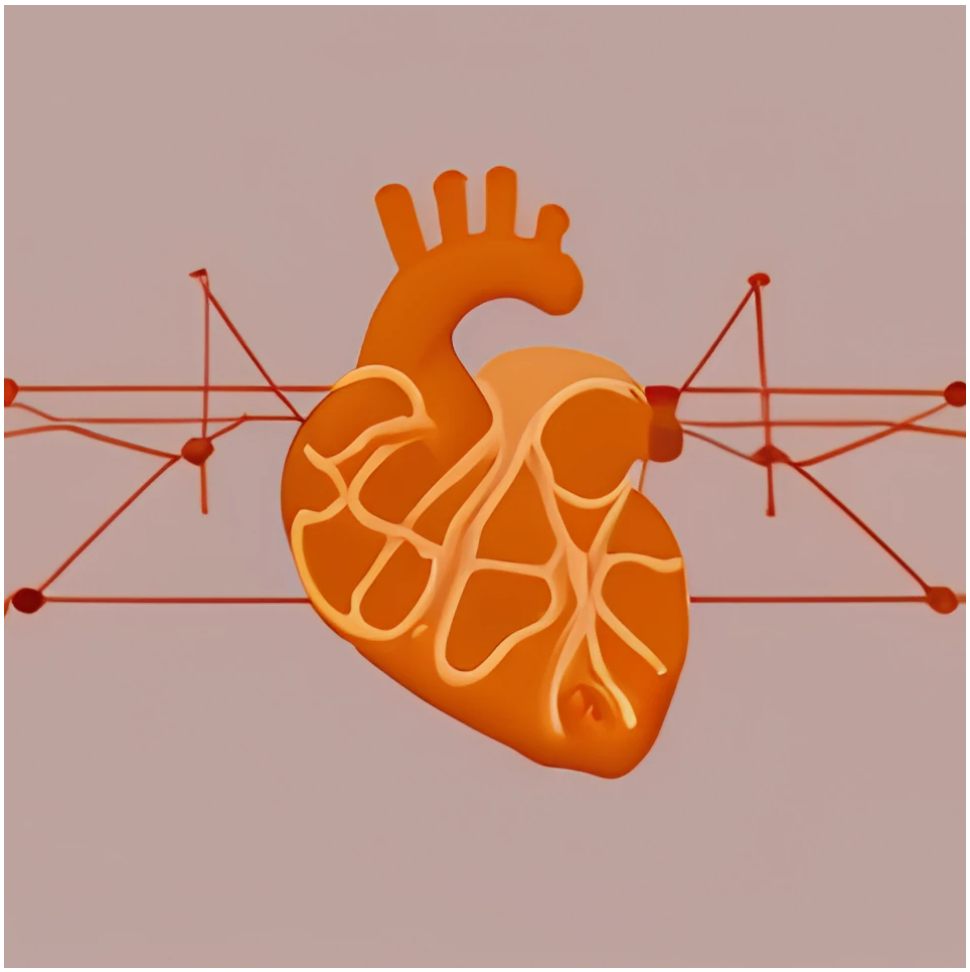
ELABORATO IN INTELLIGENZA ARTIFICIALE

Predizione sull'incidenza di Patologie Cardiache utilizzando una Rete Neurale : un'analisi esplorativa su dati clinici.



Andrea Saggio

Docente : Emanuel Weitschek

Anno Accademico : 2022/2023

Matricola Studente : 757HHHINGINFOR

UNIVERSITÀ TELEMATICA
INTERNAZIONALE UNINETTUNO

Indice

1. INTRODUZIONE

L'Intelligenza Artificiale è un campo dell'informatica che si occupa di sviluppare algoritmi e sistemi in grado di esibire comportamenti intelligenti, simili a quelli umani.

Il machine learning è una branca dell'intelligenza artificiale che si concentra sull'apprendimento automatico da dati e sull'adattamento degli algoritmi ai nuovi dati, migliorando la loro capacità di eseguire compiti specifici senza essere esplicitamente programmati per farlo.

Il Machine Learning sta rivoluzionando l'approccio alla diagnosi e alla prevenzione delle malattie cardiache, grazie alla capacità di elaborare grandi quantità di dati e di riconoscere pattern e correlazioni che spesso sfuggono all'occhio umano.

I Modelli di Machine Learning possono essere addestrati per riconoscere i fattori di rischio e prevedere la probabilità di sviluppare una malattia cardiaca utilizzando dati clinici, informazioni su stili di vita, fattori genetici e altri fattori di rischio per creare predizioni personalizzate, aiutando i medici a identificare le persone a rischio e a prescrivere interventi preventivi tempestivi.

Inoltre, i modelli di Machine Learning possono essere utilizzati anche per analizzare grandi dataset di immagini cardiache, aiutando i medici a identificare anomalie e lesioni che possono essere segnali di problemi cardiaci.

Tutto ciò sta rendendo l'approccio alla diagnosi e alla prevenzione delle malattie cardiache sempre più preciso ed efficace.

1.1 CENNI STORICI

Il Machine Learning applicato alla predizione di malattie cardiache è un campo di ricerca relativamente recente, ma ha radici storiche nell'ambito dell'informatica e dell'intelligenza artificiale.

Negli anni '50 e '60, l'informatico americano Arthur Samuel sviluppò il primo algoritmo di machine learning, chiamato "gioco delle dame" (checkers game). Questo algoritmo imparava a giocare a dama attraverso l'analisi di partite giocate in precedenza e migliorava la sua strategia man mano che acquisiva più conoscenze.

Negli anni '80, il campo del Machine Learning iniziò a essere applicato alla Medicina, con lo sviluppo di algoritmi in grado di analizzare immagini mediche e aiutare i medici nella diagnosi. Tuttavia, la mancanza di dati accurati e la complessità delle malattie cardiache hanno rappresentato una sfida per la predizione di queste patologie tramite il Machine Learning.

Negli ultimi anni, grazie alla disponibilità di grandi quantità di dati medici, all'avanzamento delle tecnologie informatiche e alla crescente consapevolezza dell'importanza della prevenzione delle malattie cardiache, il Machine Learning sta diventando sempre più rilevante nella predizione e nella prevenzione delle Patologie Cardiovascolari.

2. CONCETTI

2.1 CONCETTI DI INTELLIGENZA ARTIFICIALE (AI)

L'intelligenza artificiale (AI) è un campo dell'informatica che si occupa di creare sistemi in grado di eseguire compiti che normalmente richiedono l'intelligenza umana, come la comprensione del linguaggio naturale (NLP), la percezione visiva (Computer Vision), il ragionamento e il processo decisionale.

La Computer Vision si concentra sull'elaborazione delle immagini e sulla capacità di analizzare, interpretare e comprendere il contenuto visuale.

La NLP (Natural Language Processing) si concentra sull'elaborazione del linguaggio umano, attraverso la comprensione, la generazione e la traduzione di testo.

L'obiettivo dell'AI è di creare sistemi che possano apprendere in modo autonomo, adattarsi al cambiamento e migliorare continuamente le loro prestazioni.

L'IA può essere suddivisa in diverse categorie, tra cui:

- L'IA debole o "stretta": si riferisce a sistemi in grado di eseguire specifici compiti limitati, come il riconoscimento di immagini o la traduzione di lingue straniere.
- L'IA forte o "generale": si riferisce a sistemi in grado di eseguire compiti che richiedono una vasta gamma di abilità intellettuali, come il ragionamento, la comprensione del linguaggio naturale e la risoluzione di problemi.

L'IA ha molte applicazioni in diversi settori, tra cui la medicina, la finanza, l'automazione industriale e molti altri.

2.2 CONCETTI DI MACHINE LEARNING (ML)

Il Machine Learning è una branca dell'informatica che si occupa di sviluppare algoritmi che possono apprendere dai dati attraverso l'uso di modelli matematici e statistici come la teoria dell'informazione, l'analisi dei dati, la teoria dei grafi, la teoria della probabilità e la teoria della complessità computazionale.

L'obiettivo principale del Machine Learning è quello di costruire un modello in grado di eseguire una specifica attività senza essere esplicitamente programmato per farlo. In altre parole, l'algoritmo di machine learning utilizza i dati di input per costruire un modello matematico che possa effettuare previsioni o classificazioni su nuovi dati di input.

2.3 CONCETTI DI RETE NEURALE (NN)

Le Reti Neurali sono un tipo di algoritmo di Machine Learning che si ispirano al funzionamento del cervello umano. Una rete neurale è composta da un grande numero di unità di elaborazione, chiamate neuroni artificiali, che sono interconnesse tra di loro. Ogni neurone elabora i segnali di input che riceve dalle connessioni in ingresso e genera un segnale di output che viene trasmesso alle connessioni in uscita. La rete neurale elabora i dati attraverso una serie di trasformazioni matematiche, rappresentate dalle connessioni tra i neuroni, e adatta i pesi di queste connessioni durante il processo di addestramento, al fine di migliorare la performance della rete in un determinato compito.

Il funzionamento della Rete Neurale si basa sulla Teoria delle funzioni non lineari e della propagazione dell'errore. In pratica, la rete neurale è in grado di apprendere dalle informazioni contenute nei dati di input, attraverso l'adattamento dei pesi delle connessioni tra i neuroni, che viene eseguito durante la fase di addestramento della rete. Una volta addestrata, la rete neurale è in grado di effettuare previsioni o classificazioni su nuovi dati di input, utilizzando il modello matematico che ha appreso durante la fase di addestramento.

3. ELABORATO IN INTELLIGENZA ARTIFICIALE

3.1 INTRODUZIONE

Le malattie cardiache descrivono una serie di condizioni che colpiscono il cuore.

Le malattie cardiache sono una delle principali cause di morbilità e mortalità tra la popolazione mondiale. La previsione delle malattie cardiovascolari è considerata uno degli argomenti più importanti nella sezione della scienza dei dati clinici. La quantità di dati nel settore sanitario è enorme.

Se un paziente potesse essere identificato come una persona a rischio di sviluppare malattie cardiache, allora potrebbero essere prese misure preventive per ridurre il rischio, tra cui smettere di fumare, fare esercizio fisico regolare, mangiare sano e smettere/limitare il consumo di alcol.

3.2 SCOPO

In questo progetto di Intelligenza Artificiale applicherò tecniche di apprendimento automatico per classificare se una persona soffre o meno di Patologie cardiache.

L'obiettivo è rilevare con precisione la presenza di patologie cardiache in un paziente utilizzando tecniche di apprendimento automatico.

3.3 STRUMENTI UTILIZZATI

3.3.1 LINGUAGGIO

Linguaggio Python versione 3.10.6.

3.3.2 AMBIENTE DI SVILUPPO

IDE (Integrated Development Environment) Visual Studio code.

3.3.3 LIBRERIE

1. Pandas: Libreria Python utilizzata per la manipolazione e l'analisi dei dati, specialmente per la creazione di strutture di dati come DataFrame, utilizzati nell'elaborazione dei dati.
2. Matplotlib: Libreria Python utilizzata per la creazione di grafici e visualizzazioni dei dati.
3. Seaborn: Libreria Python basata su Matplotlib che fornisce un'interfaccia di alto livello per la creazione di visualizzazioni dei dati statistiche.
4. Scikit-learn: Libreria Python per l'apprendimento automatico, che offre una vasta gamma di algoritmi di apprendimento automatico, strumenti di valutazione e funzioni di pre-elaborazione dei dati.
5. Keras: Libreria Python per la creazione di reti neurali artificiali.

3.4 DATI UTILIZZATI

Il dataset ha 14 colonne e 303 righe e contiene informazioni mediche sui pazienti, generiche tuple del Dataset appagliono così:

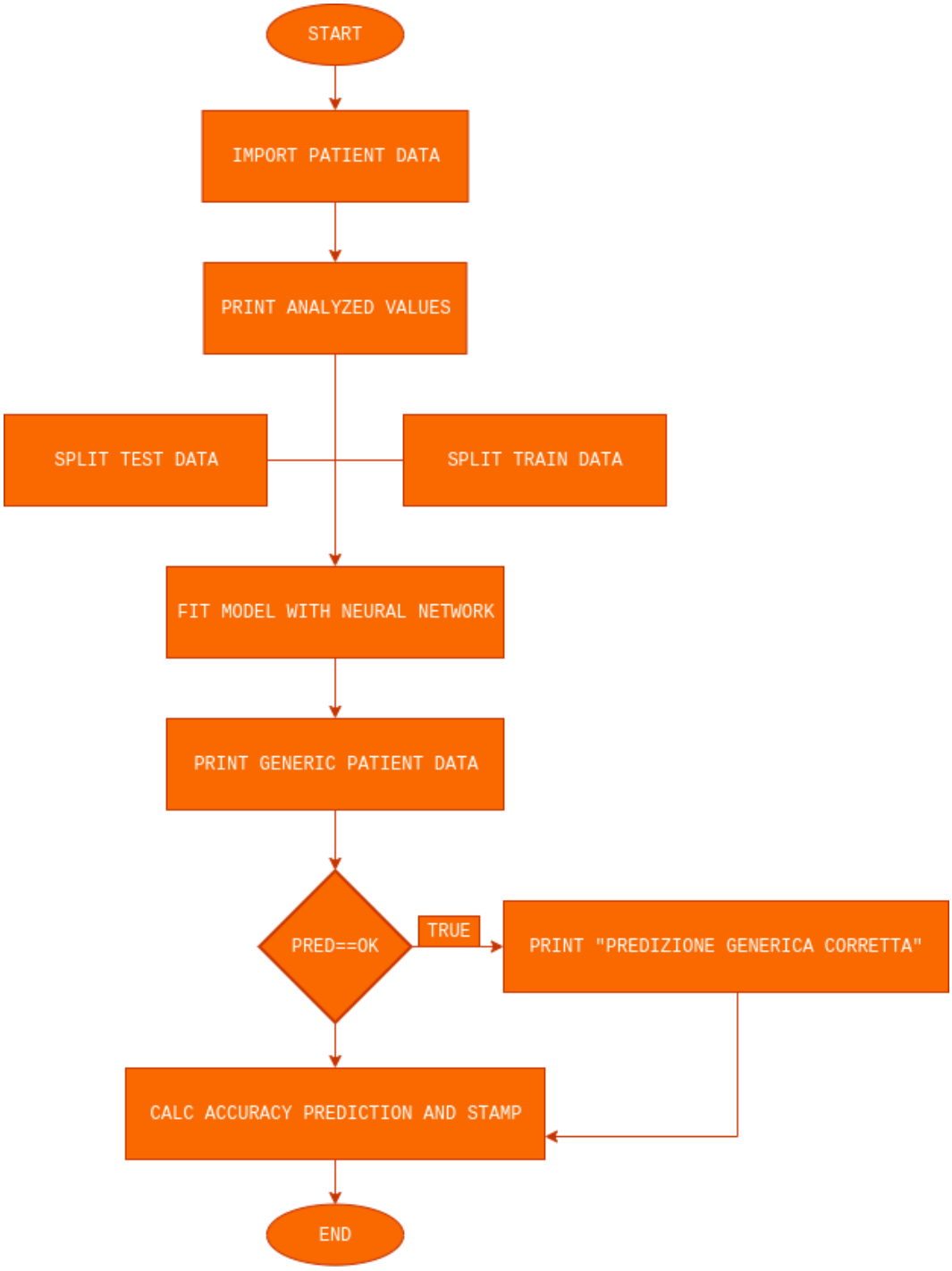
AGE	SEX	CP	TRESTBPS	CHOL	FBS	RESTECG	THALACH	EXANG	OLDPEAK	SLOPE	CA	THAL	TARGET
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
60	1	0	117	230	1	1	160	1	1.4	2	2	3	0

Il nome del Dataset è heart_research.csv ed è composto in questo modo:

Attributo	Descrizione	Tipo
Age	Età in anni	Intero
Sex	Sesso, 0=Maschio - 1=Femmina	Binario
Cp	Tipologia di dolore al petto, 0=Asintomatico - 1=Tipico Angina - 2=Atipico Angina - 3=Non Anginoso	Intero
Trestbps	Pressione sanguigna a riposo, in mmHg	Intero
Chol	Colesterolo totale, in mg/dl	Intero
Fbs	Livello di zucchero nel sangue a digiuno, 1=Alto > 120mg/dl - 0=Basso < 120mg/dl	Binario
Restecg	Risultati elettrocardiogramma a riposo, 1=Normale - 2=Anomalia onda ST-T - 3=Probabile ipertrofia ventricolare sinistra	Intero
Thalach	Frequenza cardiaca massima raggiunta durante un esercizio fisico	Intero
Exang	Angina indotta dall'esercizio fisico, 1=Sì - 0=No	Binario
Oldpeak	Depressione del segmento ST indotta dall'esercizio rispetto al riposo	Razionale
Slope	Pendenza del segmento ST dell'ECG durante l'esercizio, 0=Discendente - 1=Piatta - 2=Ascendente	Intero
Ca	Numero dei principali vasi sanguigni colorati tramite Fluoroscopia	Intero
Thal	Thallium Stress Test, 1=Normale - 2=Difetto fisso - 3=Difetto reversibile	Intero
Target	Patologia cardiaca, 0=Assente - 1=Presente	Intero

3.5 REALIZZAZIONE

3.5.1 DIAGRAMMA A BLOCCHI



3.5.2 CODICE

heart_disease_prediction.py > ...

```
1   import pandas as pd
2   import matplotlib.pyplot as plt
3   import seaborn as sns
4   from sklearn.model_selection import train_test_split
5   from sklearn.metrics import accuracy_score
6   from keras.models import Sequential
7   from keras.layers import Dense
8
```

Importazione delle librerie utilizzate nell'analisi dei dati e nell'apprendimento automatico.

```
11
12  dataset = pd.read_csv("heart_research.csv")
13
```

Utilizzo della libreria pandas per leggere un file CSV di dati e salvarlo in una variabile chiamata dataset.

In particolare, la funzione read_csv() di pandas viene utilizzata per leggere il file CSV dal percorso specificato ("heart_research.csv" in questo caso) e convertirlo in un oggetto DataFrame, che rappresenta una tabella di dati in formato tabellare.

Una volta che il file CSV viene letto e convertito in un oggetto DataFrame, è possibile manipolare e analizzare i dati utilizzando le funzioni e i metodi disponibili in pandas, ad esempio per visualizzare le prime righe dei dati, eseguire operazioni di filtraggio e raggruppamento, e altro ancora.

```
18
19  # CORRELAZIONI TRA I DATI
20  corrmat = dataset.corr()
21  top_corr_features = corrmat.index
22  plt.figure(figsize=(16,16))
23  #plot heat map
24  g=sns.heatmap(dataset[top_corr_features].corr(),annot=True,cmap="RdYlGn")
25  plt.show()
26  #pic1
27  sns.barplot(x=dataset["sex"],y=dataset["target"])
28  plt.show()
29  #pic2
30  sns.barplot(x=dataset["cp"],y=dataset["target"])
31  plt.show()
32  #pic3
```

Utilizzo di diverse funzioni della libreria seaborn per visualizzare le relazioni tra le variabili del dataset. Questi grafici possono aiutare a identificare le relazioni tra le variabili del dataset e la variabile target, nonché a comprendere la distribuzione dei dati all'interno del dataset.

```
53
54  predictors = dataset.drop("target",axis=1)
55  target = dataset["target"]
56
57  X_train,X_test,Y_train,Y_test = train_test_split(predictors,target,test_size=0.20,random_state=0)
58
59  model = Sequential()
60  model.add(Dense(11,activation='relu',input_dim=13))
61  model.add(Dense(1,activation='sigmoid'))
62  model.compile(loss='binary_crossentropy',optimizer='adam',metrics=['accuracy'])
63  model.fit(X_train,Y_train,epochs=2000)
64
```

Utilizzo dela libreria sklearn per dividere il dataset in un set di dati di addestramento (X_train, Y_train) e un set di dati di test (X_test, Y_test) utilizzando la funzione train_test_split(). Successivamente, viene definito un modello di rete neurale artificiale utilizzando la libreria keras ed infine, il modello viene addestrato (fit()) sul set di dati di addestramento (X_train, Y_train) per un numero di epoche pari a 2000. Questo significa che il modello viene esposto ripetutamente ai dati di addestramento per migliorare la sua capacità di predire la variabile target.

```
65
66  Y_pred_nn = model.predict(X_test)
67
68  print()
69  print("--- IL PAZIENTE CON QUESTI VALORI : ")
70  print(X_test.iloc[0])
71  print(f"--- HA DATO QUESTO RISULTATO (0 Assente - 1 Presente) : {Y_pred_nn[0]}")
72  #
73  print(f"--- IL VALORE REALE DEL PAZIENTE E' (0 Assente - 1 Presente) : {Y_test.iloc[0]} ")
74  a = float(Y_pred_nn[0])
75
76  b = 1-a
77
78  valore = b
79
80  if Y_test.iloc[0]==0 and a<0.500000 :
81      print("--- LA PREDIZIONE E' STATA : CORRETTA")
82
```

Viene eseguito un test generico su un paziente utilizzando il primo esempio nel set di dati di test (X_test.iloc[0]). Vengono quindi stampati i valori di input del paziente, il risultato predetto dal modello e il valore reale della variabile target del paziente.

Infine, viene eseguito un controllo della predizione fatta dal modello sul paziente di prova e se la predizione è corretta (cioè il paziente non ha la patologia e la probabilità predetta è inferiore a 0,500000), viene stampato un messaggio che indica che la predizione è stata corretta.

```
78
79  rounded = [round(x[0]) for x in Y_pred_nn]
80  Y_pred_nn = rounded
81  print()
82  score_nn = round(accuracy_score(Y_pred_nn,Y_test)*100,2)
83  print("--- La PRECISIONE DELLA PREDIZIONE GENERALE E' STATA DEL : "+str(score_nn)+" %")
84
```

Calcolo della Accuracy Generale del modello addestrato.

3.6 RISULTATI

3.6.1 OUTPUT

```
Epoch 1964/2000
8/8 [==============================] - 0s 1ms/step - loss: 0.3474 - accuracy: 0.8388
Epoch 1965/2000
8/8 [==============================] - 0s 2ms/step - loss: 0.3347 - accuracy: 0.8554
Epoch 1966/2000
8/8 [==============================] - 0s 1ms/step - loss: 0.3300 - accuracy: 0.8512
Epoch 1967/2000
8/8 [==============================] - 0s 1ms/step - loss: 0.3284 - accuracy: 0.8554
Epoch 1968/2000
8/8 [==============================] - 0s 1ms/step - loss: 0.3330 - accuracy: 0.8595
Epoch 1969/2000
8/8 [==============================] - 0s 1ms/step - loss: 0.3415 - accuracy: 0.8471
Epoch 1970/2000
8/8 [==============================] - 0s 1ms/step - loss: 0.3390 - accuracy: 0.8512
```

Di seguito è riportato il risultato di un addestramento di una rete neurale. L'addestramento è stato effettuato per un totale di 2000 epoche, e il risultato di ogni epoca viene riportato in termini di perdita e accuratezza. L'accuratezza rappresenta la percentuale di immagini classificate correttamente dalla rete neurale.

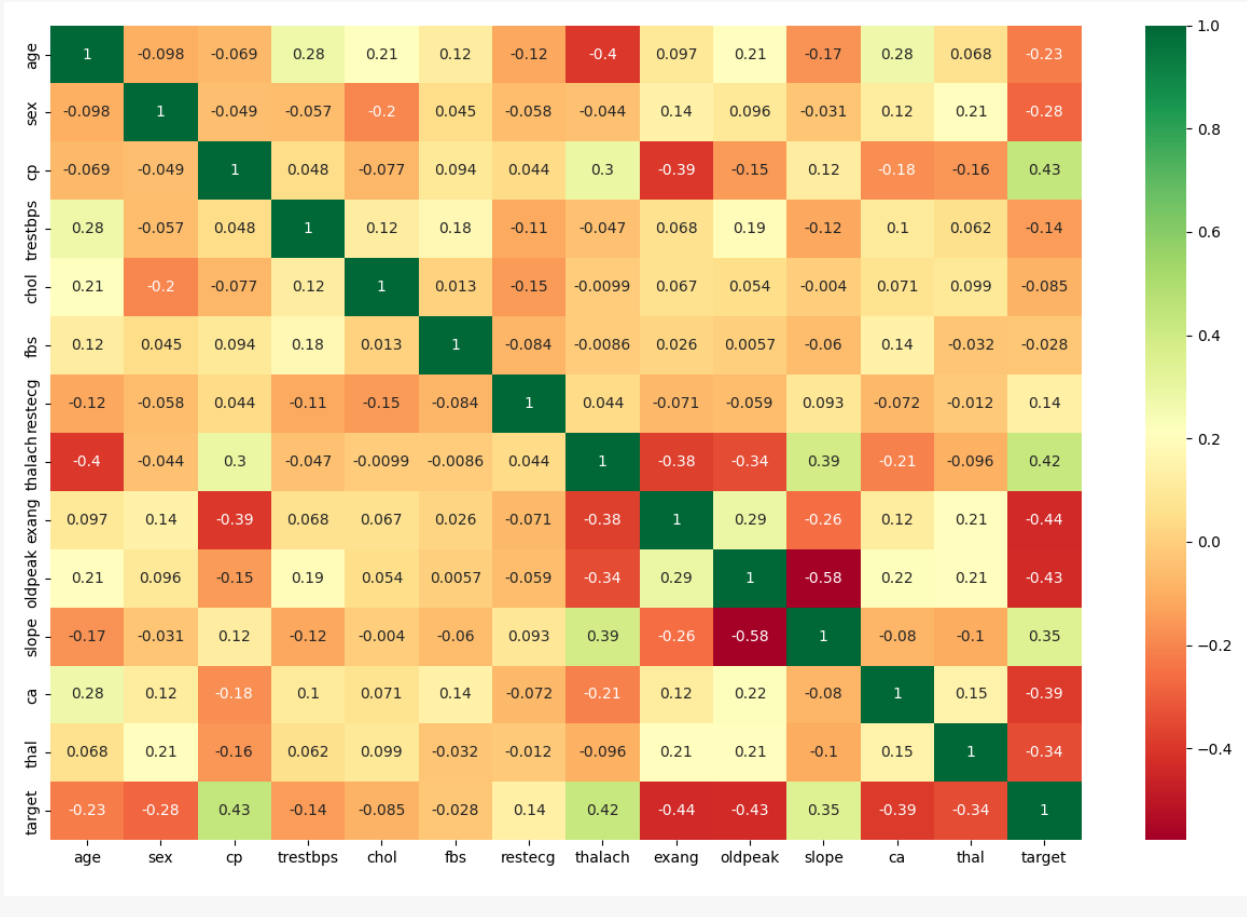
```
--- IL PAZIENTE CON QUESTI VALORI :
age          64.0
sex           1.0
cp            2.0
trestbps    125.0
chol        309.0
fbs           0.0
restecg       1.0
thalach     131.0
exang         1.0
oldpeak       1.8
slope         1.0
ca            0.0
thal          3.0
Name: 225, dtype: float64
--- HA DATO QUESTO RISULTATO (0 Assente - 1 Presente) : [0.21311358]
--- IL VALORE REALE DEL PAZIENTE E' (0 Assente - 1 Presente) : 0
--- LA PREDIZIONE E' STATA : CORRETTA
```

```
  --- La PRECISIONE DELLA PREDIZIONE GENERALE E' STATA DEL : 80.33 %
asaggio@asaggio-Vostro-3500:~/progetto_uni$
```

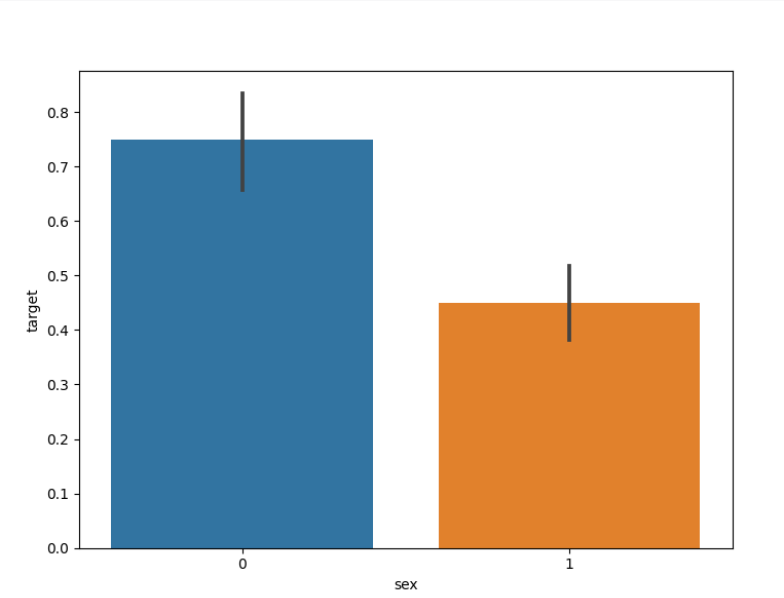
3.6.2 CORRELAZIONI TRA I DATI

L'analisi delle correlazioni tra le features del dataset è un'attività fondamentale nell'ambito dell'analisi dei dati. Essa ci permette di esplorare le relazioni tra le diverse variabili del dataset e di capire meglio come queste variabili si influenzano reciprocamente.

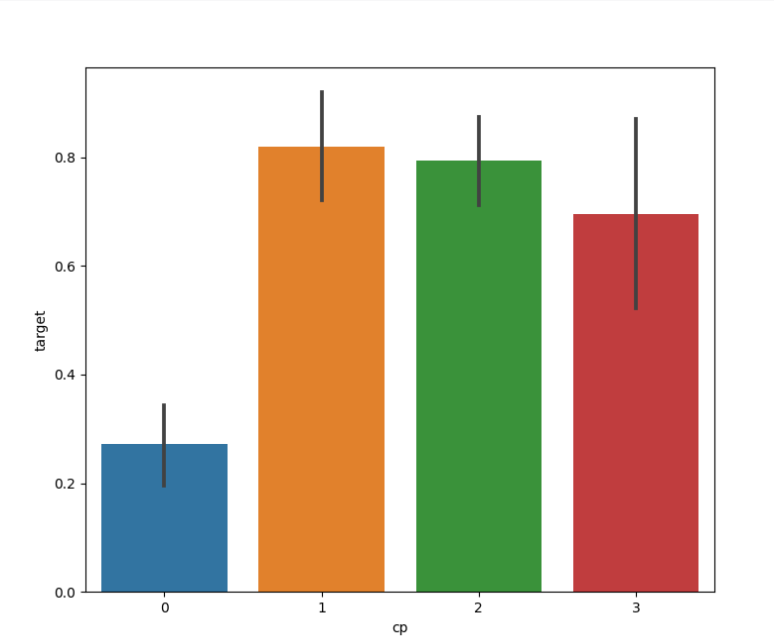
Una delle tecniche più comuni per esplorare le correlazioni tra le features è l'utilizzo di un heatmap delle correlazioni:



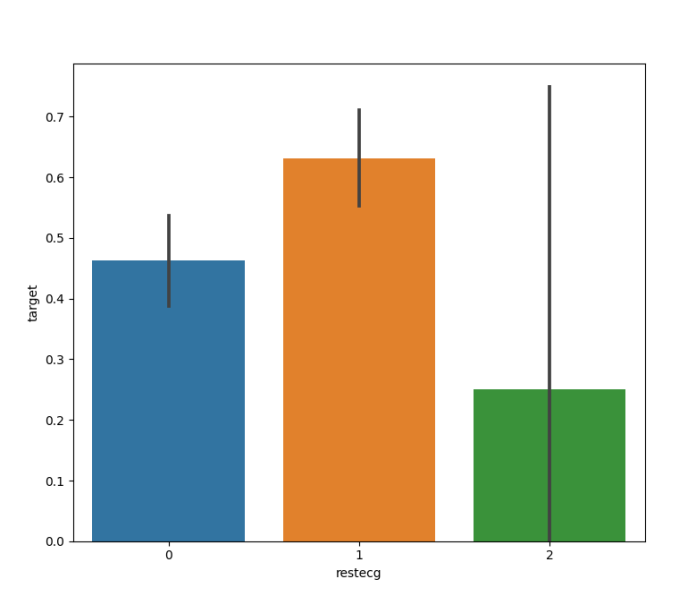
1. Heatmap delle correlazioni tra le diverse features del dataset, visualizzato tramite una matrice di colori. I valori più chiari indicano una correlazione positiva, mentre quelli più scuri indicano una correlazione negativa.



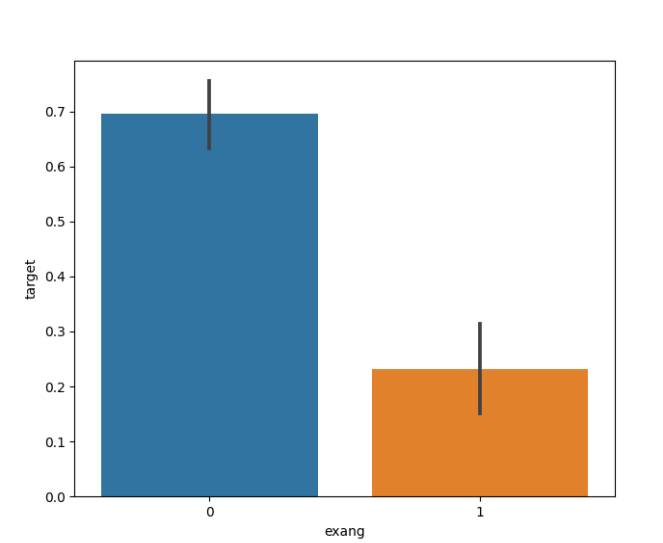
2. Barplot che mostra la relazione tra il genere dei pazienti (maschio o femmina) e la presenza di malattie cardiache.



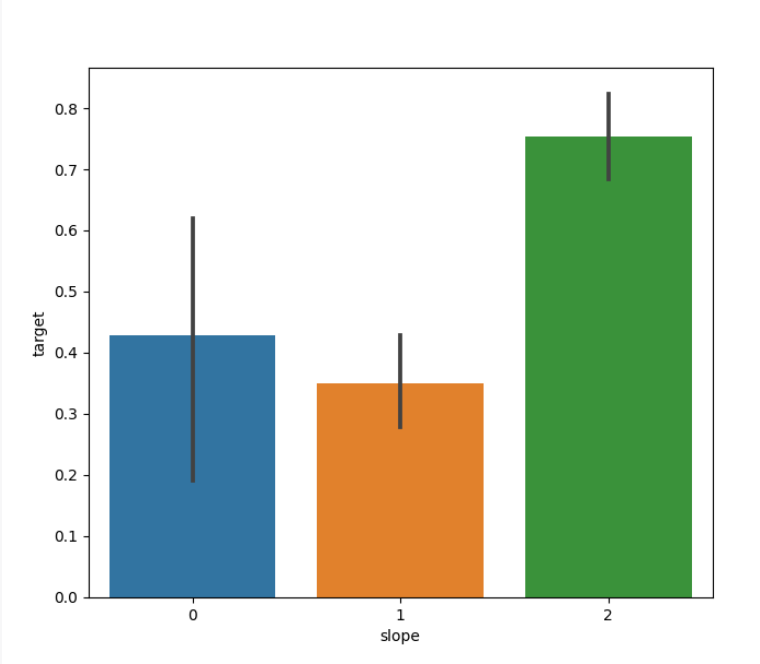
3. Barplot che mostra la relazione tra il tipo di dolore toracico (typical angina, atypical angina, non-anginal pain, asymptomatic) e la presenza di malattie cardiache.



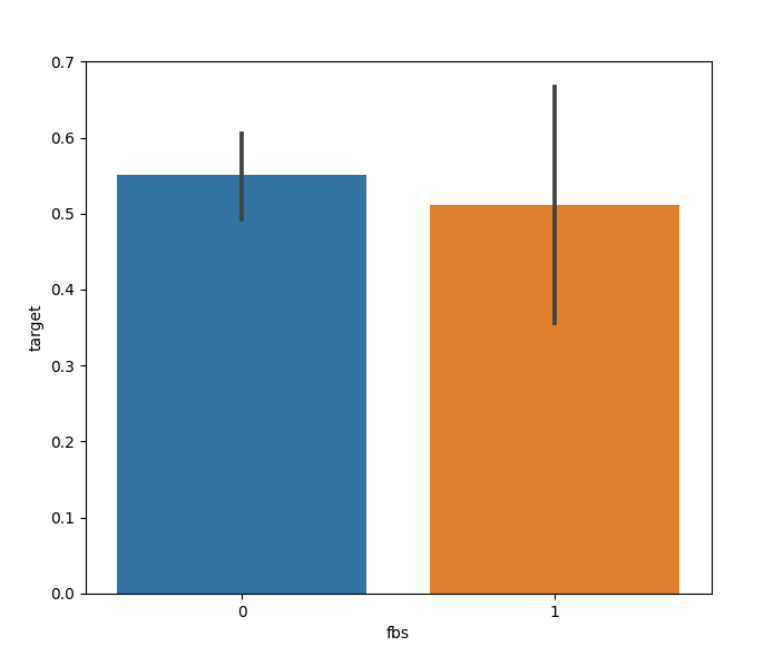
4. Barplot che mostra la relazione tra i risultati dell'elettrocardiogramma a riposo (valori 0, 1, 2) e la presenza di malattie cardiache.



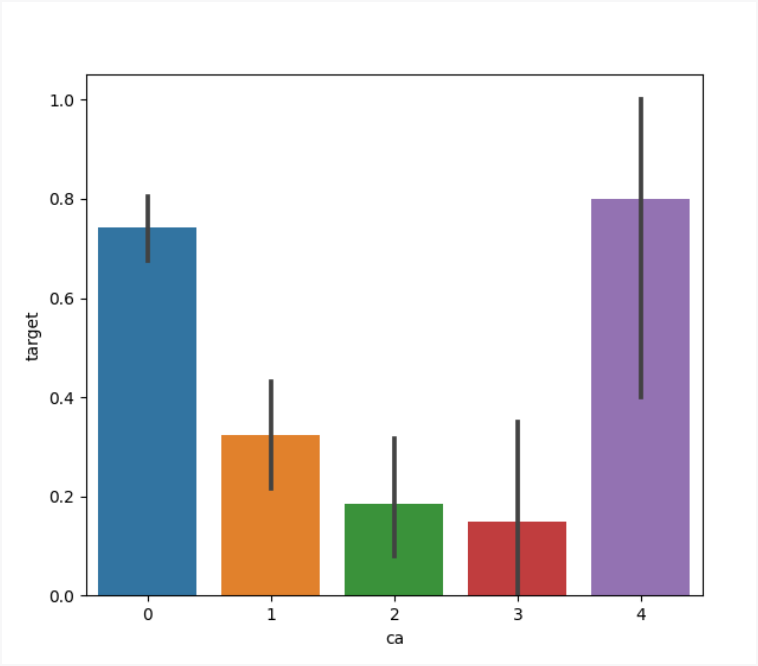
5. Barplot che mostra la relazione tra la presenza di angina indotta dall'esercizio fisico e la presenza di malattie cardiache.



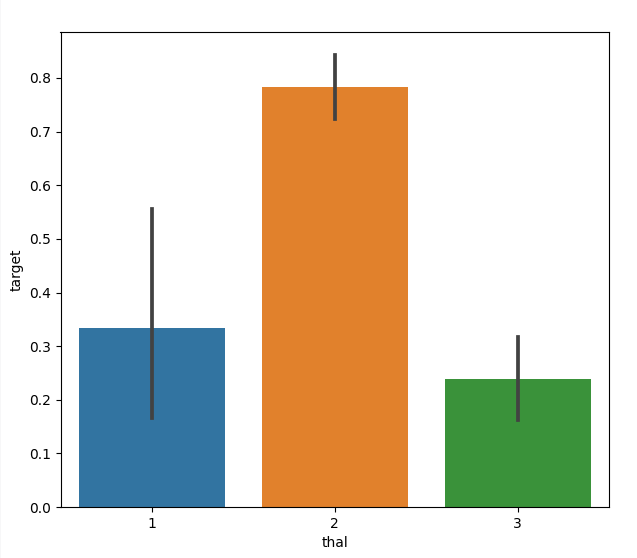
6. Barplot che mostra la relazione tra la pendenza del segmento ST durante l'esercizio fisico e la presenza di malattie cardiache.



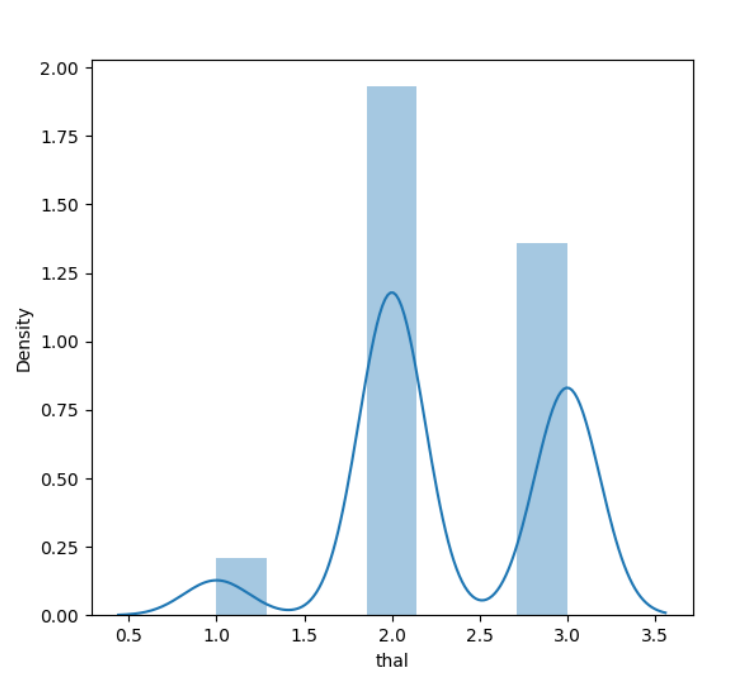
7. Barplot che mostra la relazione tra il livello di zucchero nel sangue a digiuno e la presenza di malattie cardiache.



8. Barplot che mostra la relazione tra il numero di vasi maggiori colorati durante una fluoroscopia e la presenza di malattie cardiache.



9. Barplot che mostra la relazione tra il tipo di difetto thalassemia (3 = normale; 6 = difetto fisso; 7 = difetto reversibile) e la presenza di malattie cardiache.



10. Istogramma che mostra la distribuzione dei valori della colonna thal. L'istogramma mostra la frequenza di ogni categoria all'interno del dataset e permette di capire se una particolare categoria è rappresentata in modo significativo rispetto alle altre.

4. CONCLUSIONI E SVILUPPO FUTURO

Il valore di accuracy score ottenuto è del 80.33%, il che indica che il modello ha ottenuto una buona capacità di generalizzazione sui dati di test. Tuttavia, ci sono diverse possibilità di sviluppo per migliorare il modello. Ad esempio, potrebbe essere utile esplorare diverse architetture di rete neurale, come l'aggiunta di più strati nascosti o di tecniche di regolarizzazione per prevenire l'overfitting. Inoltre, si potrebbe considerare l'utilizzo di tecniche di feature engineering per creare nuove feature dai dati esistenti. Infine, potrebbe essere utile esplorare altre tecniche di apprendimento automatico, come le foreste casuali o i support vector machine, per confrontare le performance del modello di rete neurale con quelle di altri modelli.