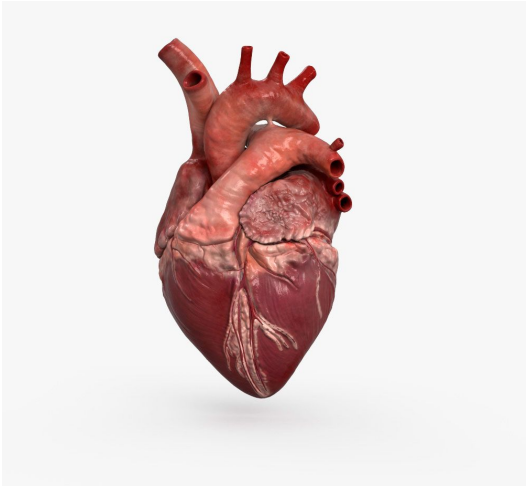


Heart Disease

Machine learning - Febbraio 2025

Realizzato da:
Andrea Spagnolo 879254
Davide Falanga 866053

Descrizione del problema

Dominio: Medico

Dataset: Heart Failure Prediction Dataset (https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data)

Obiettivo: Individuazione di un disturbo cardiaco

L'insufficienza cardiaca rappresenta la causa principale di mortalità a livello mondiale, richiedendo approcci efficaci per la diagnosi precoce e la prevenzione.

Tipologia di problema: Classificazione binaria

Exploratory Data Analysis

Il dataset si compone delle seguenti feature e target

Feature	Descrizione	Dominio
Age	età in anni del paziente	Valore intero
Sex	sesso del paziente	{M, F}
ChestPainType	tipologia di dolore al petto	{TA, ATA, NAP, ASY}
RestingBP	pressione arteriosa a riposo misurata in millimetri di mercurio (mm Hg)	Valore continuo
Cholesterol	livello di colesterolo sierico (mg/dl)	Valore continuo
FastingBS	glicemia a digiuno	Valore booleano {0, 1} 1: se > 120 mg/dl, 0: altrimenti
RestingECG	risultati dell'elettrocardiogramma a riposo	{Normal, ST, LVH}
MaxHR	massima frequenza cardiaca raggiunta (BPM)	Valore intero
ExerciseAngina	angina indotta dall'esercizio	Valore booleano {0, 1}
Oldpeak	depressione del tratto ST sull'ECG	Valore continuo
ST_Slope	categoria del segmento ST al picco dell'esercizio fisico	{UP, Flat, Down}
HeartDisease (target)	Presenza di un disturbo cardiaco	Valore booleano {0, 1}

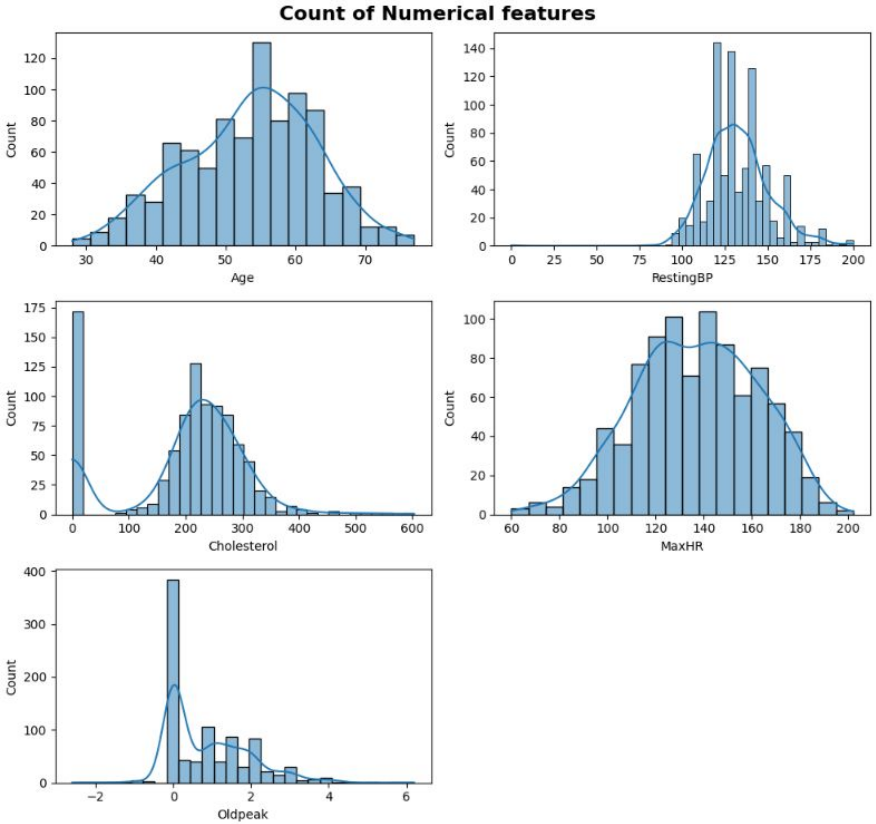
di istanze: 918

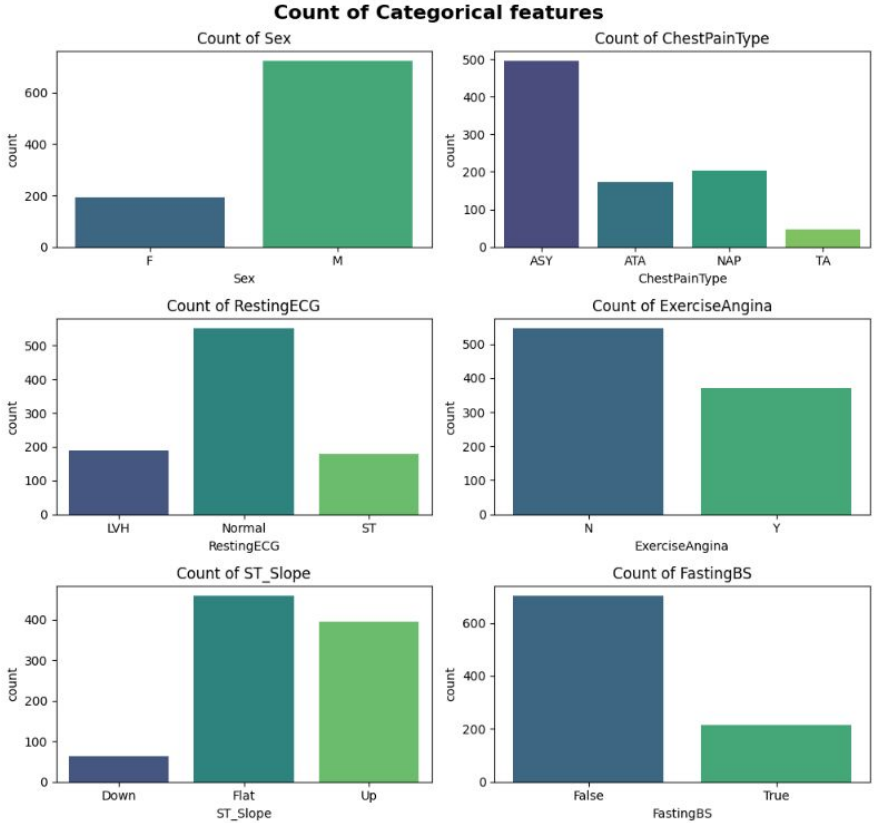
Pulizia del dataset e cast

- Rimozione dei **duplicati**, assenza di **valori mancanti**
- Cast delle variabili '**object**' a '**category**'
- Cast della variabile **HeartDisease** e **FastingBS** a **boolean**
- Codifica con **OneHotEncoder** delle feature categoriche
- Applicata una normalizzazione delle variabili numeriche con **StandardScaler**

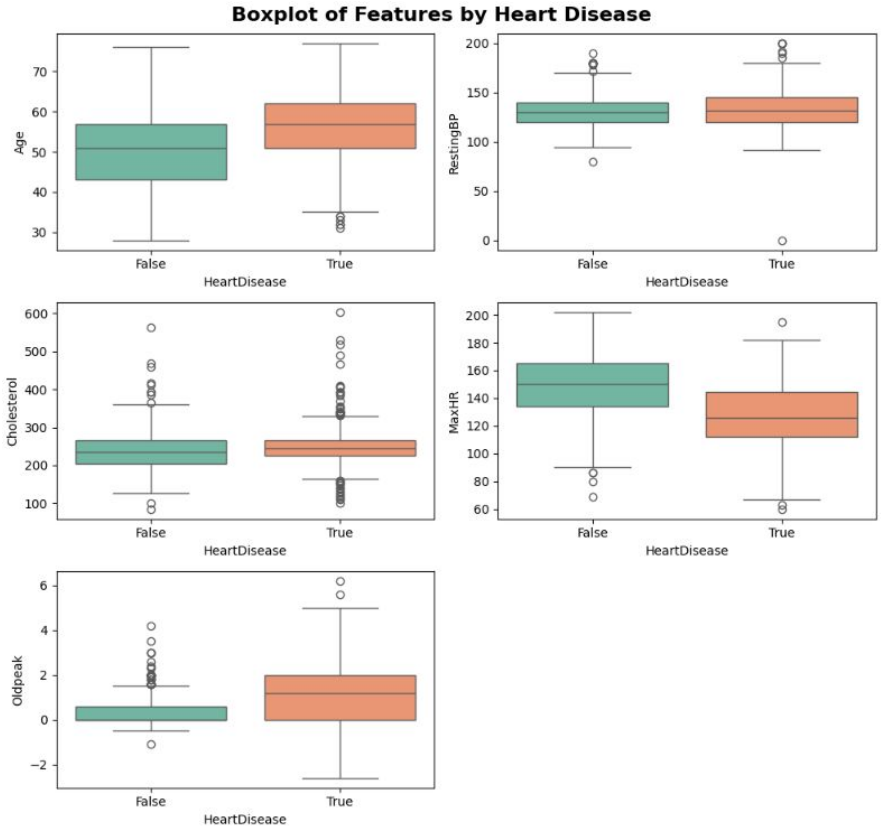
```
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             918 non-null    int64
 1   Sex             918 non-null    object
 2   ChestPainType   918 non-null    object
 3   RestingBP       918 non-null    int64
 4   Cholesterol     918 non-null    int64
 5   FastingBS       918 non-null    int64
 6   RestingECG      918 non-null    object
 7   MaxHR           918 non-null    int64
 8   ExerciseAngina  918 non-null    object
 9   Oldpeak         918 non-null    float64
 10  ST_Slope        918 non-null    object
 11  HeartDisease    918 non-null    int64
```

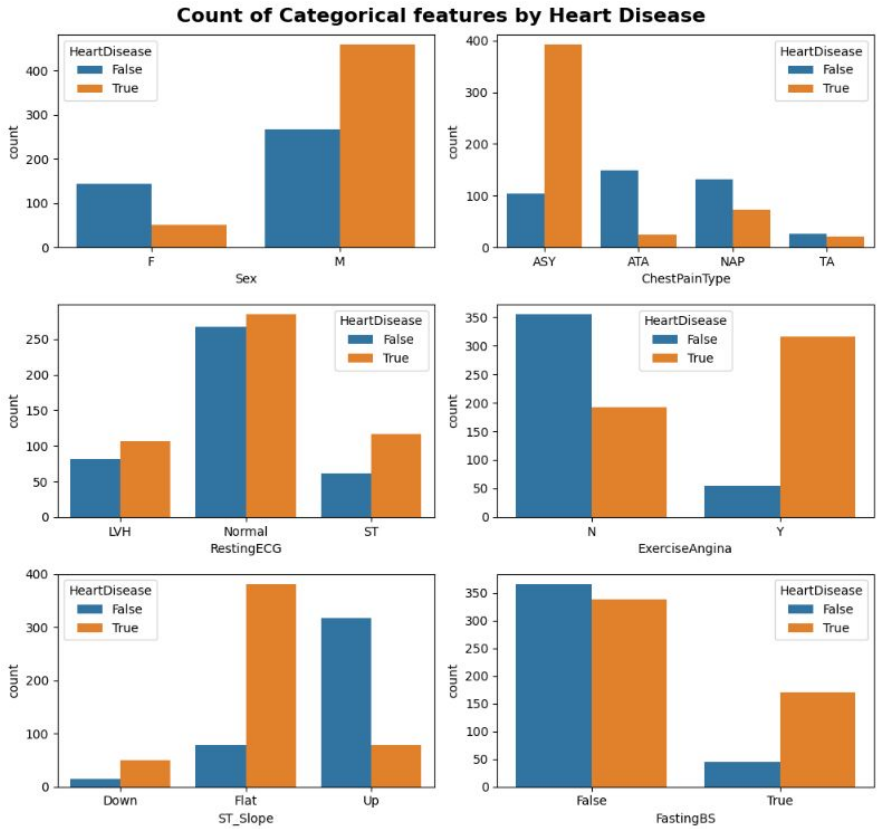
Distribuzioni dei dati



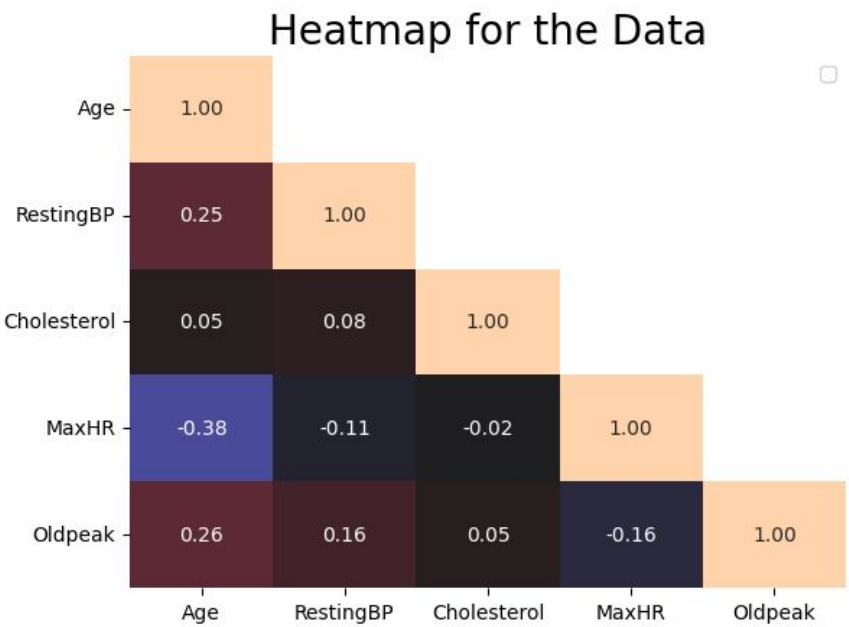


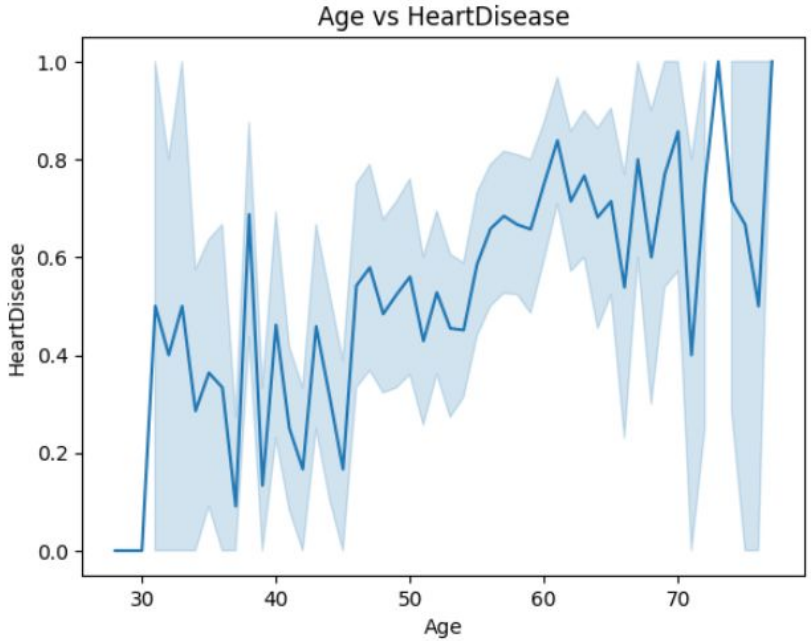
Distribuzioni dei dati in relazione al target



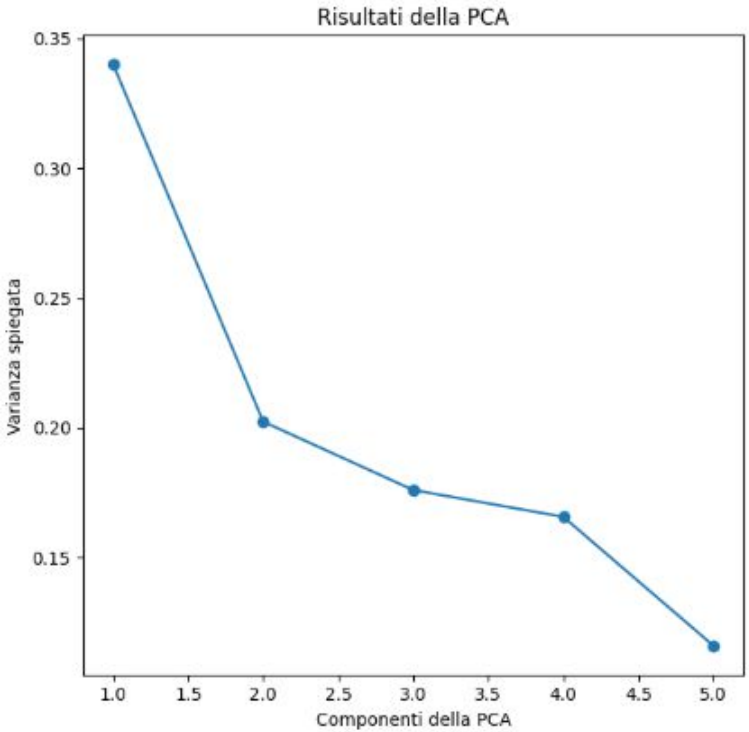


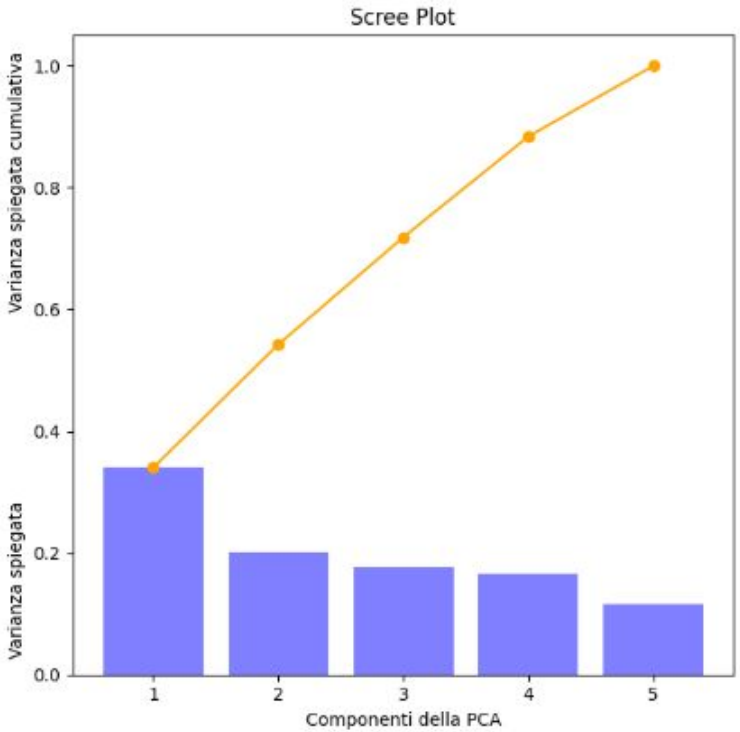
Analisi di correlazione





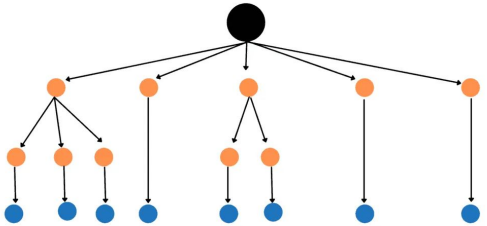
Principal Component Analysis

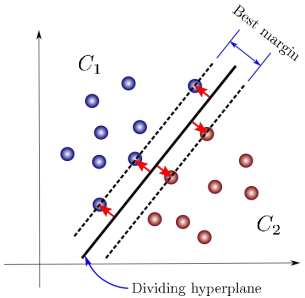


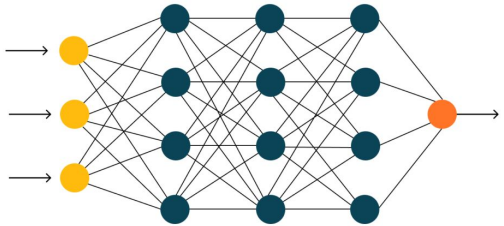


Modelli di Machine Learning

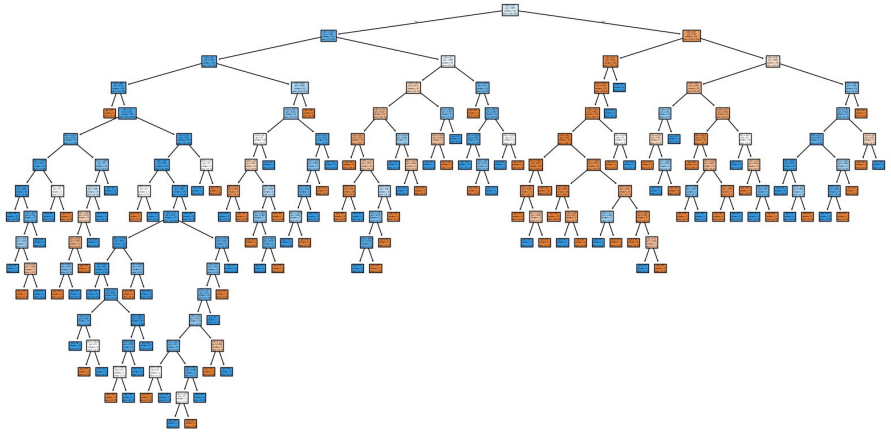
A questo punto si procede con i modelli di ML, dopo una opportuna costruzione dei dataset di training (80%) e di testing (20%).



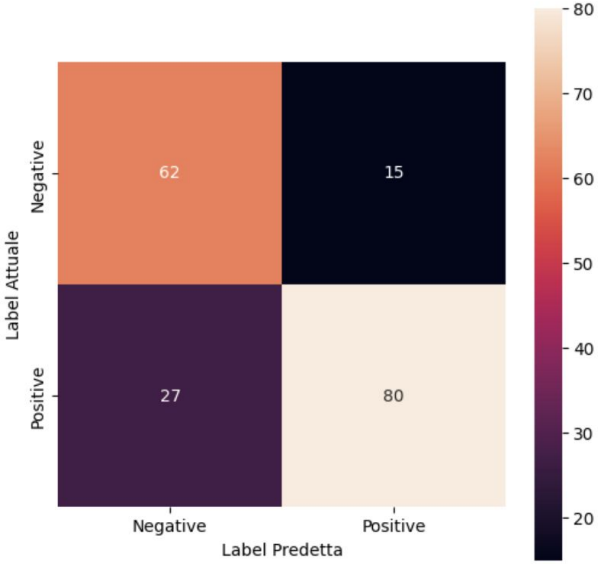




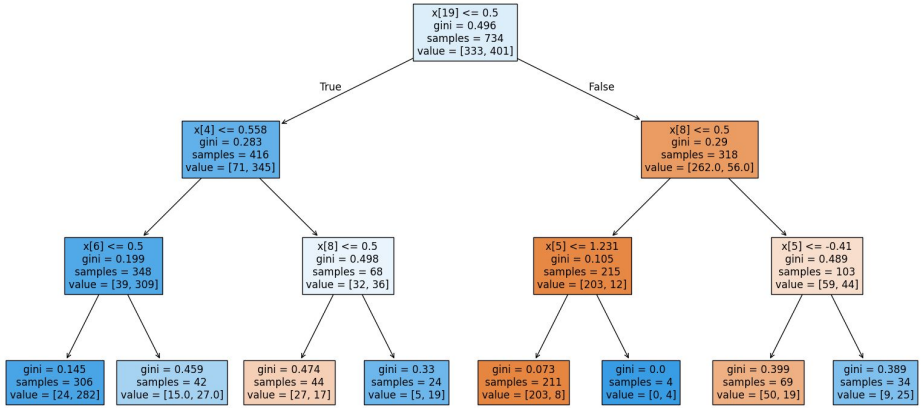
Albero di decisione

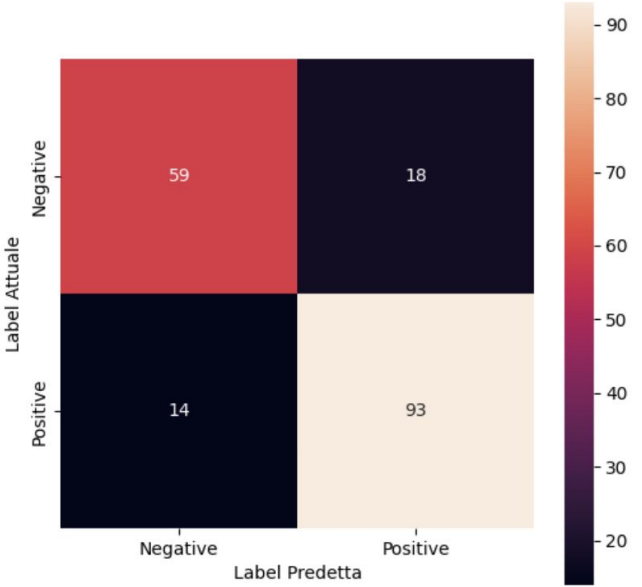


training accuracy	test accuracy
1.0	0.78



Albero di decisione ottimizzato



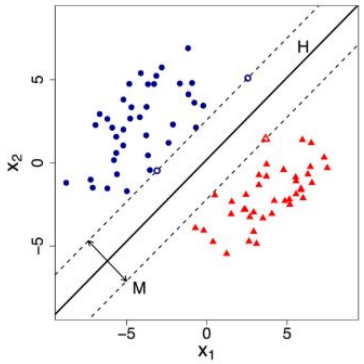


training accuracy	test accuracy
0.88 (-12%)	0.83 (+5%)

Grid Search per la ricerca degli iperparametri:

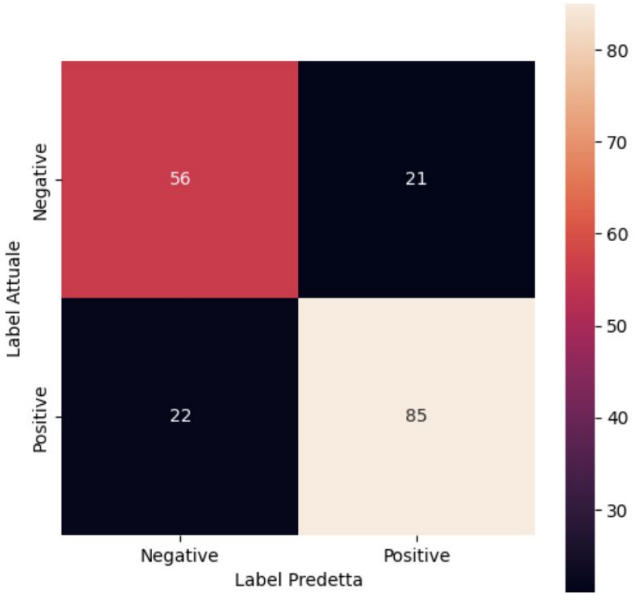
- **ccp alpha**: 0.0
- **criterion**: gini
- **max_depth**: 3
- **splitter**: best

Support Vector Machines



Iperparametri:

- kernel='rbf'
- C = 1000000



training accuracy	test accuracy
1.0	0.77

SVM ottimizzato

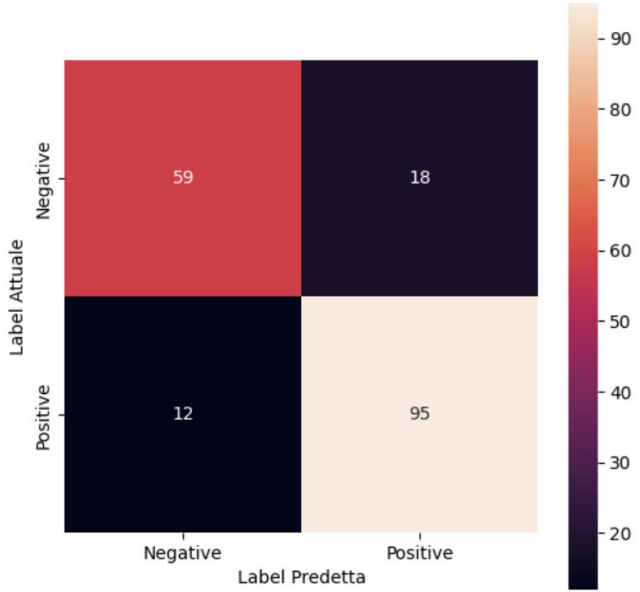
Griglia di parametri per GridSearchCV:

```
[{'kernel': ['rbf'], 'gamma': [0.01, 0.001, 0.0001, 0.00001],
'C': [0.001, 0.10, 0.1, 10, 25, 50, 100, 1000]},
{'kernel': ['poly'], 'gamma': [0.01, 0.001, 0.0001, 0.00001],
'C': [0.001, 0.10, 0.1, 10, 25, 50, 100, 1000]},
{'kernel': ['linear'], 'C': [0.001, 0.10, 0.1, 10, 25, 50, 100,
1000]}]
```

Miglior configurazione:

```
{'C': 100, 'gamma': 0.01, 'kernel': 'rbf'}
```

training accuracy	test accuracy
0.91 (-9%)	0.84 (+7%)



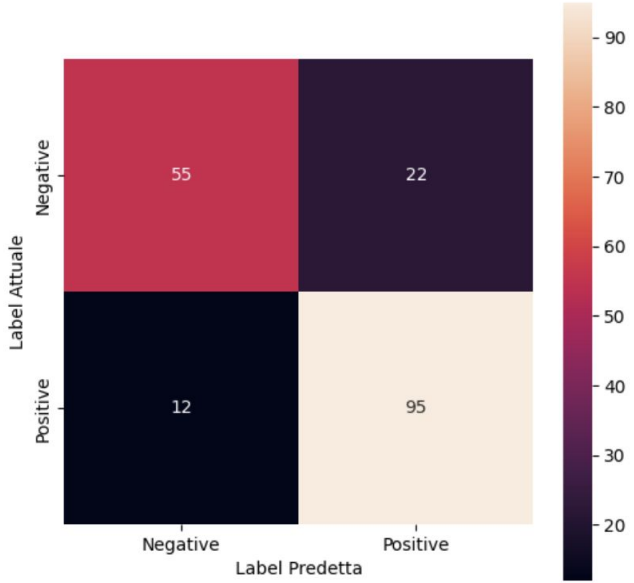
Rete neurale

Architettura rete:

- 1° strato: 100 neuroni, attivazione ReLU
- 2° strato: 50 neuroni, attivazione ReLU
- Output: 1 neurone, attivazione sigmoid

Training:

- 50 epoche
- batch size 32

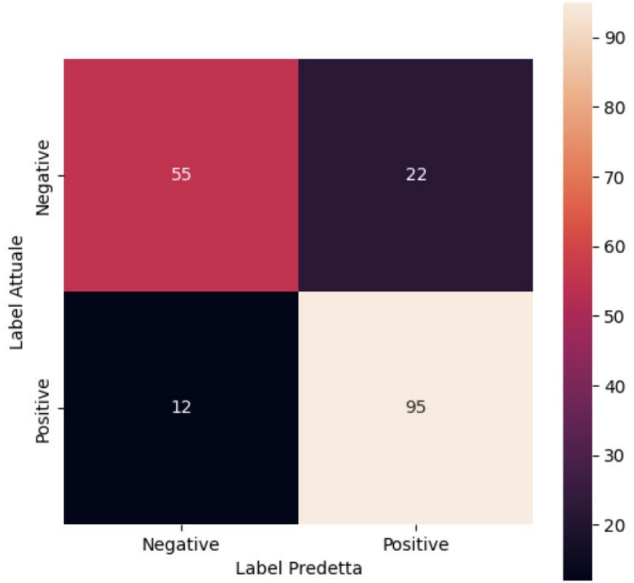


training accuracy	test accuracy	test loss	training loss
0.97	0.83	0.67	0.09

Rete neurale ottimizzata

Grid Search per la ricerca degli iperparametri

- **Attivazione**: ReLU,
- **Batch size**: 16
- **Epoche**: 50
- **Layer nascosti**: 50 neuroni
- **Ottimizzatore**: Adam
- **Dropout**: 0.2



training accuracy	test accuracy	test loss	training loss
0.92 (-5%)	0.82 (-1%)	0.41 (-38.8%)	0.23 (+155.5%)

Conclusioni

- La SVM e l’albero di decisione risultano essere i migliori compromessi garantendo buona precisione e tempi di calcolo ridotti
- La rete neurale ha raggiunto buoni risultati, ma la presenza non trascurabile di overfitting e tempi di calcolo molto onerosi, non la rendono la scelta migliore per l’analisi di questo dataset