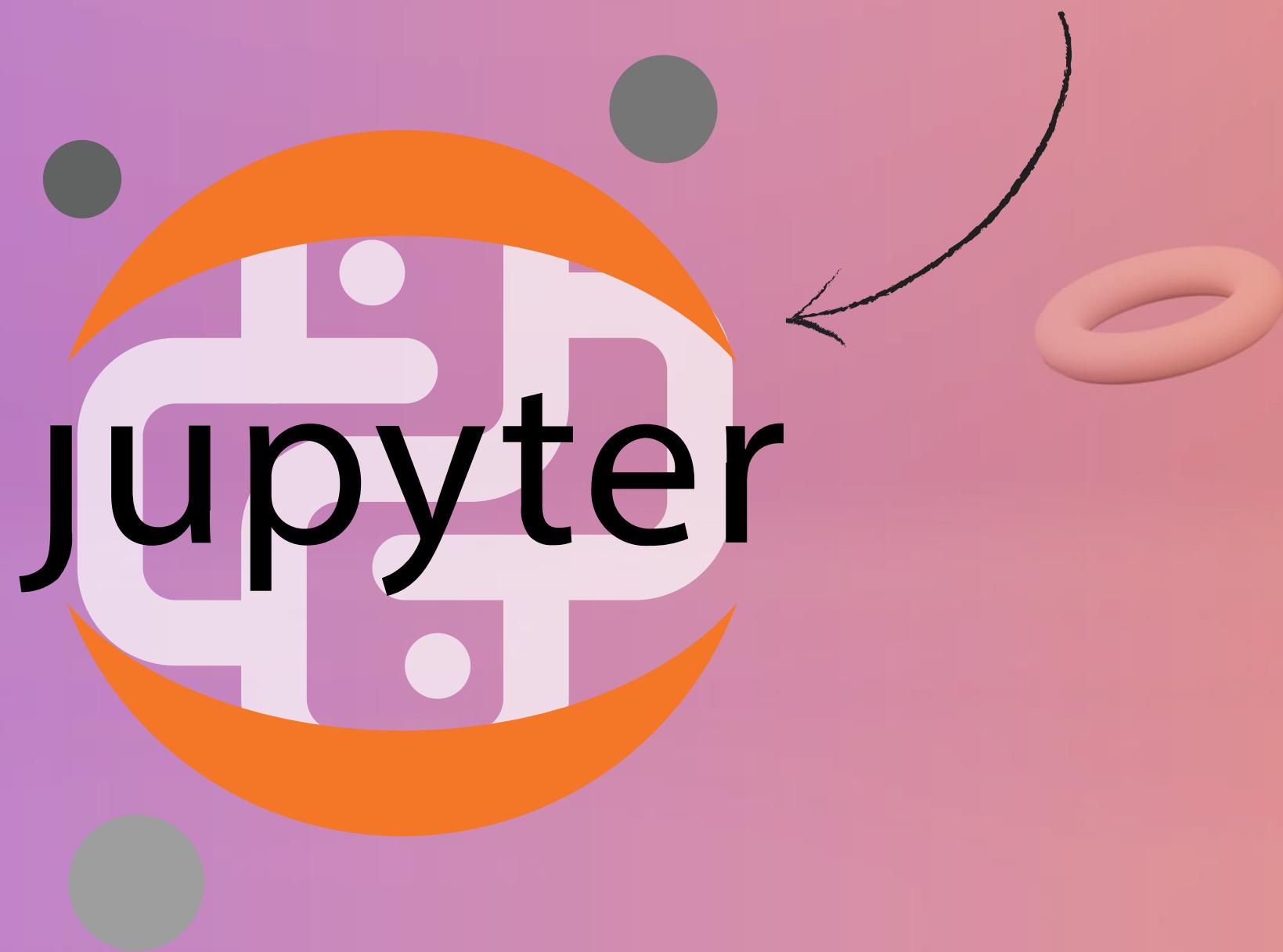


# Diabetes dataset machine learning predictions>

**Federico Gori**  
Data scientist



Questo progetto è stato sviluppato in linguaggio  
Python su Jupyter Notebook,  
**per il progetto originale clicca sul logo**



**Il Dataset verrà utilizzato per prevedere il progresso del diabete nei casi studio in base ai valori registrati nell'arco di un anno.**

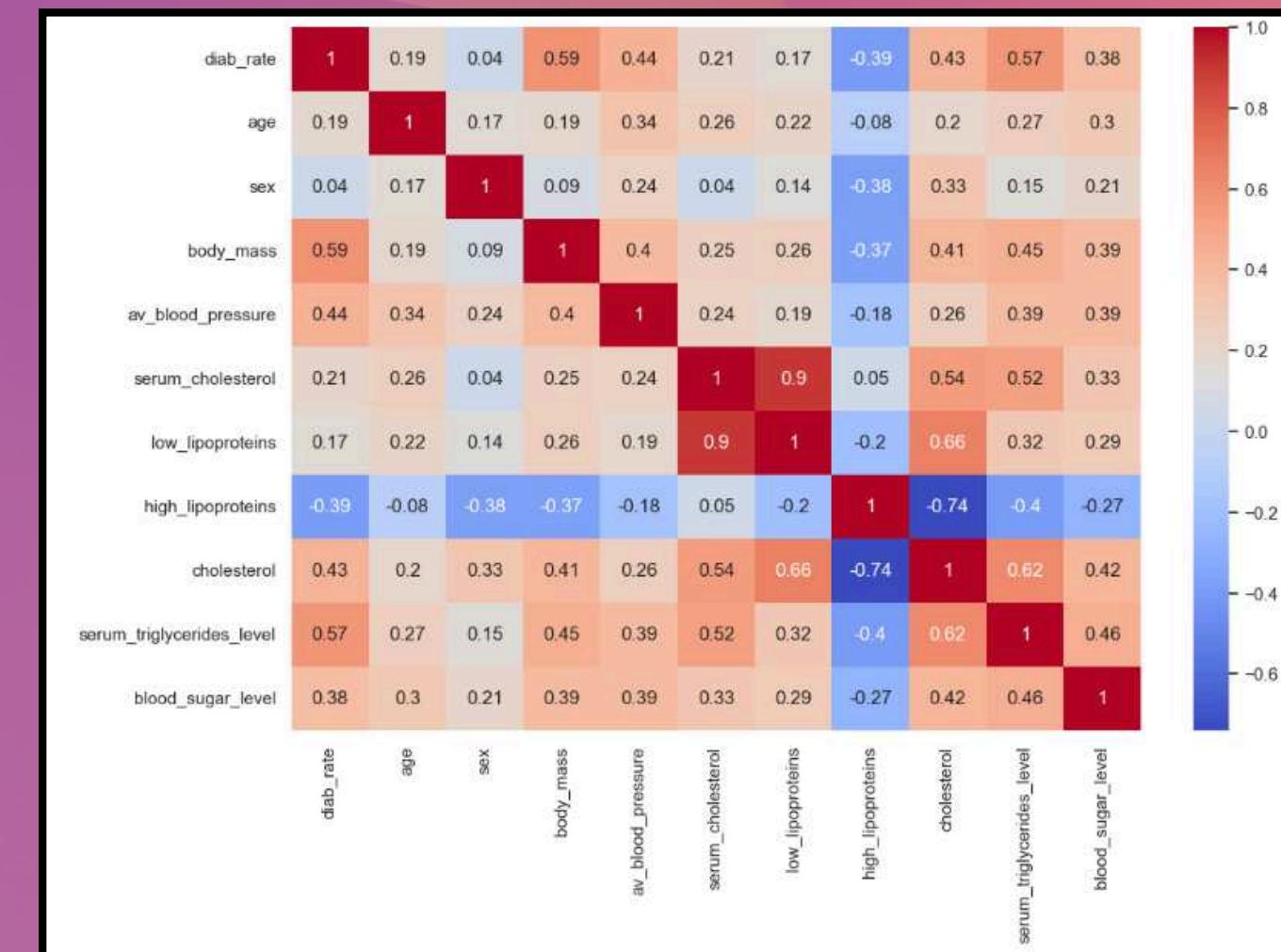
**I valori di progressione della malattia sono stati registrati attraverso un indice rappresentato da una variabile continua; per questo i modelli predittivi utilizzati faranno riferimento al criterio della regressione.**

**Le misurazioni rilevate dai 442 pazienti che compongono il dataset sono:**

- **Età**
- **Sesso**
- **BMI**: indice di massa corporea
- **BP**: pressione sanguigna media
- **S1**: misurazione totale del colesterolo nel sangue
- **S2**: misurazione delle lipoproteine a bassa densità
- **S3**: misurazione delle lipoproteine a alta densità
- **S4**: HDL/misurazione sierica del colesterolo totale
- **S5**: misurazione dei trigliceridi nel sangue
- **S6**: indice glicemico

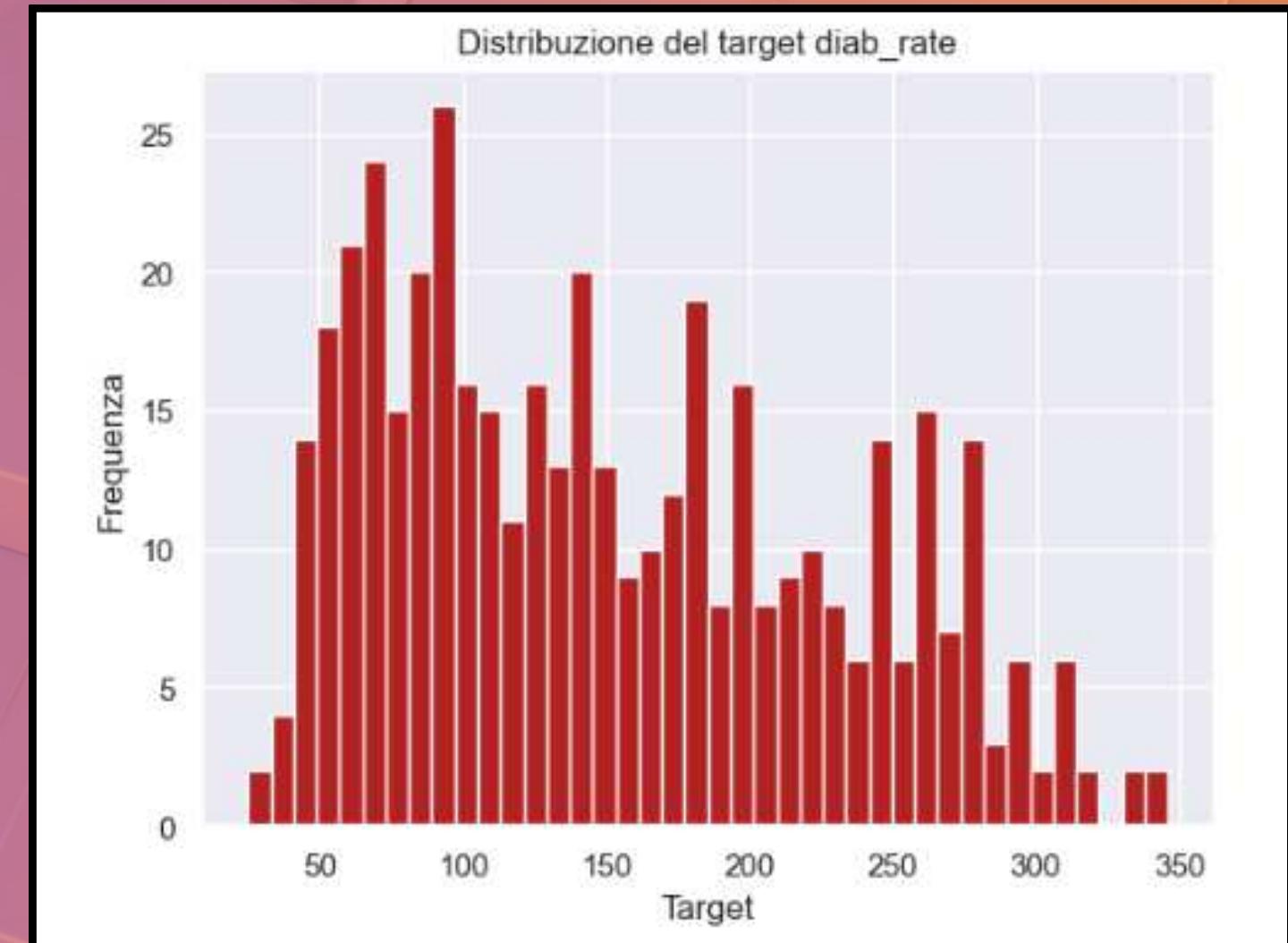
Nel verificare la correlazione lineare tra tutte le variabili del dataset si può notare che alcune features, come "low\_lipoproteins" e "serum\_cholesterol", presentano un'elevata correlazione.

Per questo saranno svolti successivamente dei test con la Features Selection per testare l'importanza o la ridondanza delle features presenti nel dataset.



**La frequenza della variabile target "diab\_rate" mostra una distribuzione asimmetrica, con picchi tra 50 e 200 nell'indice di progressione della malattia.**

**Data la natura non normale della distribuzione del target, i modelli più adatti per la sua previsione potrebbero essere di tipo non parametrico, di modo da adattarsi meglio alla tipologia dei dati e catturare in maniera più efficiente le relazioni non lineari presenti.**



## Il dataset presenta 2 tipi di features:

- Numeriche - continue
  - age
  - body\_mass
  - av\_blood\_pressure
  - serum\_cholesterol
  - low\_lipoproteins
  - high\_lipoproteins
  - cholesterol
  - serum\_triglycerides\_level
  - blood\_sugar\_level

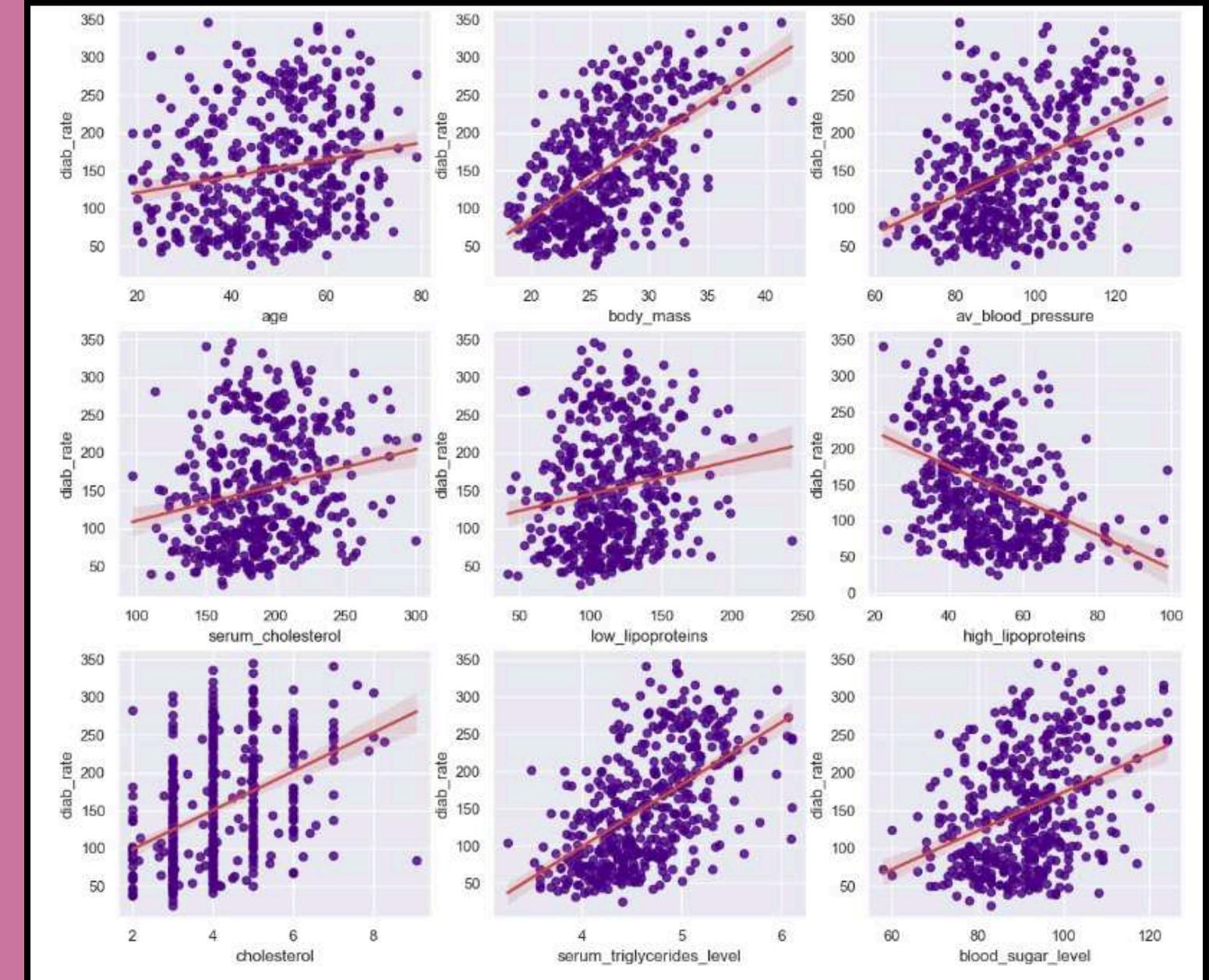
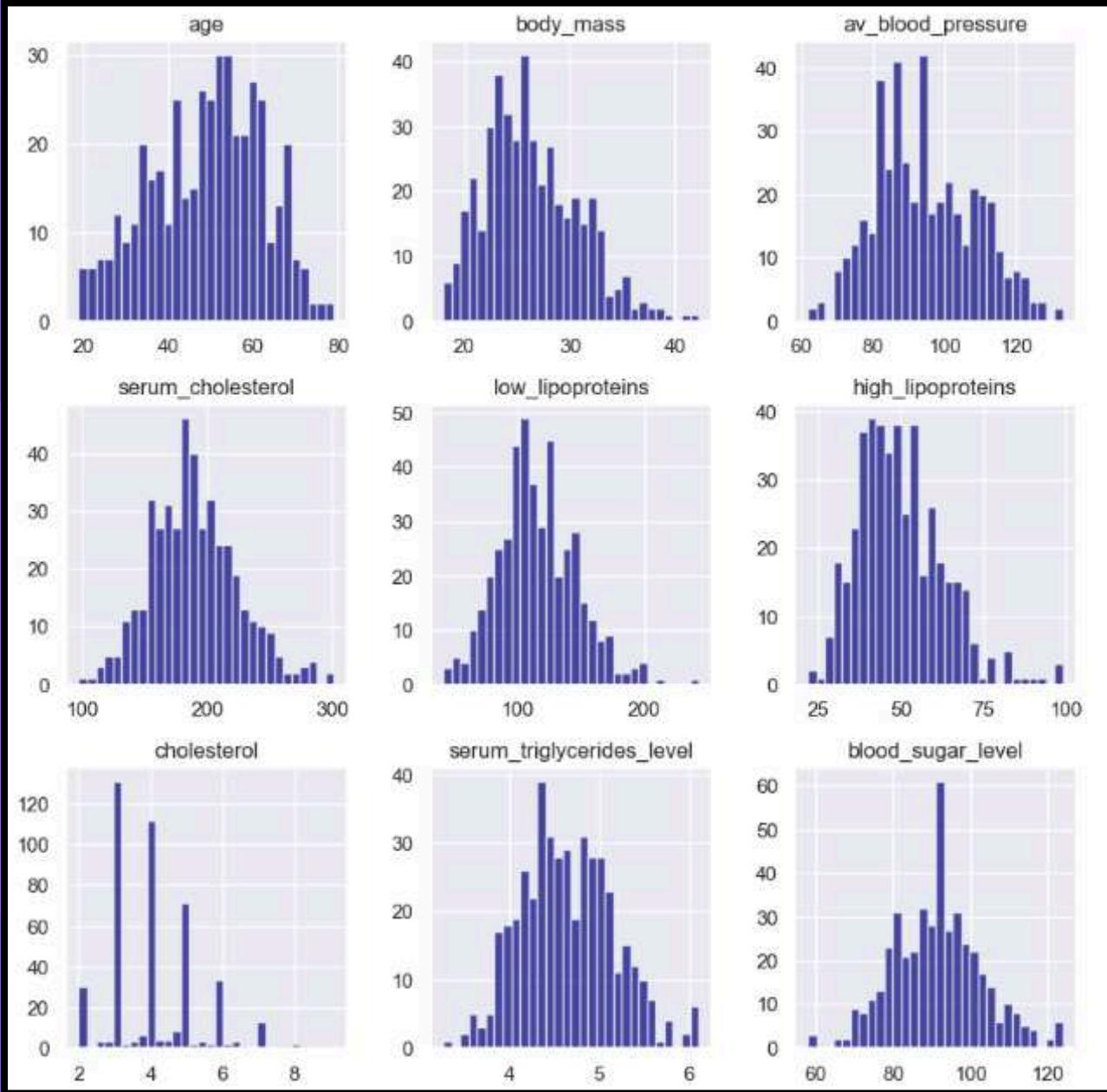
- Numeriche - discrete
  - sex

La maggior parte delle features continue mostra una distribuzione simile alla curva gaussiana.

Il rapporto delle features continue con la variabile target mostra una buona capacità predittiva delle variabili:

- `body_mass`
- `serum_triglycerides_level`
- `av_blood_pressure`
- `cholesterol`
- `blood_sugar_level`

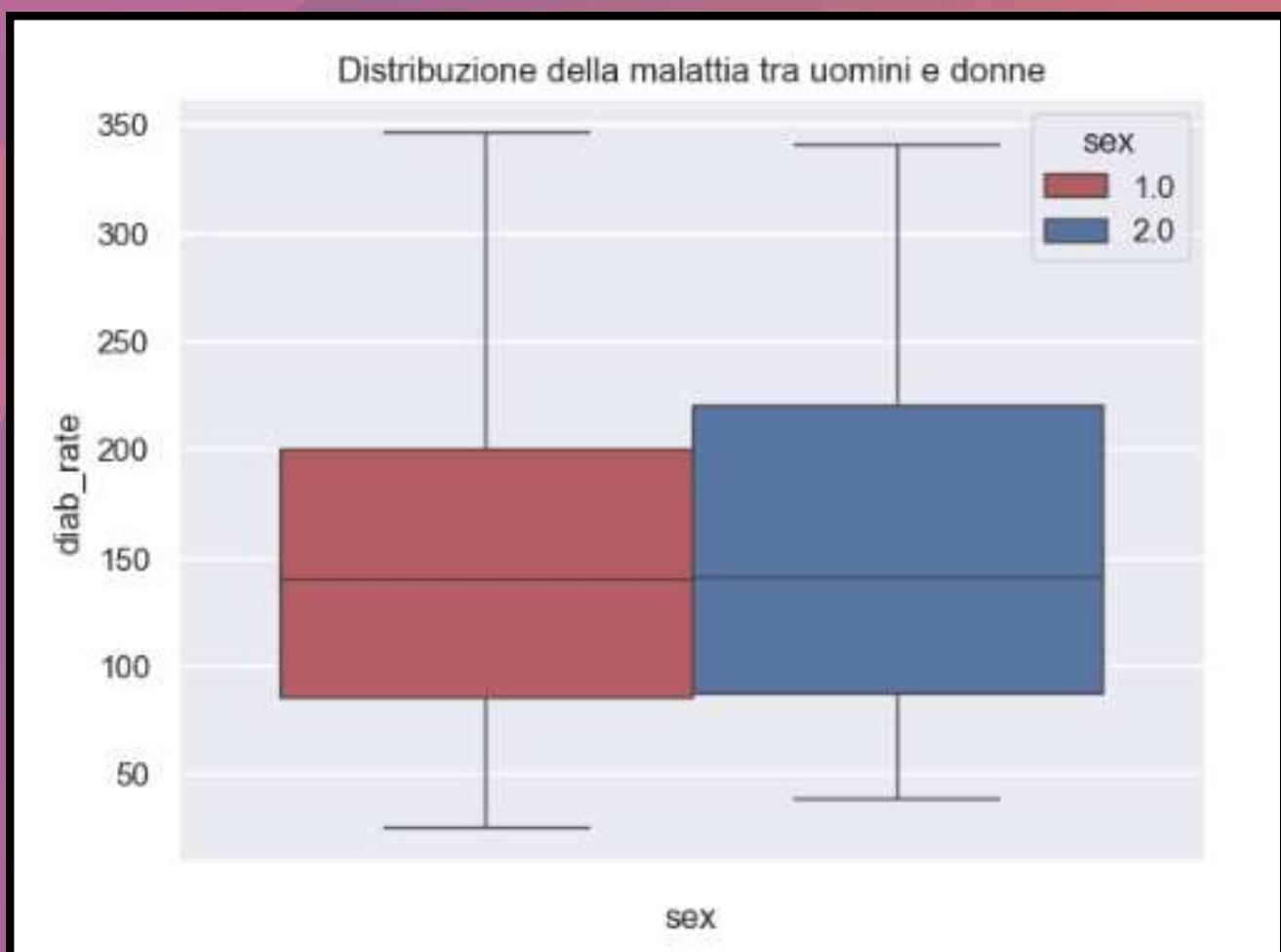
Inoltre si nota che la variabile “cholesterol” presenta un distribuzione simile ad una variabile discreta, probabilmente dovuta ad un arrotondamento dei dati; data la già significativa relazione con la variabile target non verrà trasformata ulteriormente.

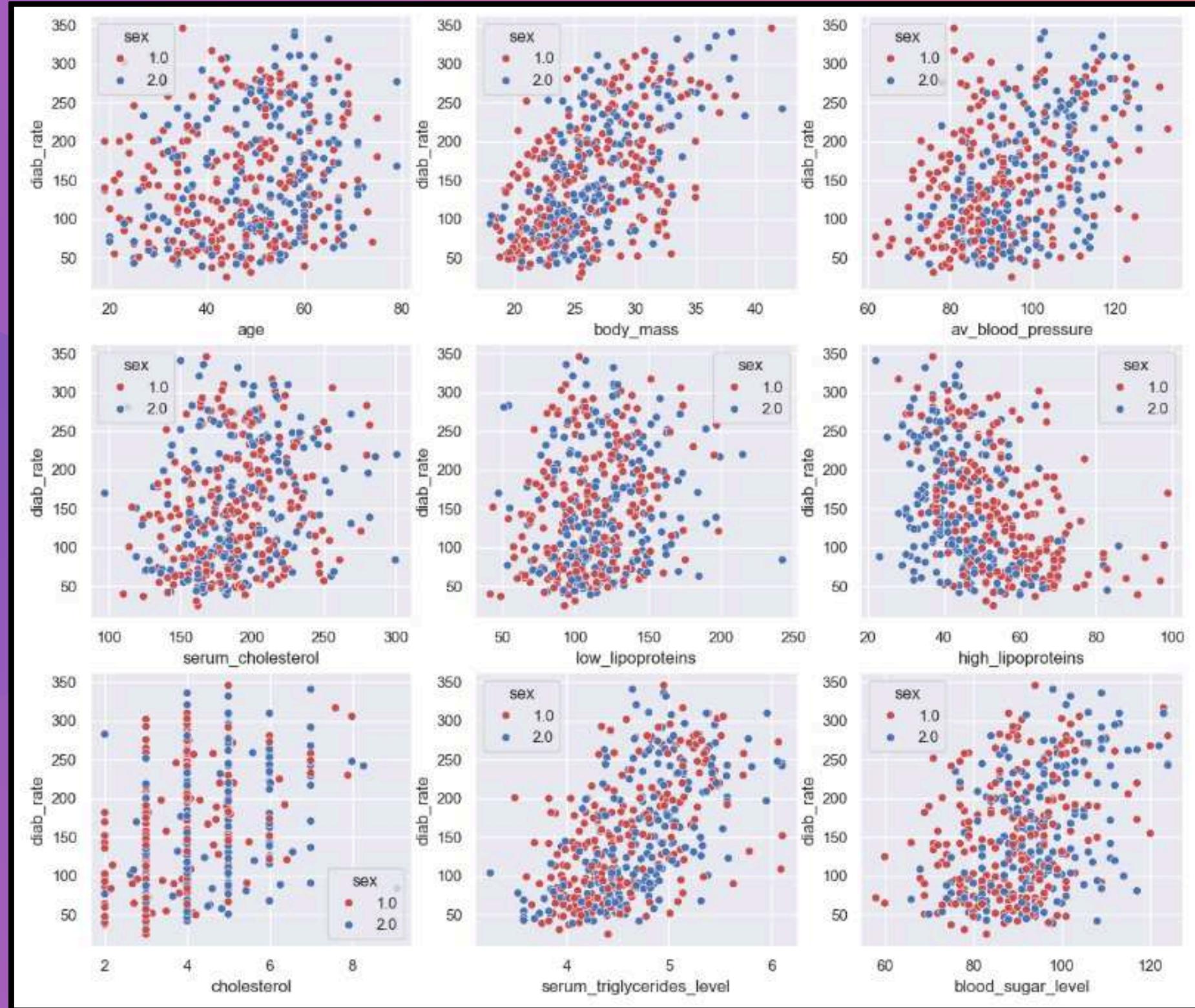


**La feature numerica discreta "sex" mostra una presenza di campioni ed una frequenza bilanciati tra uomini e donne.**

sex	count
1.0	235
2.0	207

Inoltre si nota che nel rapporto con le feature "high\_lipoproteins", "cholesterol" e "av\_blood\_pressure", la variabile "sex" riscontra una discreta influenza nella relazione tra le features con la variabile target, dimostrando che il loro rapporto può variare a seconda del genere ed essere rilevante al livello predittivo.

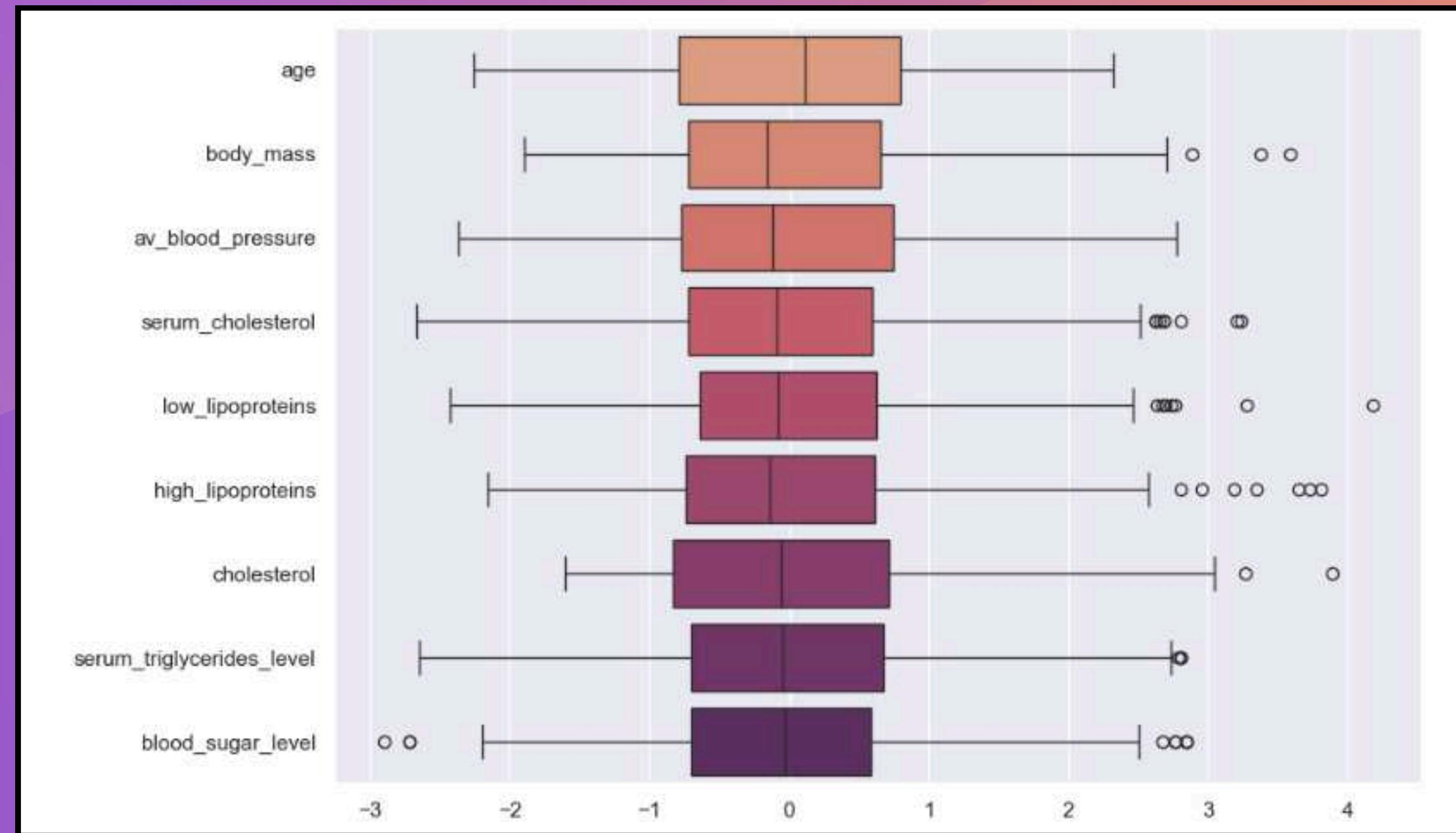




**Gli outliers presenti nel dataset non influiscono sulle predizioni, in quanto la loro presenza è molto ridotta; di conseguenza le metriche utilizzate per valutare i modelli saranno:**

- **l'indice di determinazione R<sup>2</sup> :**  
per ottimizzare la capacità predittiva dei modelli, verificandone la precisione delle predizioni
- **Mean Absolute Error :**  
per calcolare l'errore medio delle predizioni con meno attenzione agli outliers

features	outliers	percentage
age	0	0%
av_blood_pressure	0	0%
blood_sugar_level	9	2,04%
body_mass	3	0,68%
cholesterol	2	0.45%
high_lipoproteins	7	1,58%
low_lipoproteins	7	1,58%
serum_cholesterol	8	1,81%
serum_triglycerides_level	4	0,90%



**Data la natura asimmetrica della distribuzione di frequenza dei valori del target e la natura simmetrica simile alla curva gaussiana nella distribuzione dei valori della maggior parte delle features, i modelli scelti nella fase di Model Selection saranno parametrici (Linear Regression, Ridge) e non (Gradient Boosting Regressor), seguendo un approccio empirico per comprendere quale delle due tipologie gestisce in maniera ottimale il dataset o se l'unione tra di esse, attraverso un modello ensemble, possa riscontrare risultati migliori.**

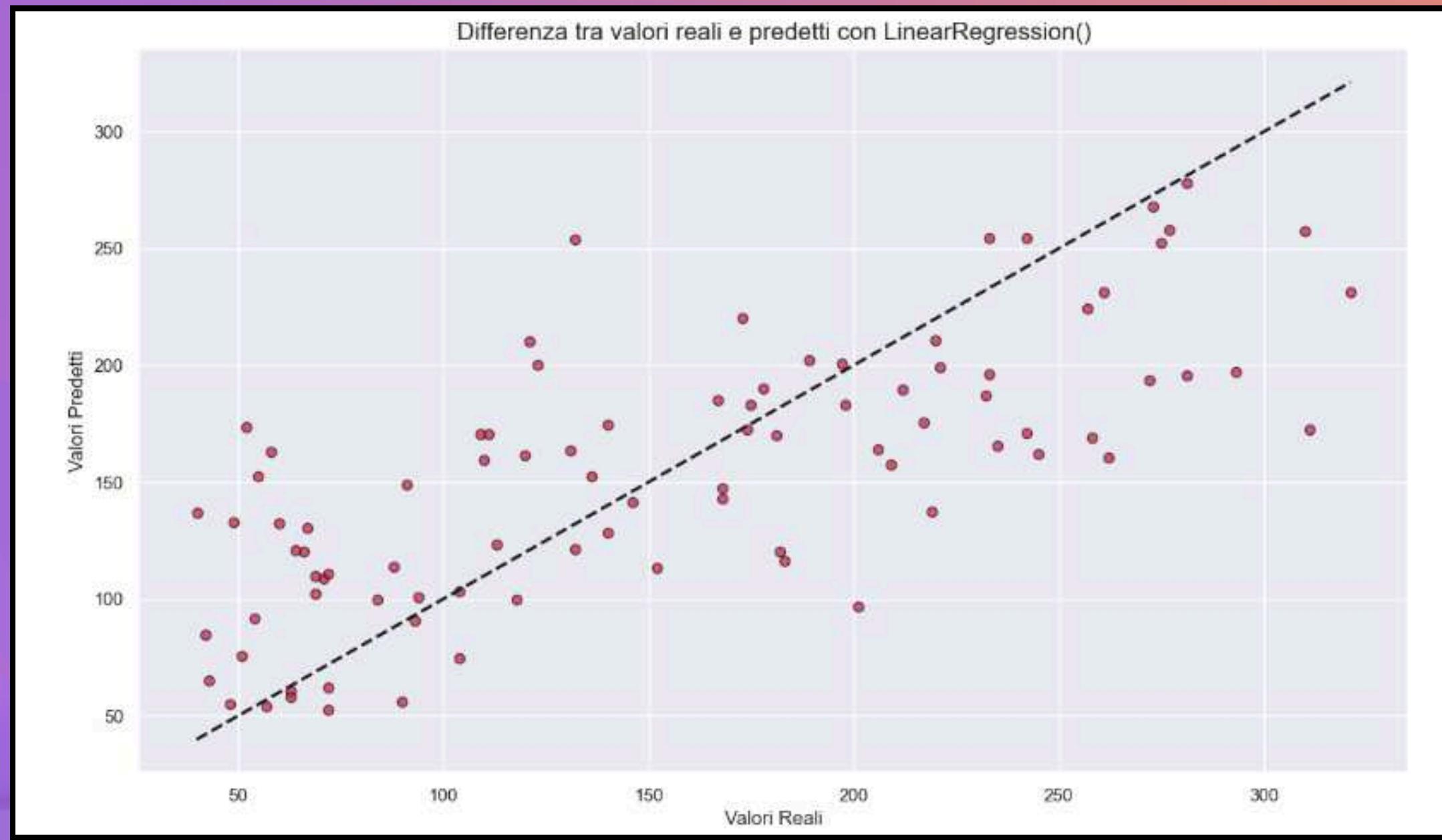
**Inoltre verrà considerato particolarmente il modello SVR, data la sua capacità di rientrare attraverso il "kernel trick" sia nei modelli catalogati come parametrici che non.**

# Pipeline & Baseline Model

Verrà creata una pipeline iniziale per stabilire un punto di partenza nella selezione dei modelli e la gestione del dataset rispetto a:

- la standardizzazione delle features attraverso la funzione **StandardScaler()**, in quanto esse presentano unità di misura differenti tra di loro
- l'impostazione di un modello baseline (**Linear Regression**) per verificare la progressione delle capacità predittive con altri modelli, basandosi su dei punteggi introduttivi dell'indice R2 e del Mean Absolute Error :

MODEL	R2 score	Over-Fitting	Mean Absolute Error
Linear Regression	0.544	0.509	42.548

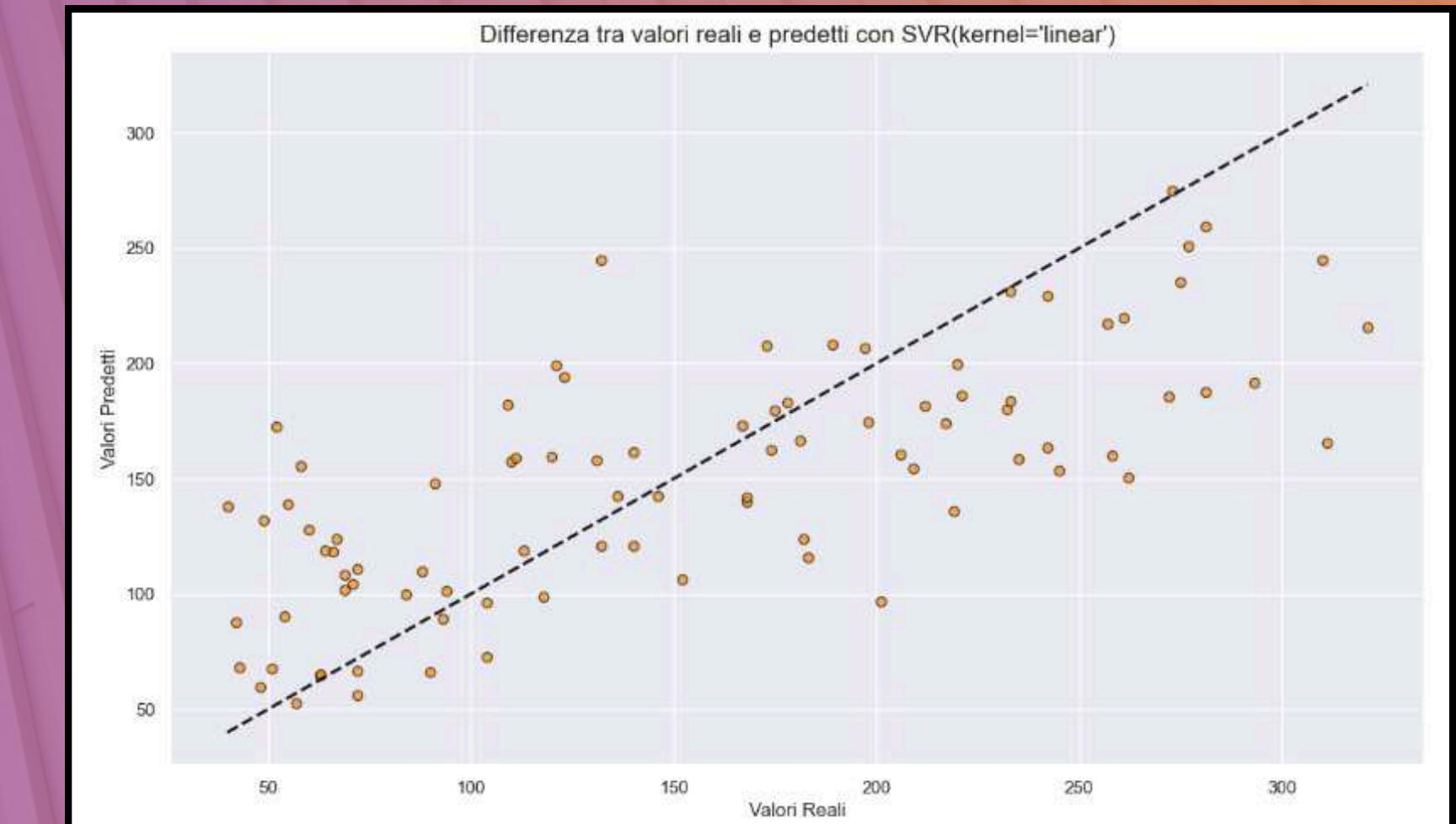
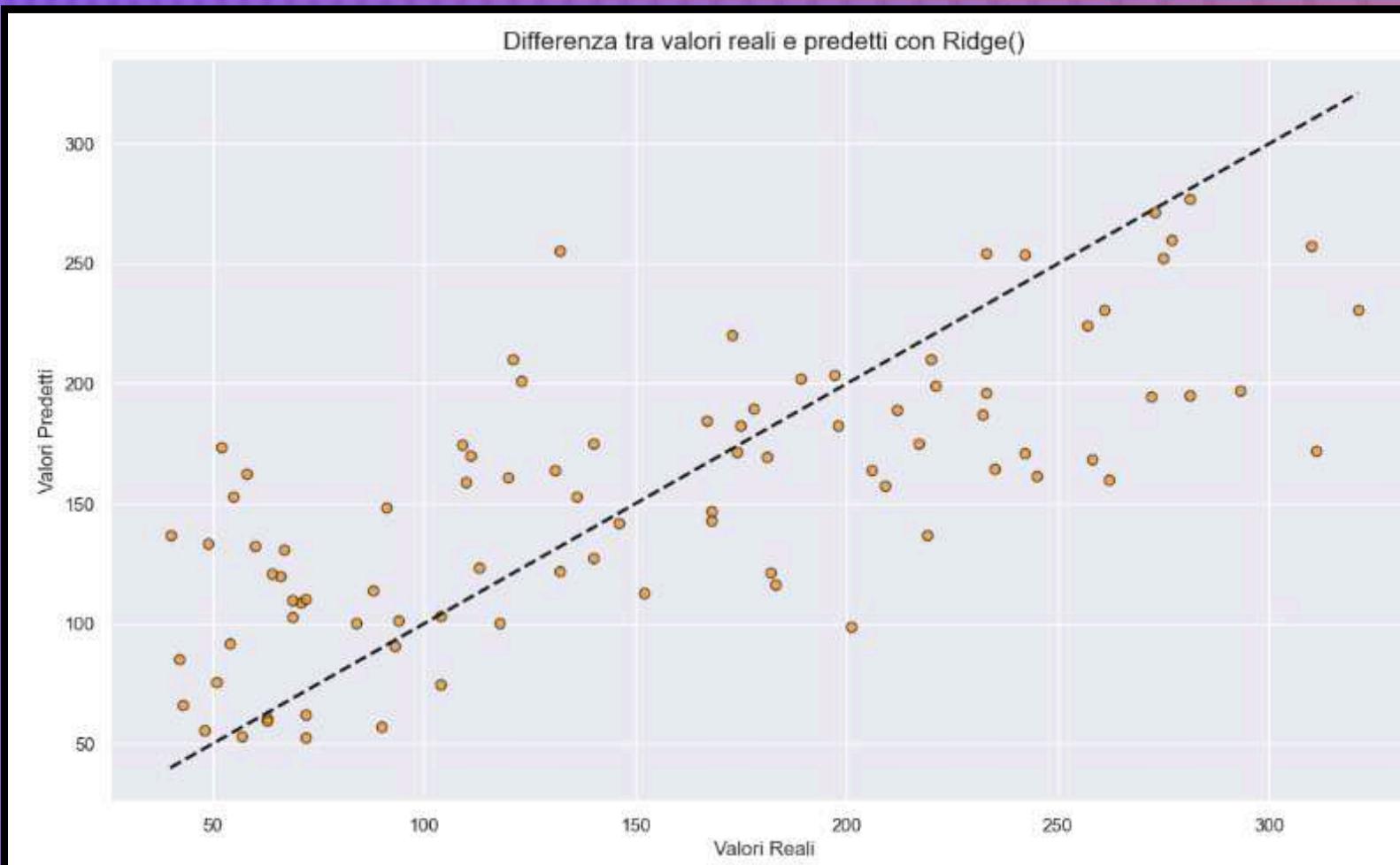


# Empirical Parametric Model Selection

I modelli parametrici scelti in base all'analisi esplorativa sono:

- la Ridge Linear Regression, per ridurre l'impatto del basso numero di Outliers presenti e garantire maggior robustezza nella regolarizzazione rispetto al modello Lasso Linear Regression
- il Support Vector Regressor di tipologia parametrica con kernel lineare.

MODEL	R2 Score	Over-Fitting	Mean Absolute Error
Ridge Linear Regression	0.541	0.509	42.656
SVR con kernel lineare	0.524	0.499	43.673

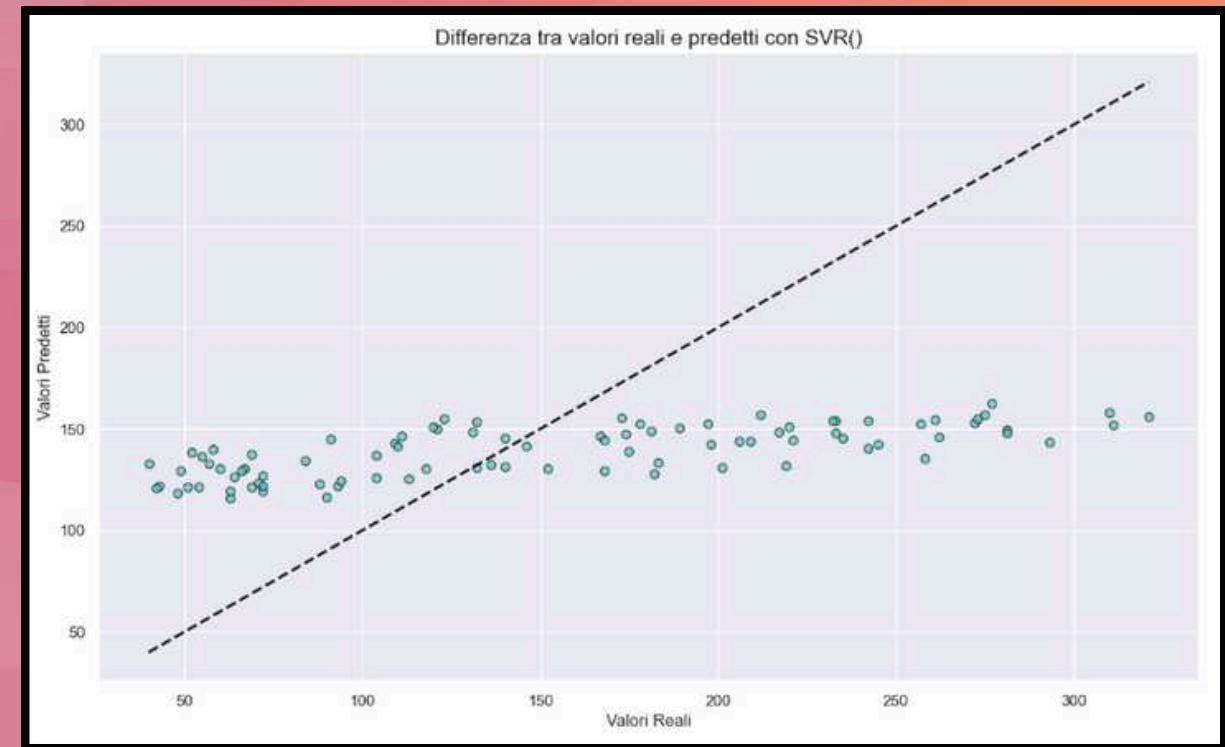
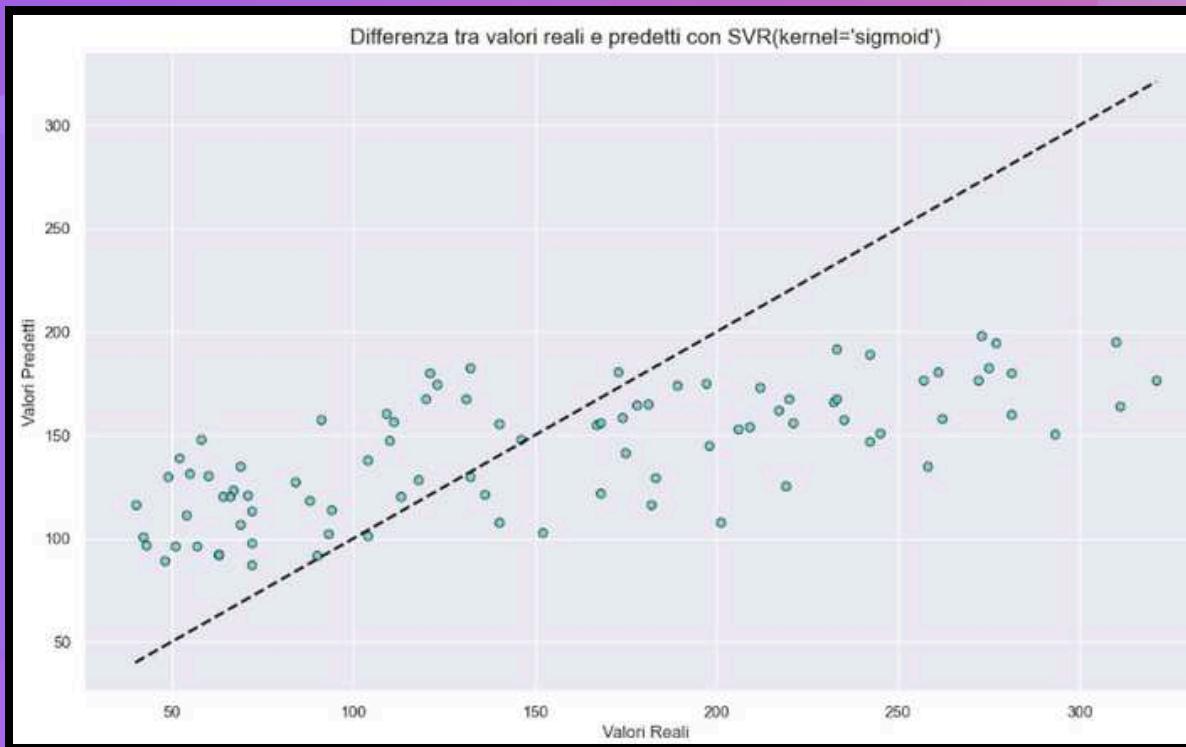
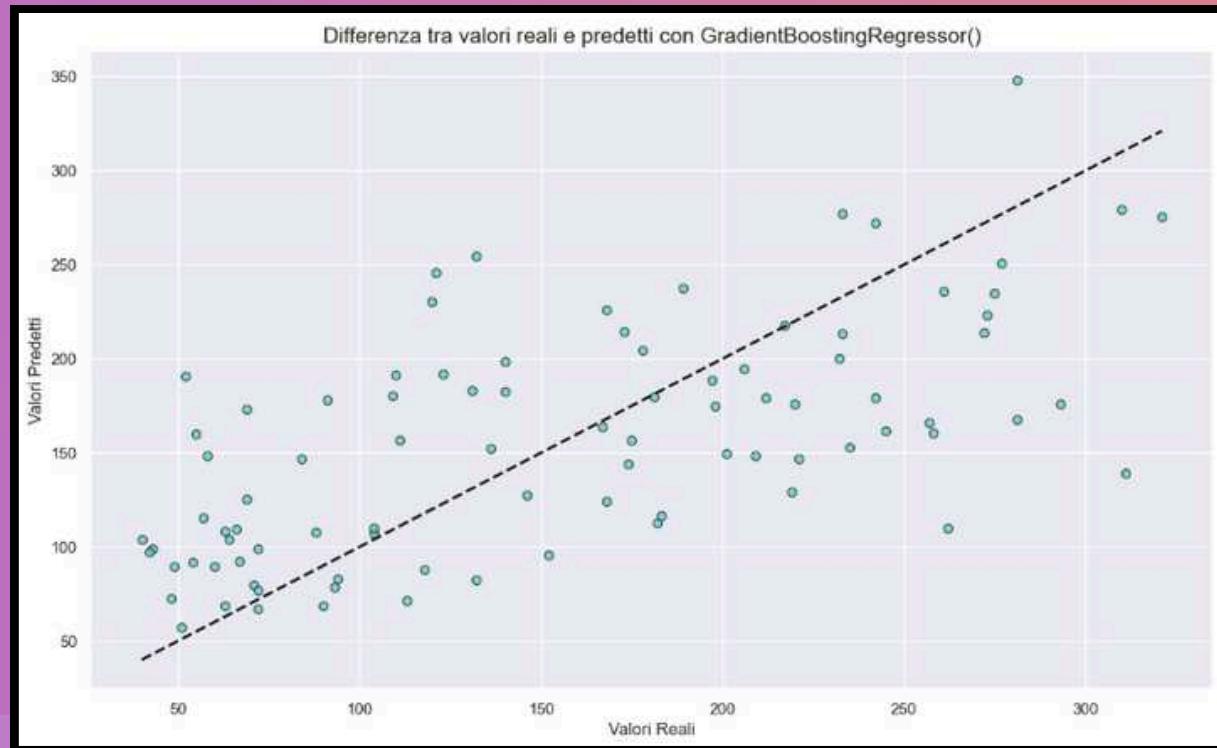


# Empirical Non Parametric Model Selection

I modelli non parametrici scelti sono:

- il Gradient Boosting Regressor per la sua minor sensibilità agli outlier rispetto a modelli come la regressione lineare e per cogliere relazioni non lineari, come quelle dimostrate dalla variabile 'sex' e per la sua migliore adattabilità a dataset di dimensione ridotta.
- il Support Vector Regressor con kernel impostato sulla funzione in base radiale e sigmoide

MODEL	R2 Score	Over-Fitting	Mean Absolute Error
Gradient Boosting Regressor	0.391	0.836	50.883
SVR con kernel sigmoide	0.365	0.356	54.192
SVR con kerne in base radiale	0.160	0.166	62.784



**I primi test empirici hanno dimostrato una maggiore adattabilità dei modelli di tipo parametrico rispetto al dataset in esame.**

**Per verificare ulteriormente questa rilevazione, si procederà con un approccio che integra features selection e hyperparameters tuning parallelamente per ottimizzare le capacità predittive e migliorare le performance rispetto al modello baseline.**

# Selection & Tuning - Parametric Models

**Data la differente importanza delle features e le differenti interazioni presenti tra di esse e la variabile target, si procederà alla features selection con la Recursive Features Elimination (RFE), in quanto questa modalità rimuove iterativamente le features meno importanti in base al modello utilizzato, selezionando in modo più efficiente il set di variabili migliori in base alla relazione con il modello stesso.**

**Rispetto all'hyperparameters tuning, gli iperparametri dei due modelli di tipo parametrico sono continui e di conseguenza la loro validazione sarà svolta attraverso la RandomizedSearchCV per testare più valori possibili.**

**Si osserva un leggero miglioramento rispetto ai modelli iniziali, sufficiente solo nel caso del modello Ridge Linear Regression a superare il punteggio *R2* ottenuto dal modello baseline:**

MODEL	R2 Score	Over-Fitting	Mean Absolute Error
Ridge Linear Regression	0.550	0.504	42.643
SVR con kernel lineare	0.532	0.501	43.243

# Selection & Tuning - Non Parametric Models

**Data la maggiore complessità dei modelli non parametrici, si procederà alla features selection con la Permutation Importance, in quanto essa non richiede l'accesso diretto ai parametri come la RFE, valutando l'importanza delle variabili in base alle performance predittive del modello utilizzato e rimanendo quindi generalmente più robusta con i modelli con alta complessità computazionale.**

**Il metodo utilizzato per l'hyperparameters tuning sarà anche questa volta la RandomizedSearchCV, data la maggiore quantità di iperparametri e di valori da testare per gli stessi nei modelli non parametrici selezionati.**

**Si osserva un miglioramento significativo rispetto ai modelli iniziali, non ancora sufficiente a sostenere le performance dimostrate dal modello baseline:**

MODEL	R2 Score	Over-Fitting	Mean Absolute Error
Gradient Boosting Regressor	0.507	0.605	44.790
SVR con kernel sigmoide	0.525	0.456	43.527
SVR con kernel in base radiale	0.538	0.491	43.777

Le performance riscontrate durante questo secondo test empirico mostrano un discreto miglioramento dei risultati ottenuti con il punteggio dell'indice *R<sup>2</sup>* , attraverso l'utilizzo della Ridge Linear Regression, ma anche un miglioramento delle capacità predittive dei modelli non parametrici, in particolare del modello Support Vector Regressor con kernel in base radiale, che mantiene un equilibrio migliore tra i risultati dell'indice di determinazione e l'Over-Fitting ottenuti su test e training set, dimostrando una capacità predittiva più attendibile sui nuovi dati.

Di conseguenza l'ultimo test empirico cercherà di migliorare le performance del modello con regolarizzazione in base Ridge, creando un modello ensemble che possa assimilare le capacità predittive dimostrate dal modello migliore appartenente alla tipologia non parametrica.

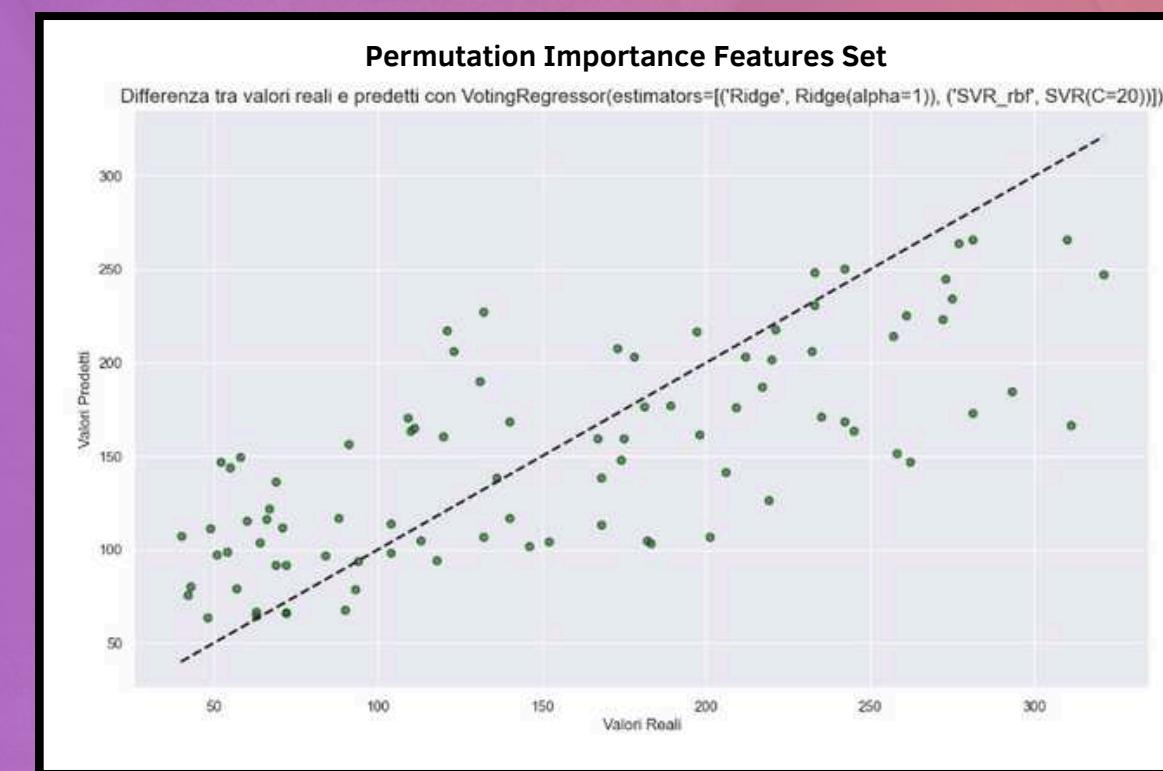
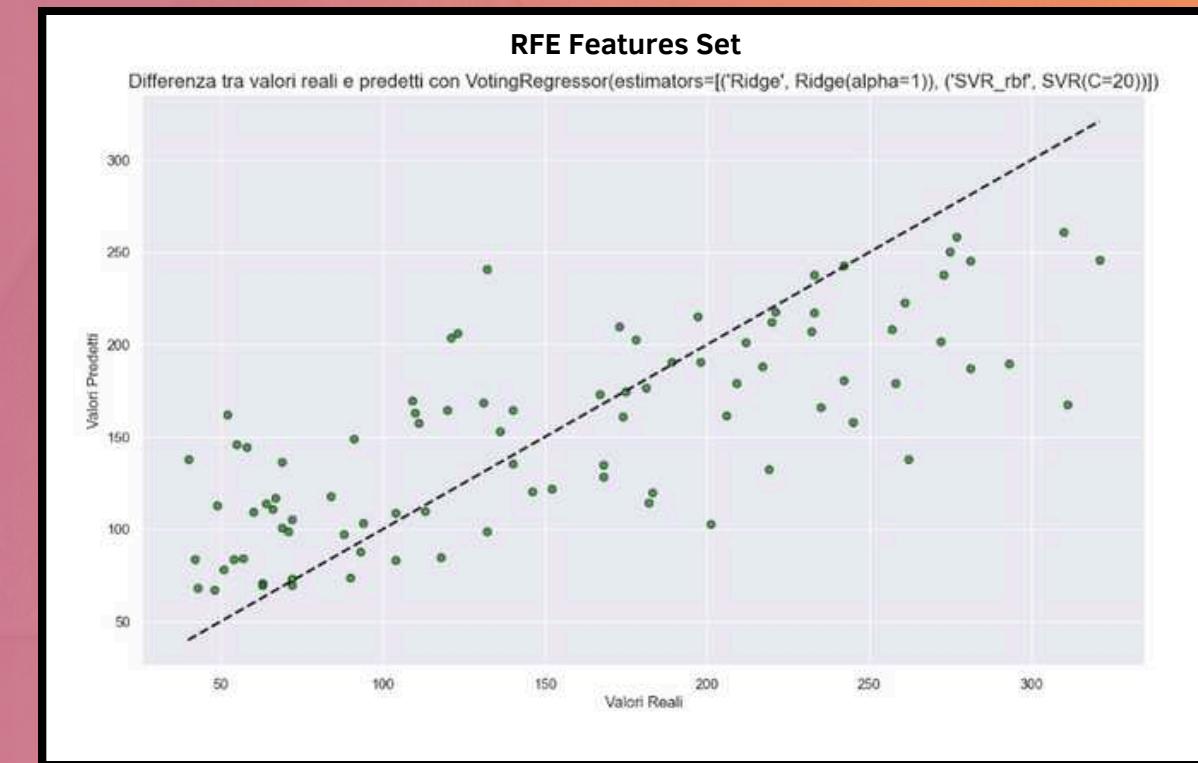
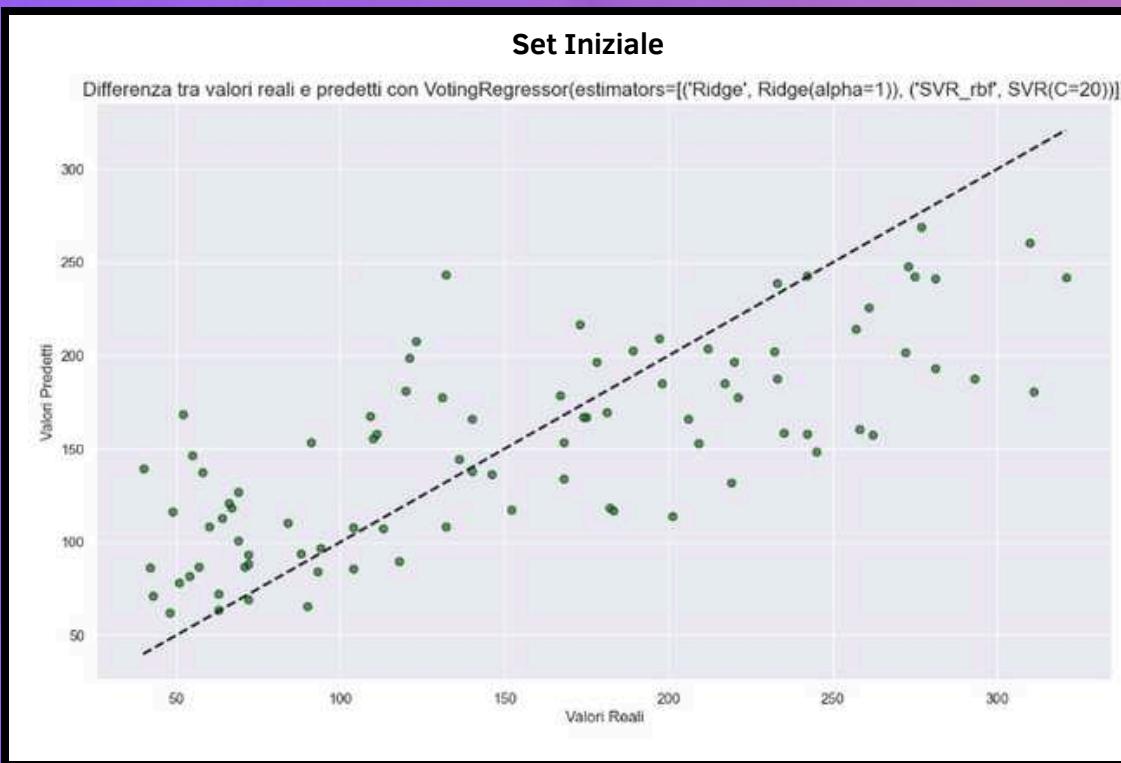
# Ensemble Voting Regressor

Con lo scopo di migliorare le prestazioni complessive dei due modelli individuati precedentemente, verrà creato un modello unico attraverso la tecnica ensemble del Voting Regressor, combinando le predizioni dei due metodi di regressione, parametrico e non, per ottenere con caratteristiche diverse una previsione unica basata sulla media ponderata delle previsioni individuali.

Il modello ensemble verrà poi testato con il training e test set base ed i training e test set di features isolati attraverso la RFE e la Permutation Importance per i modelli Ridge e SVR con kernel in base radiale, che compongono il modello ensemble stesso.

**Il modello Voting Regressor addestrato con i training e test set selezionati dalla RFE migliora significativamente il punteggio dell'indice  $R^2$  e riduce maggiormente l'errore medio assoluto sulle predizioni rispetto al modello baseline e agli altri test empirici svolti con il Voting Regressor:**

MODEL	R2 Score	Over-Fitting	Mean Absolute Error
Voting Regressor con features set iniziale	0.562	0.557	42.126
Voting Regressor con features set RFE	0.571	0.534	41.216
Voting Regressor con features set Permutation Importance	0.549	0.504	43.161



ML MODEL	Features Set	R2 Score	Over-Fitting	MAE
Linear Regression	Set Base	0.544	0.509	42.548
Ridge Linear Regression	RFE	0.550	0.504	42.643
SVR con kernel lineare	RFE	0.532	0.501	43.243
Gradient Boosting Regressor	Permutation Importance	0.507	0.605	44.790
SVR con kernel sigmoide	Permutation Importance	0.525	0.456	43.527
SVR con kernel in base radiale	Permutation Importance	0.538	0.491	43.777
Voting Regressor con Ridge - SVR	Ridge con RFE	0.571	0.534	41.216

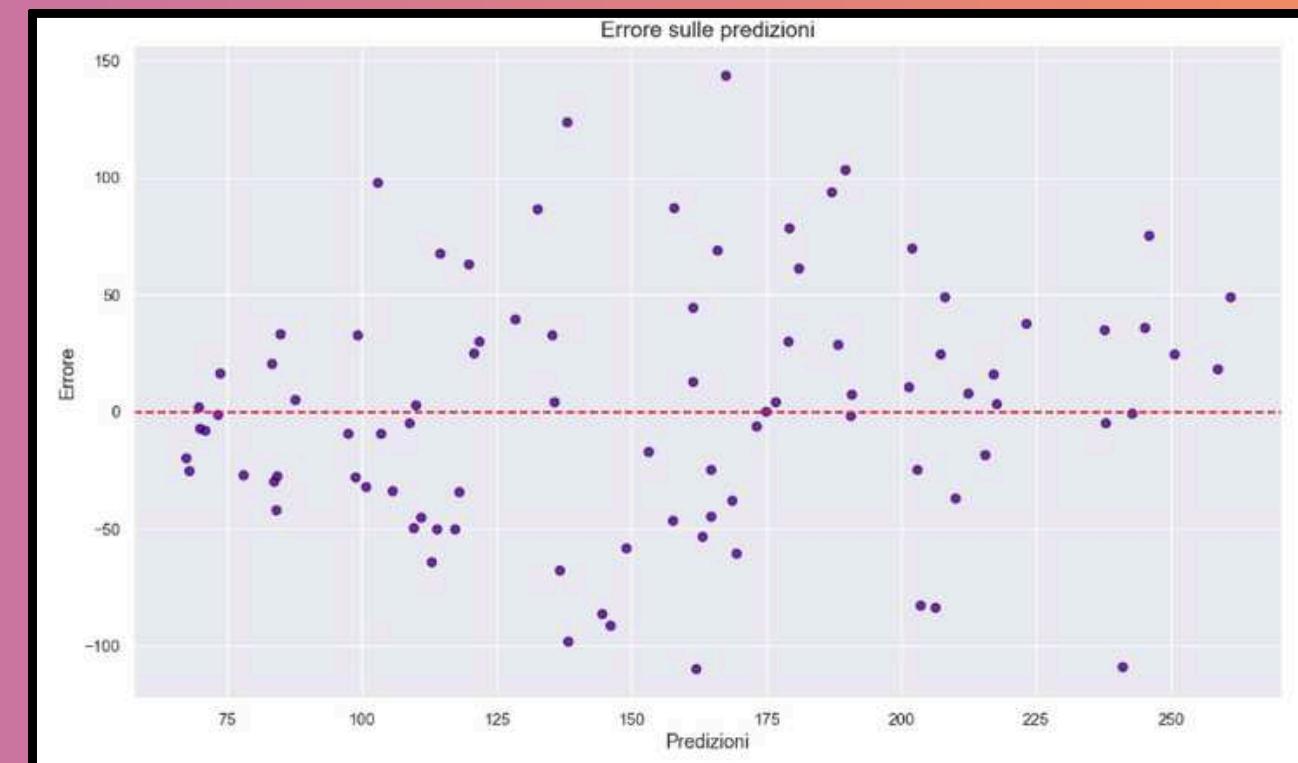
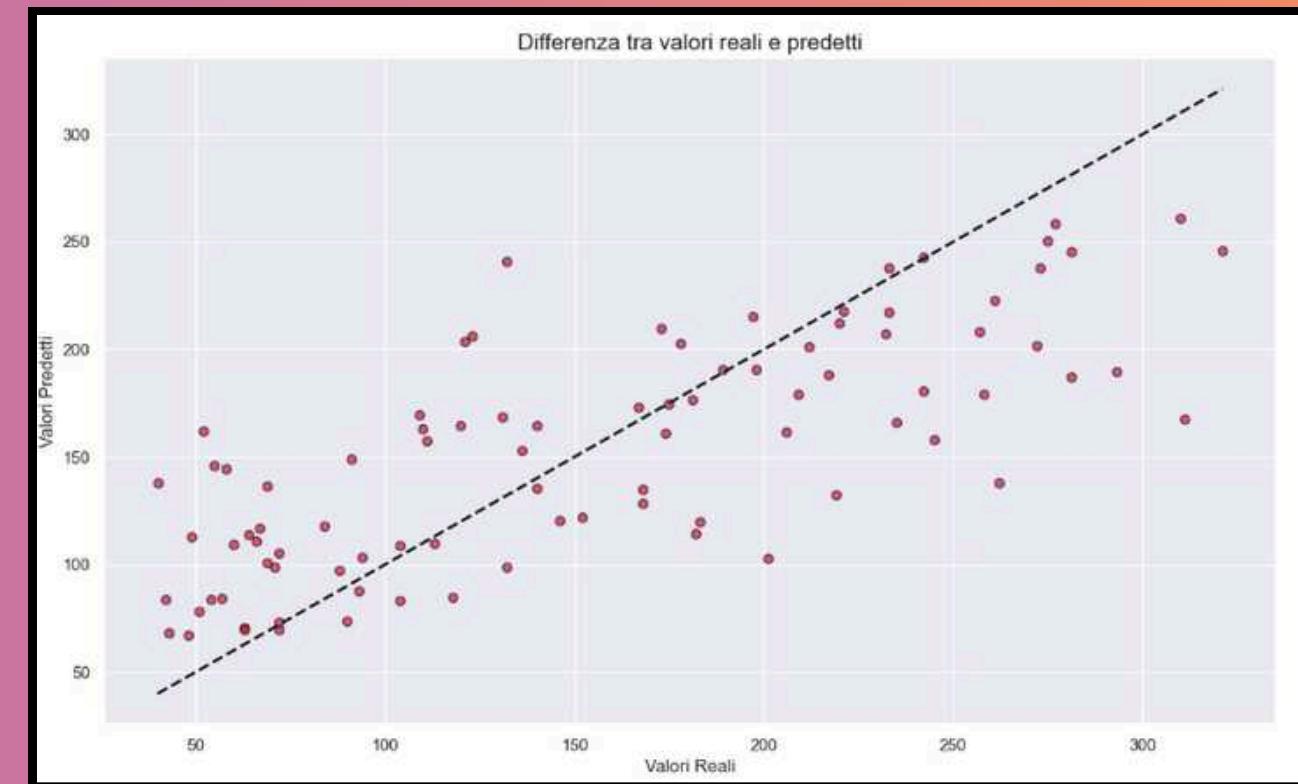
# Conclusioni

In definitiva, il modello migliore riscontrato per la previsione dell'indice di progressione del diabete è il Voting Regressor allenato con le features selezionate attraverso il metodo RFE applicato sul modello Ridge Linear Regression, migliorando le proprie capacità rispetto al modello baseline con:

- l'aumento del 3% del punteggio dell'indice  $R^2$  (da 0.544 a 0.571)
- la conservazione dell'equilibrio tra i risultati dell'indice  $R^2$  e l'Over-Fitting (da 0.544-0.509 a 0.571-0.534 )
- la riduzione del 3% dell'errore assoluto sulle predizioni MAE (da 42.548 a 41.216)

## Features Set:

- **sex**
- **body\_mass**
- **av\_blood\_pressure**
- **serum\_cholesterol**
- **low\_lipoproteins**
- **serum\_triglycerides\_level**



Grazie per la visione