RELAZIONE PROGETTO STATISTICA NUMERICA

a.a 2023/2024

Il dataset scelto è stato raccolto dal servizio di assicurazione sanitaria nazionale in Corea.
Lo scopo di questo dataset è:

1. Analizzare i parametri del corpo
2. Classificazione di fumatori e bevitori

L'obiettivo è prevedere se una persona assume alcool, dati in input degli attributi che rappresentano le caratteristiche e i parametri corporei del soggetto.

Link: https://www.kaggle.com/datasets/sooyoungher/smoking-drinking-dataset

---

Pre-processing

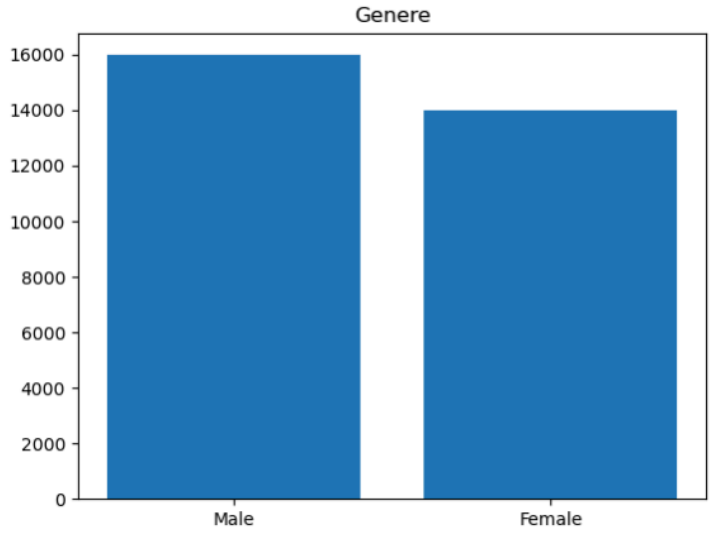
Date le grandi dimensioni del dataset, ho eliminato delle righe, in modo da avere una maggiore velocità di esecuzione.
Il dataset partiva composto da 991'346 righe (persone analizzate); dopo la rimozione, sono sceso a 30'000.
Ho rimosso 4 colonne che non ho valutato utili per la predizione della variabile target. Rappresentano rispettivamente la qualità visiva (occhio destro e sinistro) e uditiva (orecchio destro e sinistro).
Ho verificato che non ci fossero valori NaN nelle righe, attraverso la funzione `dropna(axis=0)`.
Infine, ho rimpiazzato i valori della variabile target in 0 e 1, al posto di Y e N. In questo modo ho una variabile numerica su cui posso effettuare operazioni.

Exploratory Data Analysis (EDA)

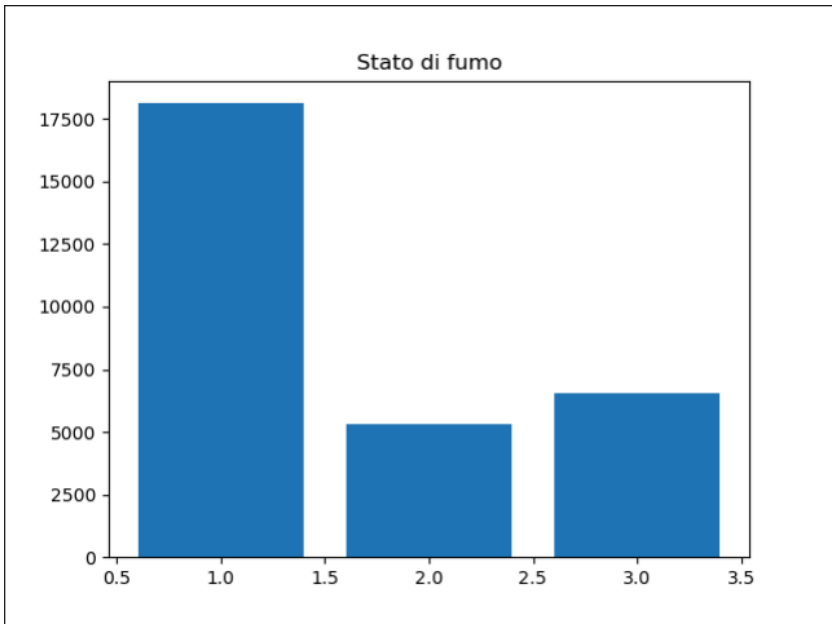
Inizialmente ho effettuato un'ispezione sui dati più generici, quindi età, altezza e peso.
Questi sono i risultati:

- **Età** - Media: 47.57 Mediana: 45.0
- **Peso** - Media: 63.29 Mediana: 60.0
- **Altezza** - Media: 162.3 Mediana: 160.0

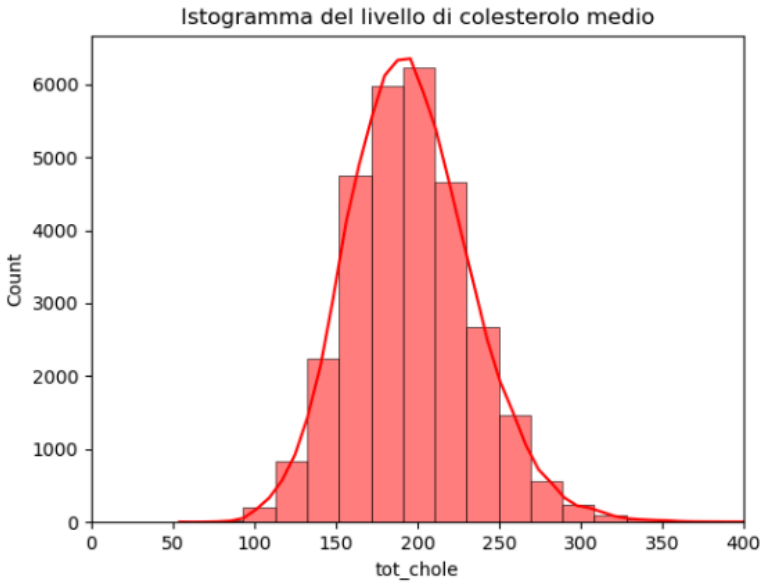
Per quanto riguarda il genere, il numero di maschi e femmine è pressoché uguale.



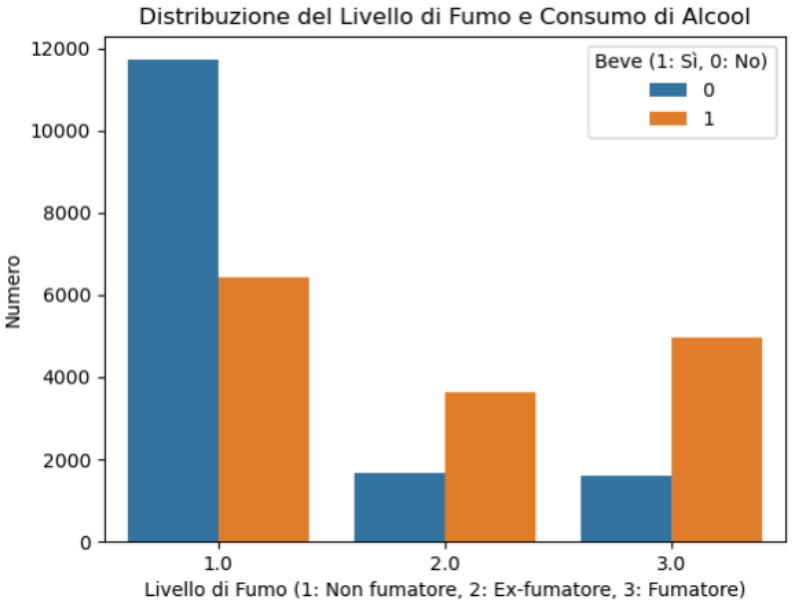
Stato di fumo (1 - non fumatore , 2 - ex-fumatore , 3 - fumatore)



Si può notare una distribuzione normale per quanto riguarda i livelli di colesterolo.
Livello di colesterolo medio: 195.66771

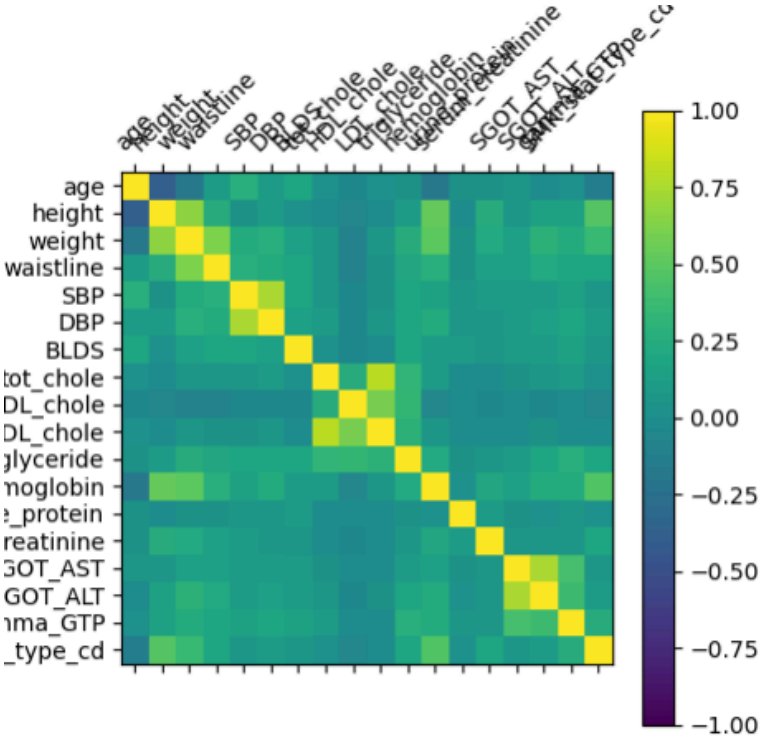


Un aspetto di cui ho tenuto conto durante questa fase è la relazione tra fumatori e bevitori.



Dal grafico, si può notare come se una persona non fuma è meno probabile che beva, mentre negli ex-fumatori e fumatori viceversa.

Come ultimo punto in questa fase, ho generato una matrice di correlazione.



Dalla matrice di correlazione, ho individuato i parametri maggiormente correlati che ho utilizzato dopo nella fase di Regressione.

Splitting

In questa fase, ho diviso il dataset in train set e test set con una dimensione del 78% delle righe per il train set.
Infine, ho diviso il train set in due per ottenere il validation set, in modo da poter sperimentare tutti i possibili iperparametri.
Ho cercato una percentuale per avere dimensioni simili nel test set e validation set.

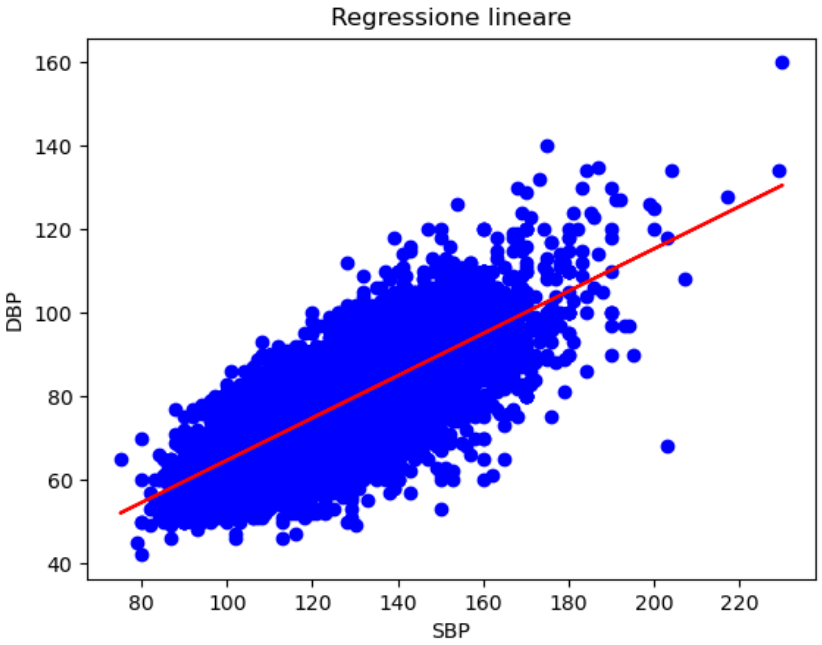
Regressione

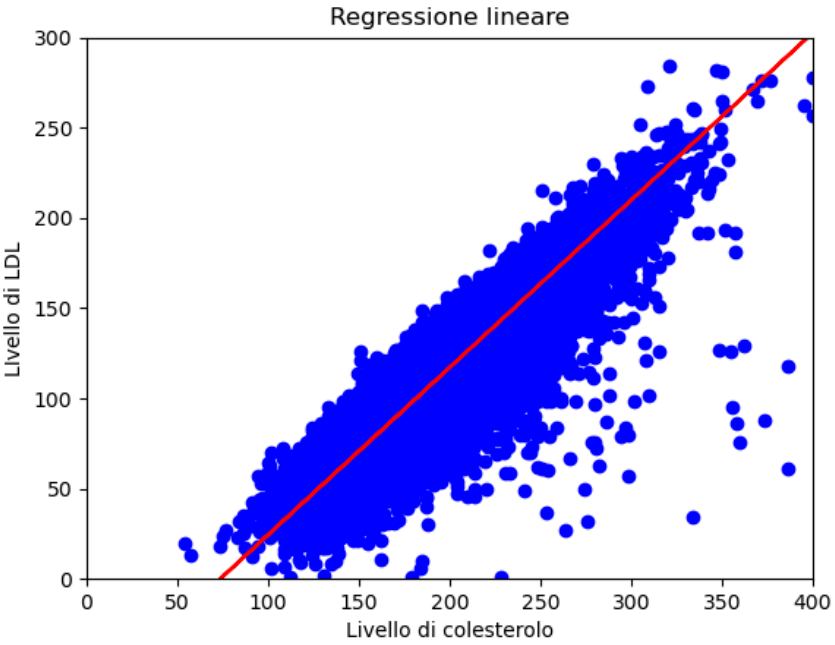
Dalla matrice di correlazione,ho individuato le 2 coppie di valori maggiormente correlate:

- SBP (pressione sanguigna sistolica) e DBP (pressione sanguigna diastolica)
- Livello di colesterolo e livello di lipoproteine a bassa densità (LDL)

La prima coppia è correlata al 74%
La seconda coppia è correlata al 79.9%

La pressione sanguigna sistolica rappresenta la pressione del sangue nelle arterie quando il cuore si contrae e pompa il sangue nel sistema circolatorio.
La pressione sanguigna diastolica rappresenta la pressione nelle arterie quando il cuore si rilassa tra una battito e l'altra.
L’indice LDL, invece, indica la capacità del corpo di trasportare il colesterolo nel sangue.





Risultati tra SBP e DBP :

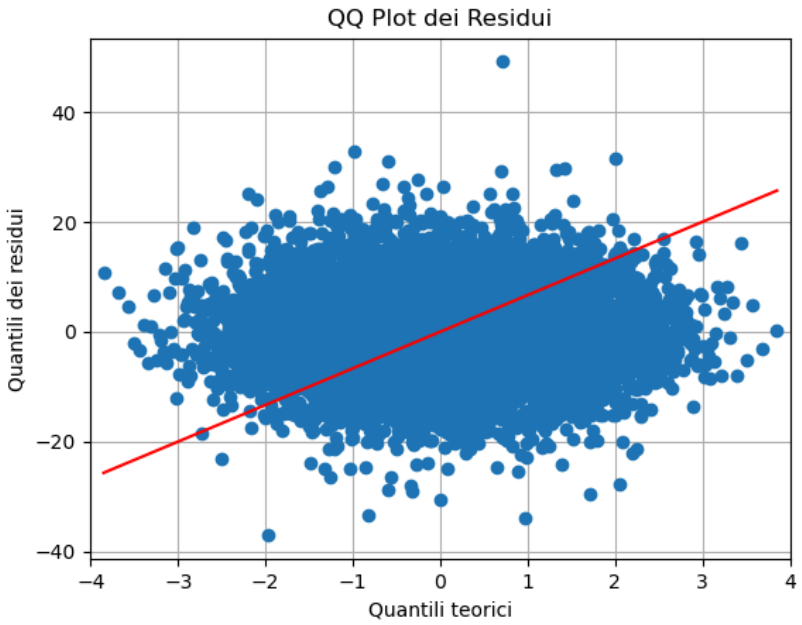
r^2 : 0.205
MSE: 44.5193

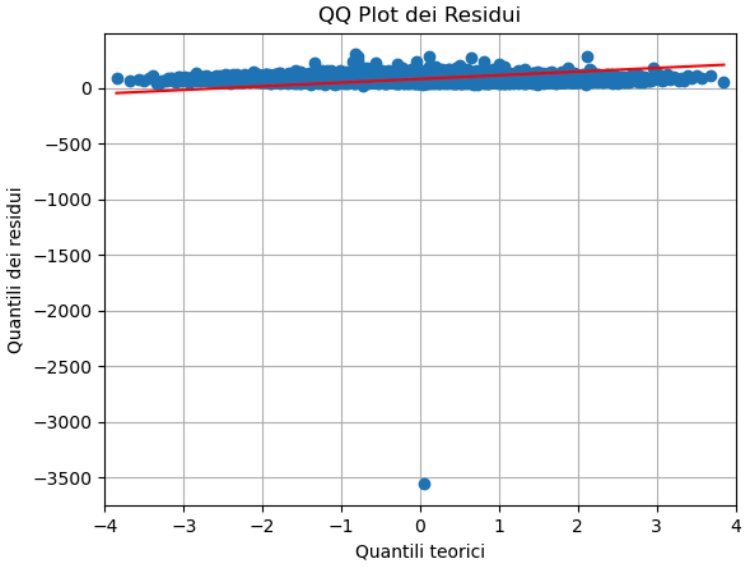
Risultati tra livello di colesterolo e LDL :

r^2 : 0.337
MSE: 1049.718

Questi risultati fanno capire come, nonostante gli attributi siano correlati tra di loro, i modelli di regressione lineare non sono adeguati per le relazioni tra le due coppie di valori.
L'indice r^2 molto basso indica che il modello utilizzato non è adatto per i dati in questione.
Mentre, valori MSE (Mean Squared Error) elevati indicano che le previsioni del modello sono lontane dai valori reali.

Per quanto riguarda l'analisi di normalità dei residui, ho tracciato dei qq-plot per capire se avessero una distribuzione normale.
Nel primo caso, non c'è una distribuzione normale.
Mentre nel secondo caso, potrebbe esserci ma è impedita da un outlier. Per questo ho eliminato l'elemento e ho stampato un nuovo grafico, ma nonostante ciò la distribuzione non è normale.





Addestramento del modello

Provando ad addestrare con una Logistic Regression, il modello non converge.

Di conseguenza, ho provato con due kernel in particolare della classe SVC: linear e rbf.

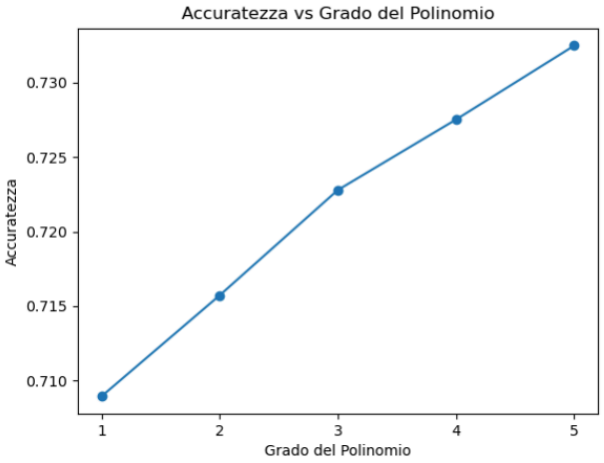
```
[0 1 0 ... 1 1 0]
ME: 4964
MR: 0.28284900284900283
Acc: 0.7171509971509972
[0 1 0 ... 1 1 0]
ME: 4918
MR: 0.2802279202279202
Acc: 0.7197720797720798
[0 1 0 ... 1 1 0]
```

Kernel linear
Accuratezza: 71.71%.

Kernel rbf (gamma=1)
Accuratezza: 71.97%

Hyperparameter Tuning

In questa fase, ho utilizzato un altro kernel della classe SVC: il poly.
Questo tipo di kernel accetta un altro parametro che è il grado, che vado ad impostare iterativamente in un ciclo for.



Si può notare come all'aumentare del grado, aumenti anche l'accuratezza. Ma bisogna stare attenti, perché un'elevata accuratezza potrebbe portare ad una situazione di overfitting.

Valutazione delle performance

Accuratezza sul test set: 0.72382 | 72.38%

Matrice di confusione	Vero Positivo	Vero Negativo
Esito Positivo	2567	752
Esito Negativo	1071	2211

Positive predictive value = 2567 / (2567+752) ≈ 0.774 | 77.4%
Bevitori che effettivamente bevono

Negative predictive value = 2211 / (2211 +1071) ≈ 0.674 | 67.4%
Non bevitori che effettivamente non bevono

Sensibilità = 2567 / (2567+1071) ≈ 0.706 | 70.6%
Esiti positivi effettivi che sono stati correttamente identificati dal modello

Specificità = 2211 / (752+2211) ≈ 0.746 | 74.6%
Esiti negativi effettivi che sono stati correttamente identificati dal modello

---

Studio statistico sui risultati della valutazione

Accuratezza media: 0.72295 | 72.3%

Deviazione Standard: 0.00546
Le accuratezze ottenute nelle diverse valutazioni sono molto vicine alla media, con una variazione media di circa 0.546%

Intervallo di confidenza 95% : (0.71956, 0.72633)
Si è al 95% sicuri che la vera accuratezza media del modello si trovi in questo intervallo.
Non avendo una distribuzione normale, ho usato la distribuzione t di Student.