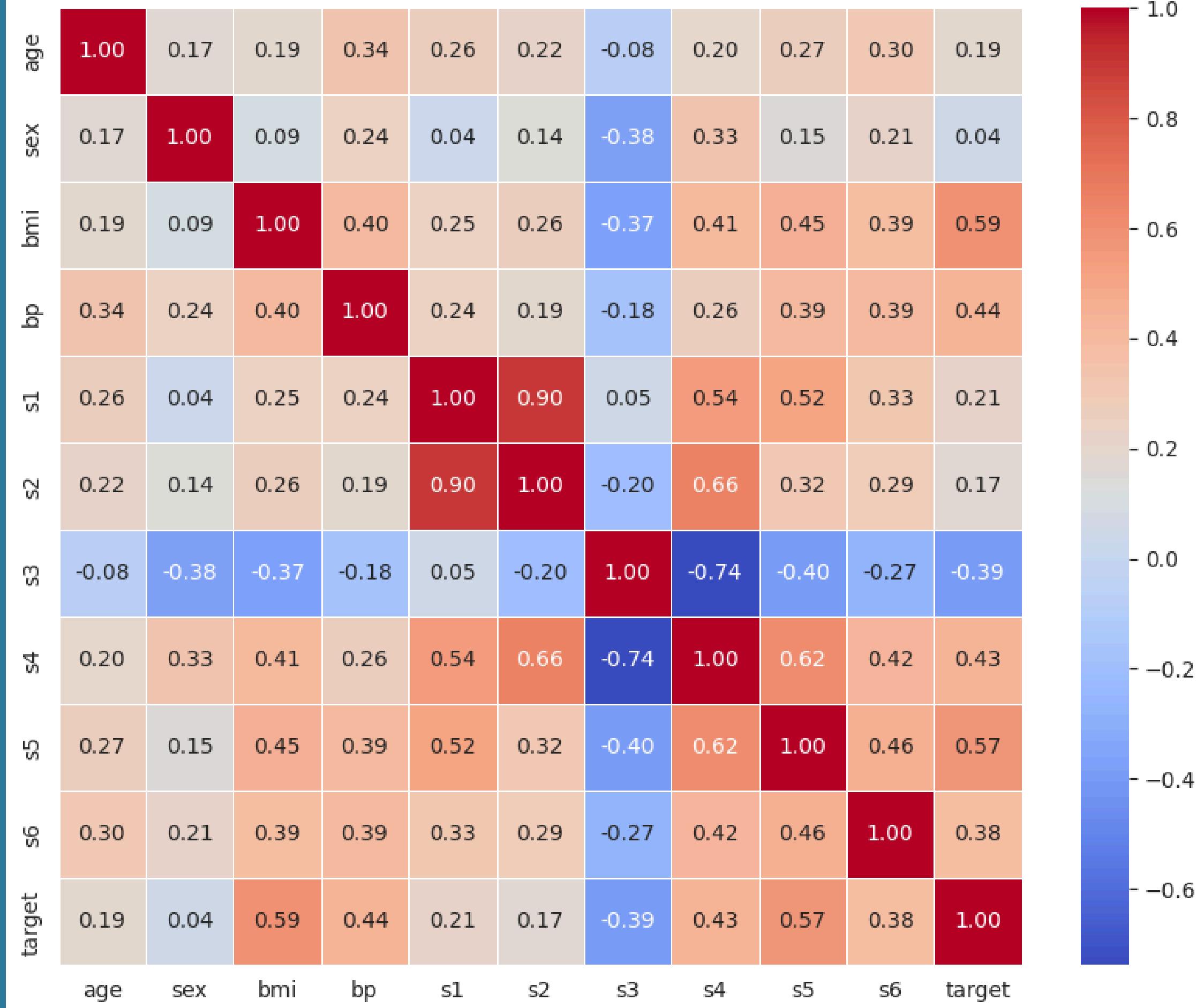
## Implicazioni per la Modellazione

- Modello lineare adatto per 'bmi', 'bp', 's5', 's6'.
- Per altre variabili:
  - valutare modelli non lineari (es. random forest).
  - considerare interazioni tra predittori.

## Sintesi

- Identificati predittori lineari forti.
- Alcune variabili richiedono approcci più sofisticati.
- Attenzione particolare agli outlier.

Matrice di Correlazione tra le Variabili del Dataset Diabete



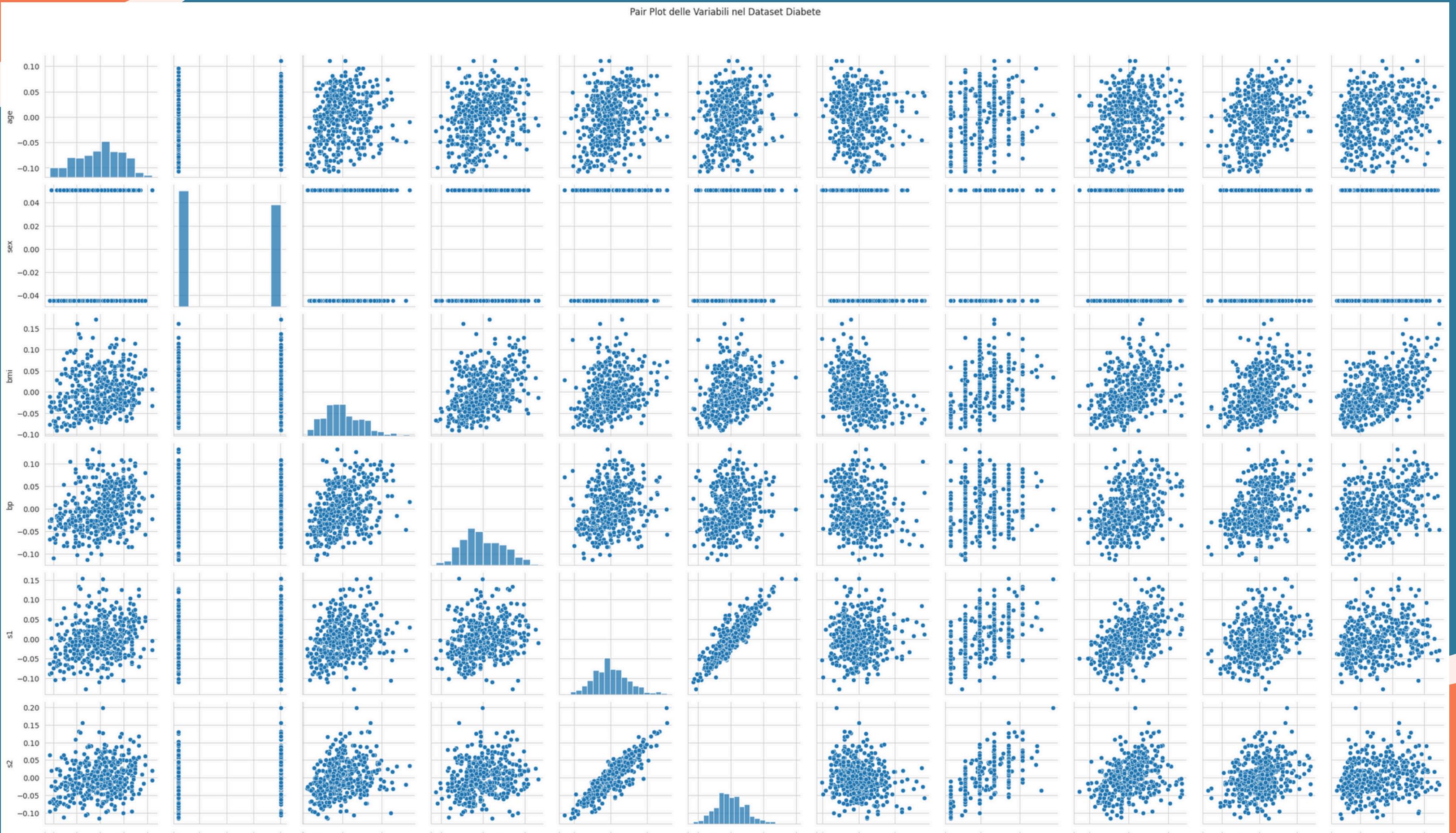
# ANALISI DELLA MATRICE DI CORRELAZIONE

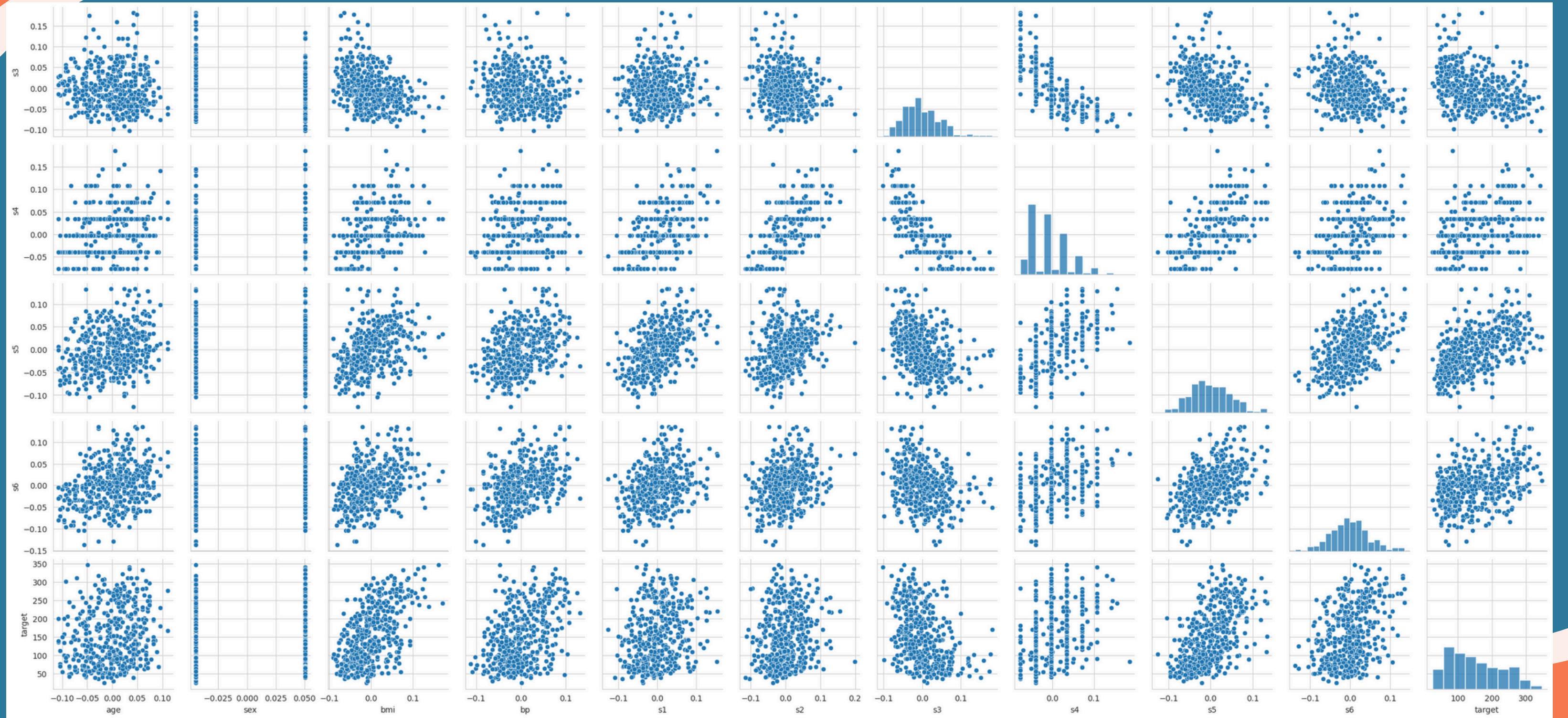
L'analisi della matrice di correlazione tramite heatmap permette di valutare quantitativamente le relazioni lineari tra tutte le variabili del dataset sul diabete.

- Le variabili più fortemente correlate con la progressione del diabete sono 'bmi' (0.59), 's5' (0.57), 'bp' (0.44) e 's6' (0.38), confermando il loro ruolo come predittori principali.
- 's3' (-0.39) mostra una correlazione negativa coerente con la funzione protettiva dell'HDL.
- Altre variabili, come 'age', 'sex', 's1', 's2' e 's4', presentano correlazioni più deboli.
- L'analisi evidenzia anche la presenza di multicollinearità tra le variabili sieriche, in particolare:
  - 's1' e 's2' (0.90), 's2' e 's4' (0.66), 's4' e 's5' (0.62)
  - correlazioni negative significative come 's3' e 's4' (-0.74)

Queste interdipendenze suggeriscono che, nell'uso di modelli lineari, i coefficienti potrebbero risultare poco interpretabili senza tecniche di regolarizzazione come Ridge o Lasso.

Inoltre, l'analisi evidenzia quali variabili potrebbero influenzare maggiormente la progressione del diabete e quali combinazioni di variabili richiedono attenzione nella modellazione. La heatmap, quindi, non solo conferma i predittori principali, ma fornisce anche indicazioni sulla struttura interna del dataset e sulle possibili interazioni tra le misurazioni sieriche.





# ANALISI DEL PAIR PLOT

Il pair plot fornisce una panoramica globale del dataset, combinando distribuzioni univariate (diagonale) e relazioni bivariate (fuori diagonale).

- Distribuzioni univariate: molte variabili mostrano andamenti quasi normali; 'sex' è binaria; 's4' ha distribuzione irregolare; la target è concentrata su valori bassi con coda a destra.
- Relazioni con la target: chiara correlazione positiva per 'bmi', 'bp', 's5', 's6'; correlazione negativa per 's3' (HDL); le altre variabili presentano legami più deboli.
- Relazioni tra predittori: evidenti correlazioni lineari tra variabili sieriche ('s1', 's2', 's3', 's4', 's5'), con pattern sia positivi che negativi ciò conferma della multicollinearità.
- Pattern visivi: alcune coppie di variabili mostrano nuvole di punti strette e allungate (correlazione forte), altre dispersioni ampie (relazioni deboli o non lineari).
- Outlier: osservabili in più variabili, con potenziale impatto sui modelli lineari.

Conclusione:

Il pair plot agisce come una mappa completa del dataset, utile per:

1. confermare distribuzioni e relazioni con la target,
2. individuare multicollinearità tra predittori,
3. rilevare outlier e possibili interazioni da approfondire in fase di modellazione.

# CONSIDERAZIONI E APPROCCI PER LA MODELLAZIONE BASATI SULL'EDA

Basandosi sull'analisi esplorativa dei dati, osservando distribuzioni, relazioni bivariate e outlier, emergono alcune considerazioni chiave per la modellazione:

- Dati scalati e outlier:
- I dati sono già centrati e scalati; gli outlier identificati nei box plot potrebbero rappresentare valori reali estremi o possibili errori. Rimuoverli senza conoscenza del dominio potrebbe far perdere informazioni importanti.
- Sensibilità dei modelli agli outlier:
- Modelli come la regressione lineare standard possono essere influenzati dai valori estremi, alterando significativamente i coefficienti stimati.
- Multicollinearità:
- La heatmap evidenzia forte correlazione tra alcune variabili sieriche, suggerendo multicollinearità che può ridurre interpretabilità e stabilità dei modelli lineari.

Approcci suggeriti:

- Random Forest Regressor: robusto agli outlier, non sensibile alla multicollinearità e capace di catturare relazioni non lineari.
- Lasso Regressor: regressione lineare con penalità L1; forza alcuni coefficienti a zero, selezionando automaticamente le feature più rilevanti e creando modelli interpretabili.
- Support Vector Regressor (SVR): cattura relazioni complesse e può ridurre l'impatto degli outlier grazie al margine di tolleranza definito dai parametri.

Sintesi:

L'EDA suggerisce di combinare modelli robusti agli outlier e regolarizzati, sfruttando approcci sia lineari che non lineari per ottenere predizioni accurate e interpretabili.



# ANALISI RISULTATI DEI MODELLI

Dopo aver addestrato e ottimizzato Lasso, SVR e Random Forest Regressor sul set di addestramento e valutato le performance sul set di validazione, emergono alcune evidenze importanti.

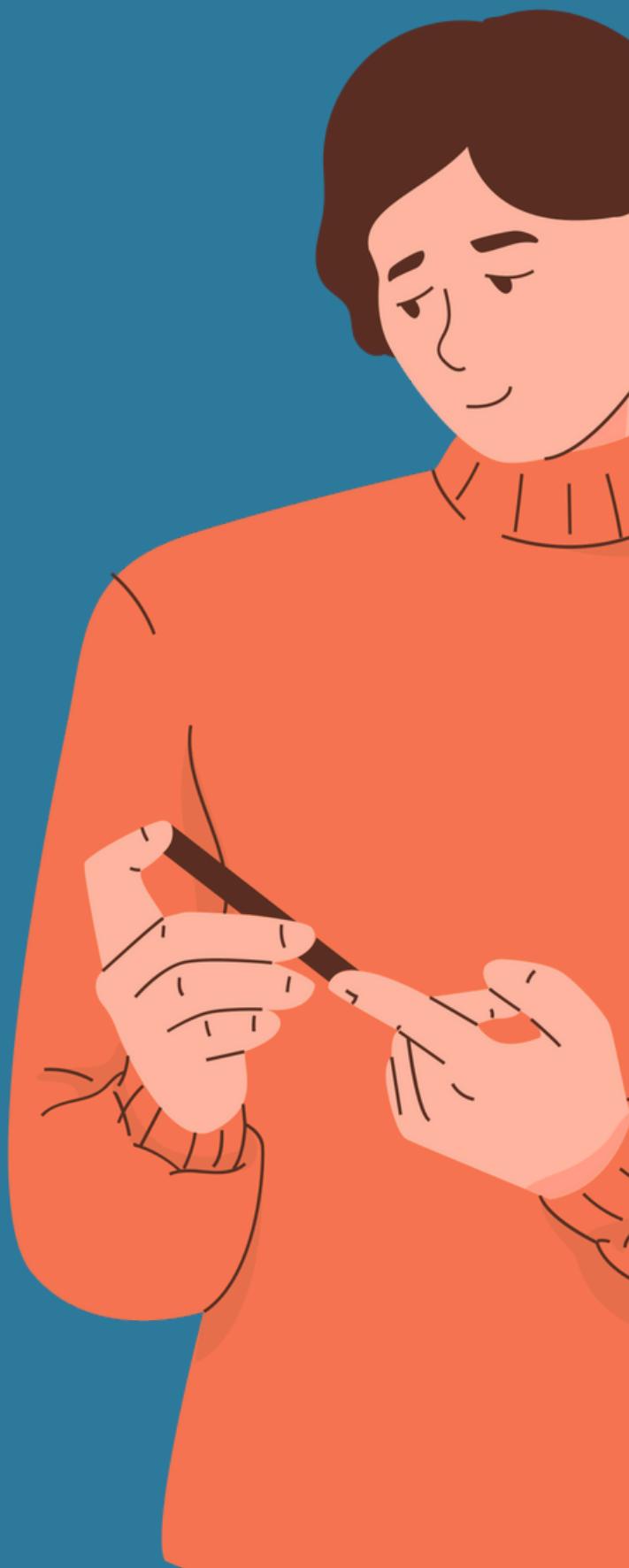
Le metriche sul set di validazione mostrano che il Lasso Regressor ottiene il MSE più basso (2847.94) e il MAE più ridotto (44.40), indicando che le sue previsioni sono, in media, più vicine ai valori reali della progressione del diabete. Anche il valore di  $R^2$  più alto (0.4703) suggerisce che Lasso spiega circa il 47% della variabilità della variabile target, superiore a SVR e Random Forest.

Pur non essendo la differenza enorme rispetto a Random Forest, il Lasso mostra una leggera superiorità e offre vantaggi teorici importanti: grazie alla penalità L1, riduce automaticamente i coefficienti delle feature meno rilevanti e può azzerarne alcune, aiutando a gestire la multicollinearità evidenziata nell'EDA. Questo rende il modello più interpretabile e stabile, un aspetto utile soprattutto considerando che i dati sono scalati e centrati.

Considerazioni aggiuntive:

- Random Forest cattura relazioni non lineari ma non seleziona automaticamente le feature.
- SVR può gestire relazioni complesse ma, nel set attuale, mostra performance leggermente inferiori.
- La scelta di Lasso permette di bilanciare accuratezza e interpretabilità, facilitando ulteriori miglioramenti e analisi approfondite.

In sintesi, il Lasso Regressor rappresenta la scelta più adatta per proseguire con l'ottimizzazione e lo sviluppo del modello predittivo sulla progressione del diabete.



# Tecniche per Migliorare le Performance del Lasso Regressor

Per ottimizzare ulteriormente il Lasso Regressor, saranno adottate due strategie principali:

## 1. Ottimizzazione estesa degli iperparametri

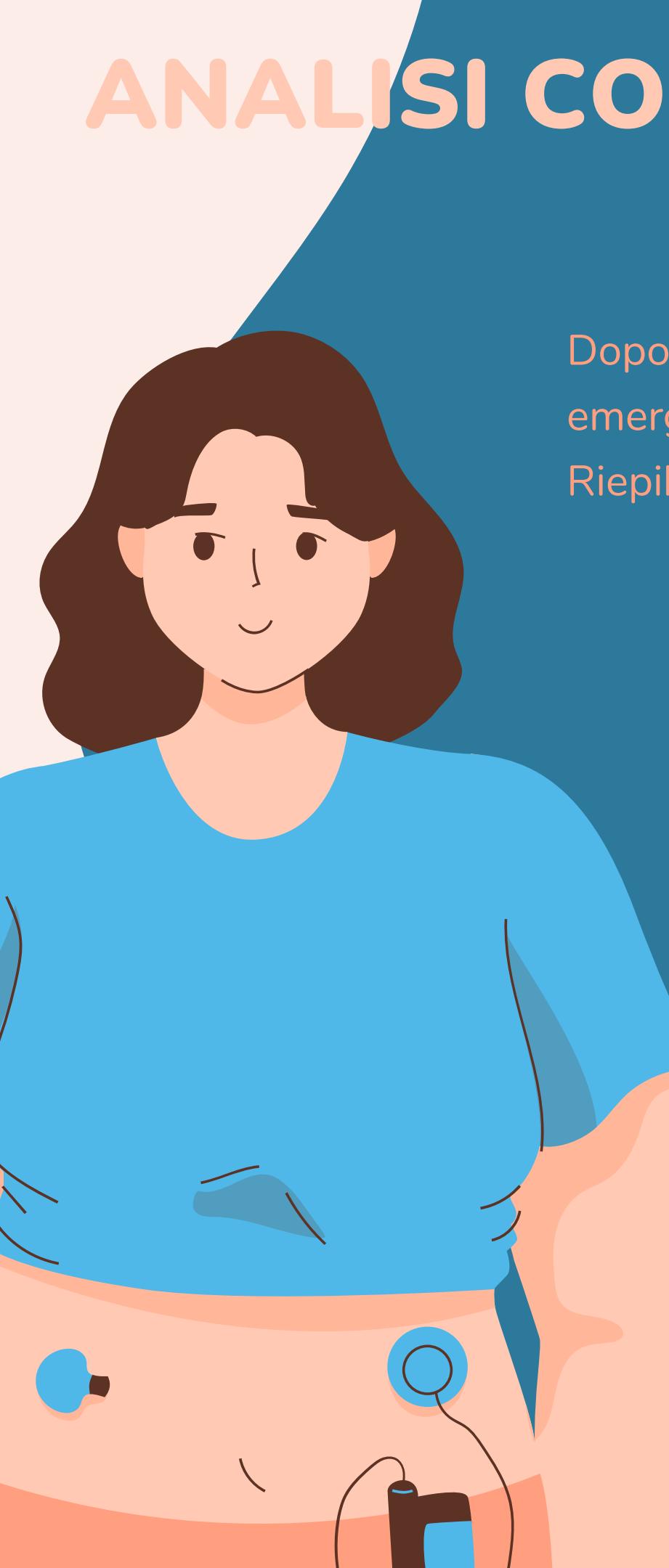
- Rieseguirò la Grid Search per trovare il miglior valore di alpha, usando una griglia più ampia e densa in scala logaritmica, con cross-validation a 10 fold. Questo approccio permette di esplorare più a fondo lo spazio dei parametri e identificare il livello ottimale di regolarizzazione, sia con che senza le nuove feature ingegnerizzate.

## 2. Feature engineering mirato

- Creerò nuove feature combinando le variabili più correlate con la target: 'bmi', 'bp', 's5', 's6'.
- Verranno aggiunti termini di interazione (prodotti tra le variabili) per catturare possibili relazioni non lineari.
- L'obiettivo è rappresentare meglio la complessità dei dati, consentendo al modello lineare di sfruttare informazioni aggiuntive.
- Procederò quindi con il Lasso Regressor, sperimentando entrambe le strategie.

Mi aspetto che l'aggiunta di feature di interazione migliori la capacità predittiva, mentre un'ottimizzazione più approfondita degli iperparametri permetterà di trovare il giusto equilibrio tra regolarizzazione e adattamento ai dati.

# ANALISI COMPARATIVA DELLE OTTIMIZZAZIONI DEL LASSO REGRESSOR



Dopo aver confrontato il Lasso Regressor con due diversi approcci di ottimizzazione degli iperparametri, emergono alcune osservazioni importanti.

Riepilogo delle performance sul set di validazione:

## Prima ottimizzazione:

- Alpha: 1.0
- MSE: 2847.94
- MAE: 44.40
- R<sup>2</sup>: 0.4703

## Ottimizzazione estesa:

- Alpha: 0.596
- MSE: 2857.47
- MAE: 44.53
- R<sup>2</sup>: 0.4685

## Considerazioni aggiuntive:

- La seconda ottimizzazione ha esplorato uno spazio di alpha più ampio e con maggiore granularità e utilizzato cross-validation a 10 fold, fornendo una stima più robusta delle performance.
- Le piccole differenze nelle metriche di validazione possono essere dovute alla variabilità dello split dei dati o alla casualità della CV.

## Decisione:

Sulla base delle metriche sul set di validazione, si procederà con il Lasso Regressor della prima ottimizzazione per la valutazione finale sul set di test.

Tuttavia, la seconda ottimizzazione rimane utile come riferimento per strategie future su dataset più grandi o più variabili.

# ANALISI COMPARATIVA DELLE OTTIMIZZAZIONI DEL LASSO REGRESSOR CON E SENZA FEATURE ENGINEERING

Dopo aver confrontato tre versioni del Lasso Regressor—ottimizzazione iniziale, ottimizzazione estesa degli iperparametri e feature engineering mirato con ottimizzazione estesa—emergono alcune osservazioni importanti.

Riepilogo delle performance sul set di validazione:

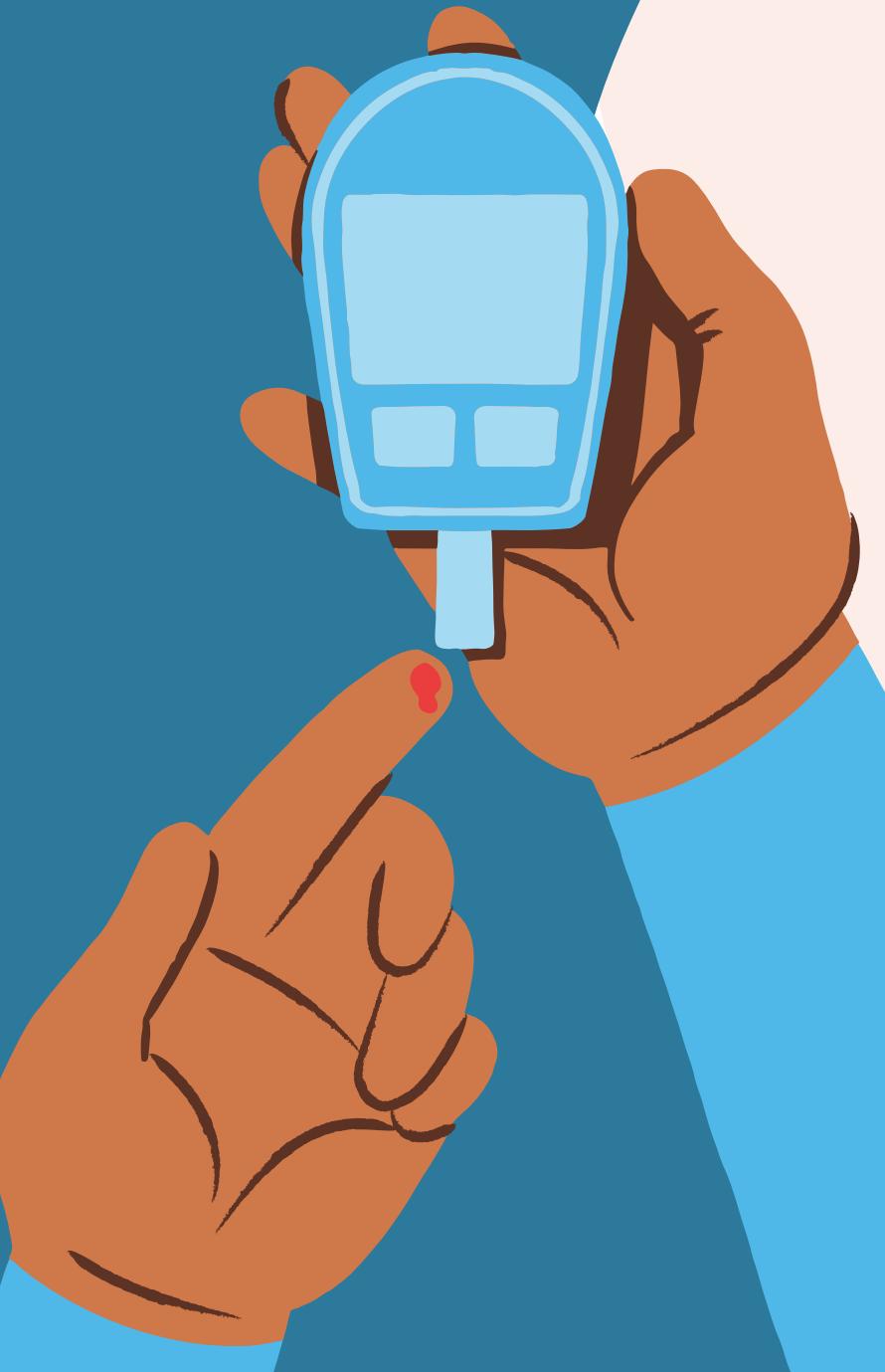
- Lasso ottimizzazione iniziale: MSE 2847.94, MAE 44.40, R<sup>2</sup> 0.4703
- Lasso ottimizzazione estesa: MSE 2857.47, MAE 44.53, R<sup>2</sup> 0.4685
- Lasso feature engineering + ottimizzazione estesa: MSE 3082.82, MAE 45.75, R<sup>2</sup> 0.4266

Analisi dei risultati:

- L'aggiunta di feature di interazione mirate non ha migliorato le performance; anzi, MSE, MAE e R<sup>2</sup> sono peggiori. Questo indica che le interazioni create non hanno fornito informazioni aggiuntive utili o hanno introdotto rumore e multicollinearità non gestita dal modello.
- Tra i due Lasso senza feature engineering, la prima ottimizzazione risulta leggermente più performante sul set di validazione. La seconda, pur esplorando uno spazio di iperparametri più ampio e usando cross-validation a 10 fold, non migliora la generalizzazione.

Conclusioni:

Il Lasso Regressor della prima ottimizzazione è il modello più adatto per la generalizzazione sul set di validazione. Le relazioni tra le variabili sembrano prevalentemente lineari, e i termini di interazione specifici testati non aggiungono valore predittivo.



# RISULTATI FINALI E ANALISI DEL LASSO REGRESSOR SUL SET DI TEST

Ho valutato il Lasso Regressor sul set di test utilizzando l'iperparametro alpha=1.0, che si era rivelato leggermente migliore durante la prima fase di validazione. I risultati confermano le performance osservate in precedenza, fornendo una stima realistica del comportamento del modello su dati completamente nuovi.

Il modello ha ottenuto un MSE di circa 2825, un MAE di circa 43 e un R<sup>2</sup> di circa 0.47. Questo significa che le previsioni del modello si discostano mediamente di 43 unità dai valori reali della progressione del diabete, e che il modello riesce a spiegare circa il 47% della variabilità nei dati. Le performance molto simili a quelle sul set di validazione indicano che non c'è stato un overfitting significativo e che il modello generalizza bene su nuovi dati.

Nonostante queste buone performance, resta una parte significativa della variabilità non spiegata, circa il 53%. Questo può dipendere da diversi fattori: le feature disponibili catturano solo in parte la complessità biologica della progressione del diabete, potrebbero esserci relazioni non lineari o interazioni tra variabili che il modello lineare fatica a catturare, oppure una componente della variabilità potrebbe essere intrinsecamente casuale o non osservabile.

In sintesi, il Lasso Regressor si conferma un modello solido e interpretabile, utile per fornire stime preliminari della progressione del diabete. Per migliorare ulteriormente le performance, si potrebbero esplorare dataset più ricchi, approcci di feature engineering più sofisticati o modelli in grado di catturare relazioni non lineari e interazioni più complesse tra le variabili.



# THANK YOU!

LINK:

<https://drive.google.com/file/d/1kbZNhb9Zonr8rswAmPwgbQIWoWhRpYr5/view?usp=sharing>