TITANIC

Prevedere la Sopravvivenza con il Machine Learning

INTRODUZIONE AL PROBLEMA

Il progetto si basa sul famoso dataset del Titanic. L'obiettivo è costruire un modello di Machine Learning per prevedere se un passeggero è sopravvissuto o meno al naufragio, basandosi su alcune delle sue caratteristiche.

IL PROCESSO DI MACHINE LEARNING

1. Caricamento e Analisi dei Dati

L’analisi dei dati è stata il primo passo del progetto. È stata condotta un’esplorazione completa del dataset per comprenderne la struttura, i tipi di variabili e la presenza di eventuali valori mancanti. Sono state generate statistiche di base e visualizzazioni per evidenziare la sopravvivenza in base a genere, classe e porto d’imbarco. È stata inoltre analizzata la distribuzione dell’età secondo le categorie socio-demografiche, e sono state create heat map per osservare le correlazioni tra le variabili.

2. Suddivisione del data set

Successivamente, il dataset è stato suddiviso in train e test set per garantire una valutazione affidabile del modello. Sucessivamente e stato ulteriormente diviso in Train, Validation e Test

3. Grid Search CV

Una fase cruciale è stata la gestione dei valori mancanti. In particolare, sono stati trattati i dati assenti nelle colonne Age ed Embarked. Per la variabile Age, sono state confrontate diverse strategie: sostituzione con la media, con la mediana globale o con la mediana calcolata per genere e classe. I risultati di queste strategie sono stati valutati tramite Grid Search per identificare la soluzione migliore.

IL PROCESSO DI MACHINE LEARNING

4. Sostituzione valori mancati

I base ai risultati della Grid Search del punto quattro sono stati sostitui i valori mancanti con la mediana dell'età calcolata per genere raggruppato per classe.

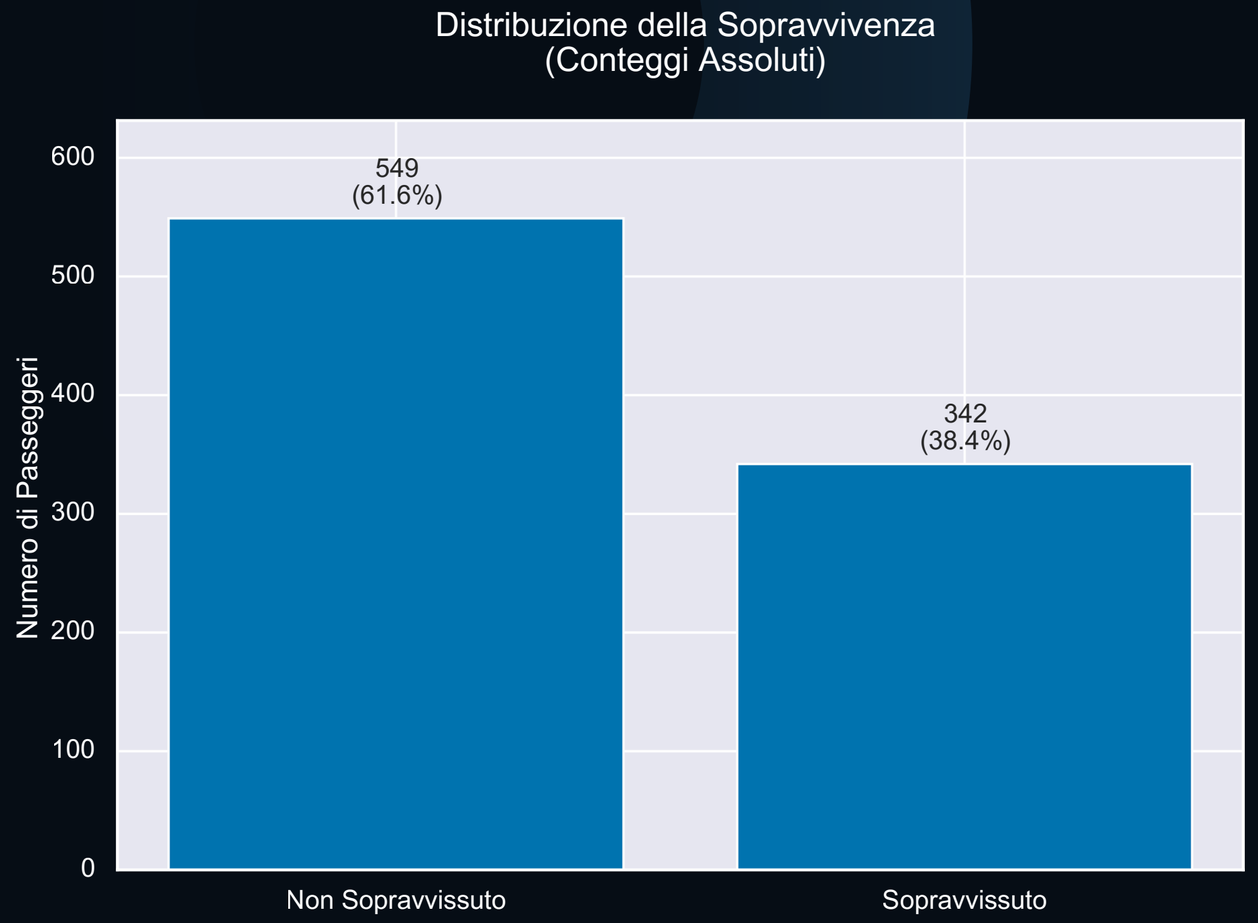
5. Encoding Variabili Categoriche con One-Hot

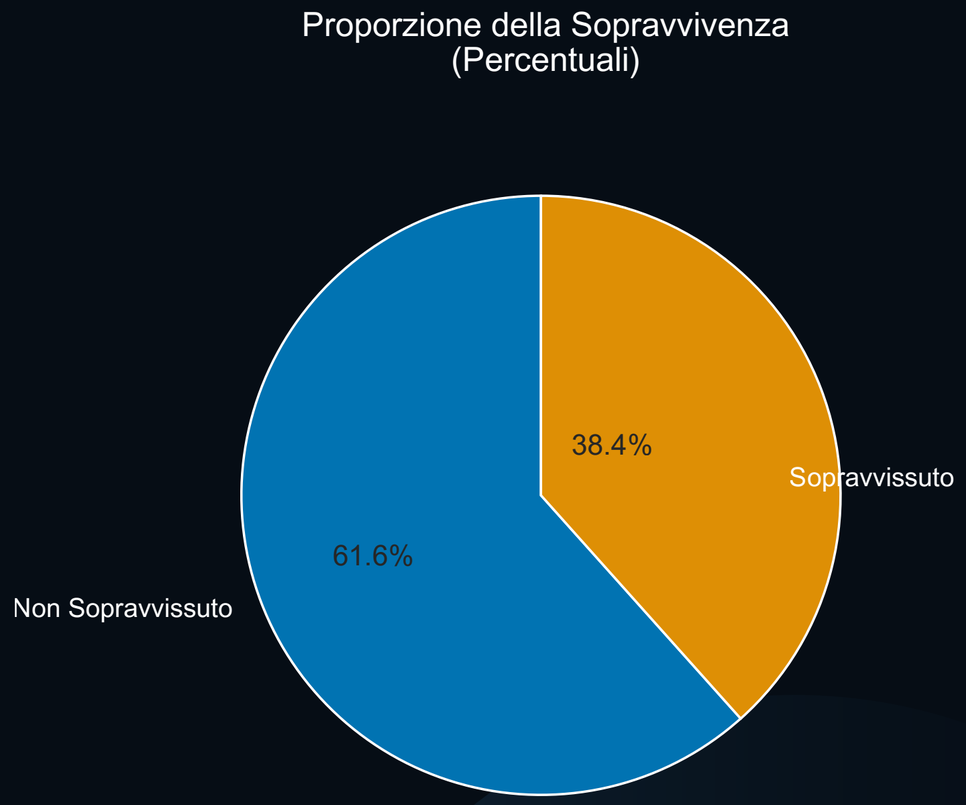
Le variabili categoriche sono state poi codificate tramite One-Hot Encoding, rendendo il dataset pronto per l'addestramento del modello. Questo passaggio ha garantito che tutte le informazioni rilevanti fossero utilizzabili in modo coerente dall'algoritmo.

6. Implementazione Decision Tree Classifer

Infine, è stato implementato un Decision Tree Classifier. Il modello è stato addestrato e validato per determinare la profondità ottimale dell'albero, e le performance sono state valutate sul test set. I risultati finali hanno permesso di trarre conclusioni significative sulla capacità del modello di predire la sopravvivenza dei passeggeri.

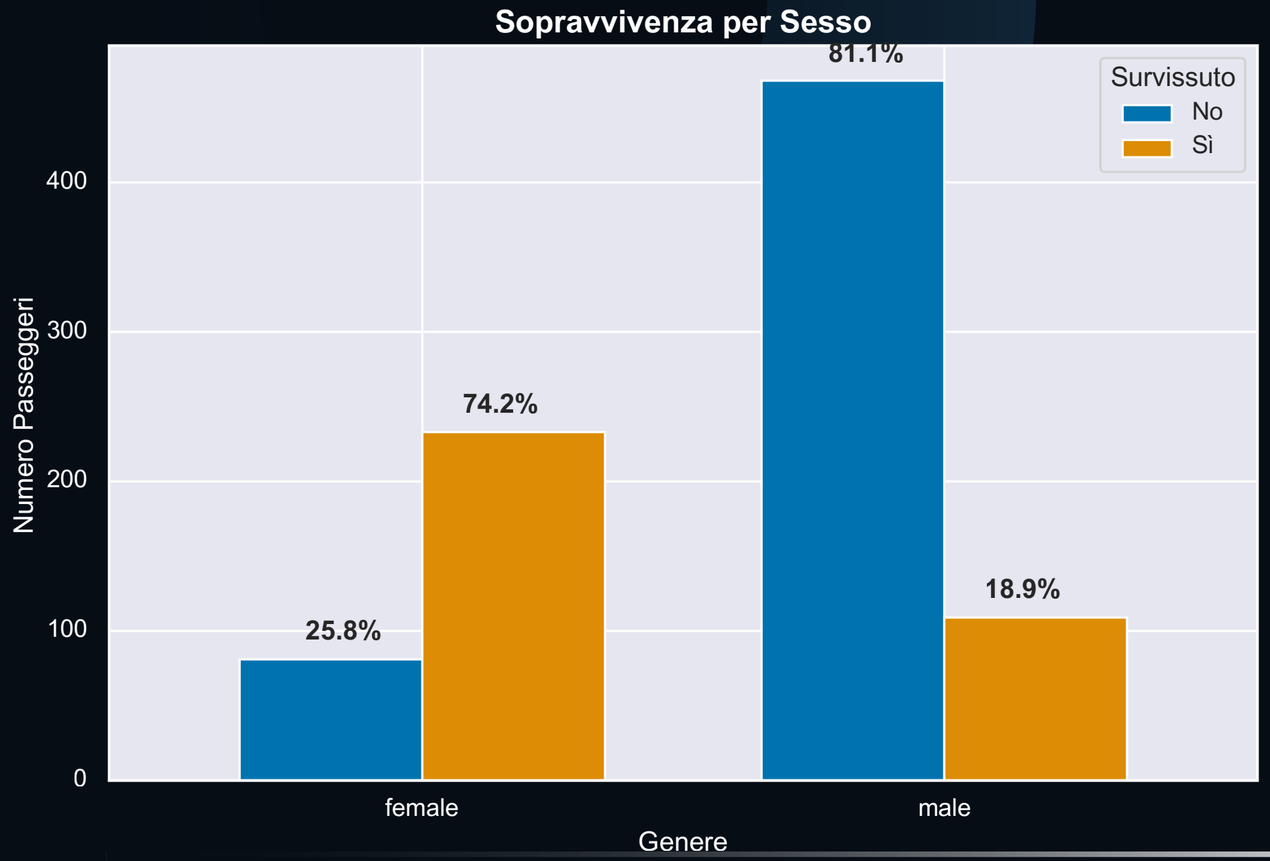
Analisi della Sopravvivenza dei Passeggeri del Titanic

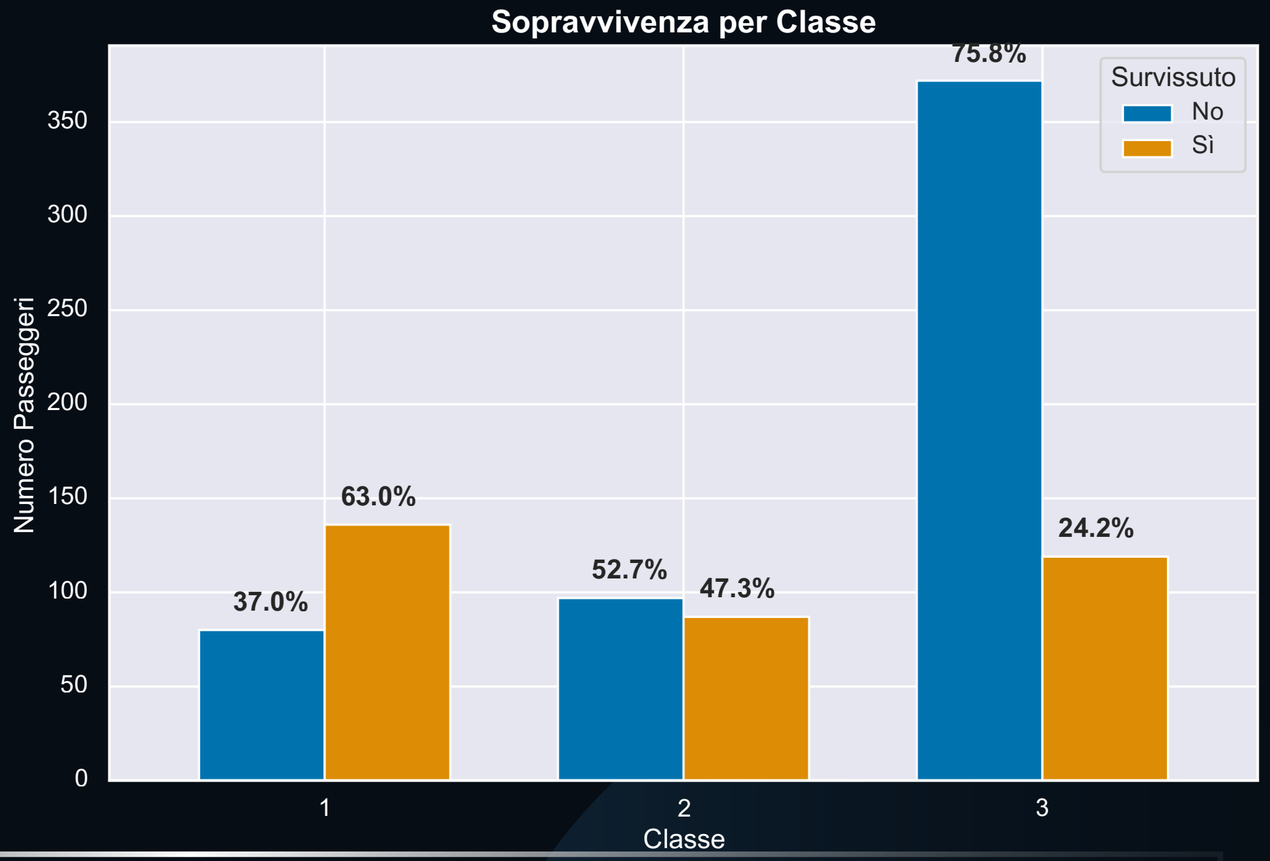




Il grafico mostra la distribuzione della variabile “Sopravvivenza” sul campione (n=891), evidenziando una netta prevalenza dei non sopravvissuti: 549 casi (61,6%) contro 342 sopravvissuti (38,4%). La distribuzione è sbilanciata, con probabilità marginali di sopravvivenza pari a 0,384 e di non sopravvivenza pari a 0,616. Non ci sono valori mancanti, e i conteggi coincidono con il totale dei record. Per approfondire, la variabile sarà analizzata per sottogruppi come sesso, classe, età e porto d’imbarco, per individuare eventuali pattern sistematici.

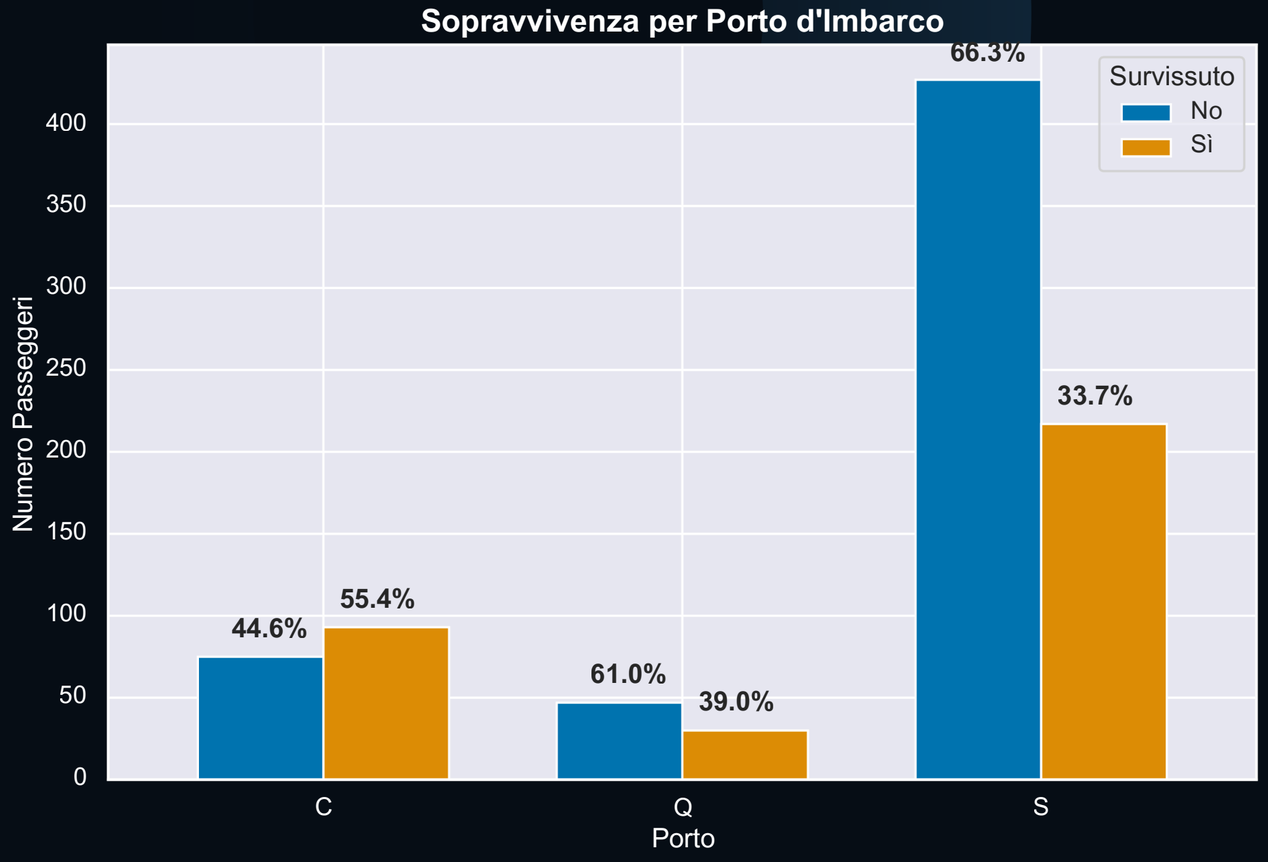
Sopravvivenza Titanic: confronto per Genere e Classe

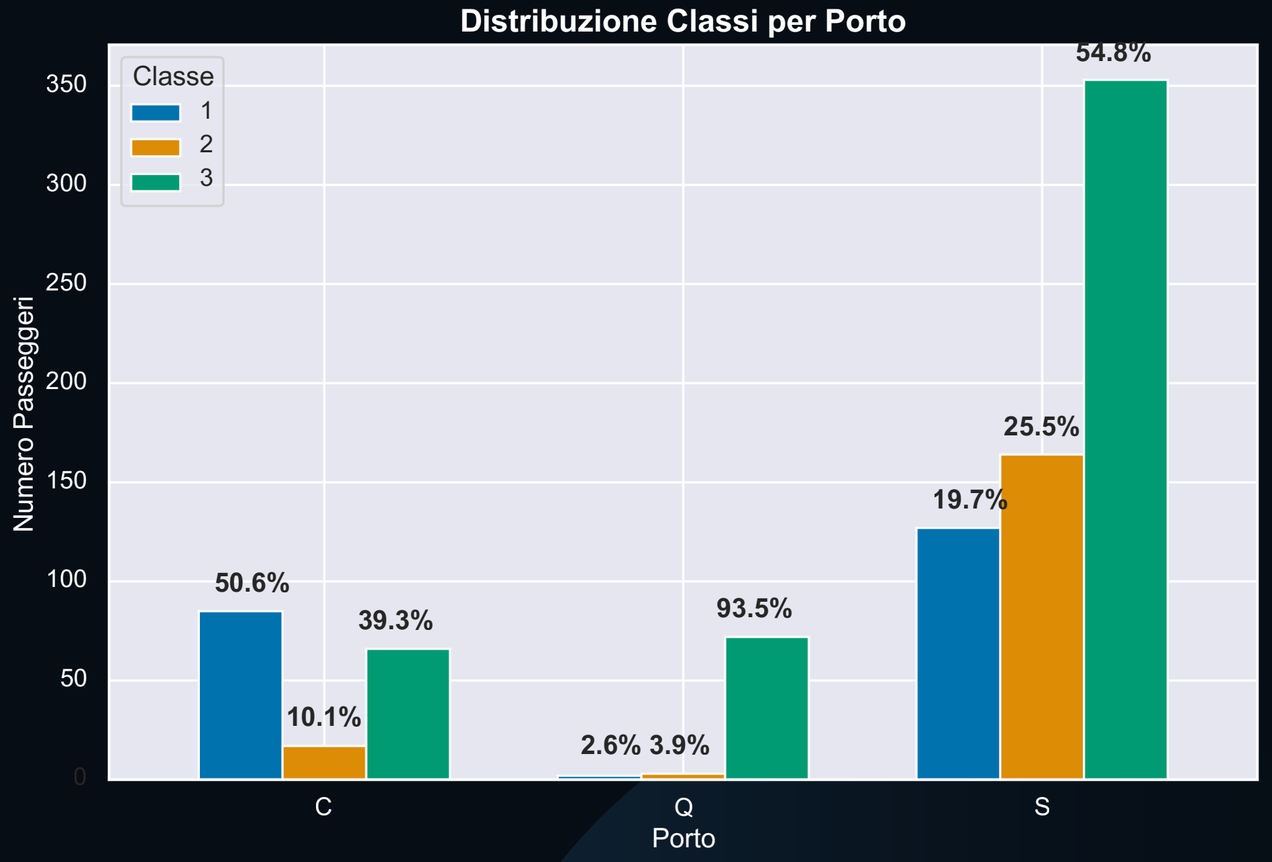




L’analisi mostra che il genere ha avuto un ruolo determinante nella sopravvivenza: il 74,2% delle donne è sopravvissuto, contro solo il 18,9% degli uomini. Anche la classe sociale ha influenzato le probabilità di salvezza: i passeggeri di prima classe hanno avuto una sopravvivenza del 63%, quelli di seconda classe circa il 47%, mentre i passeggeri di terza classe solo il 24,2%. Queste differenze riflettono non solo lo status sociale, ma anche fattori logistici della nave: i ponti più alti e l’accesso rapido alle scialuppe favorivano i passeggeri di prima classe, mentre quelli di terza classe erano in aree periferiche e più difficili da evacuare. In sintesi, donne e passeggeri delle classi superiori avevano probabilità di sopravvivere significativamente più alte rispetto a uomini e terza classe, confermando l’impatto di fattori sociali e strutturali.

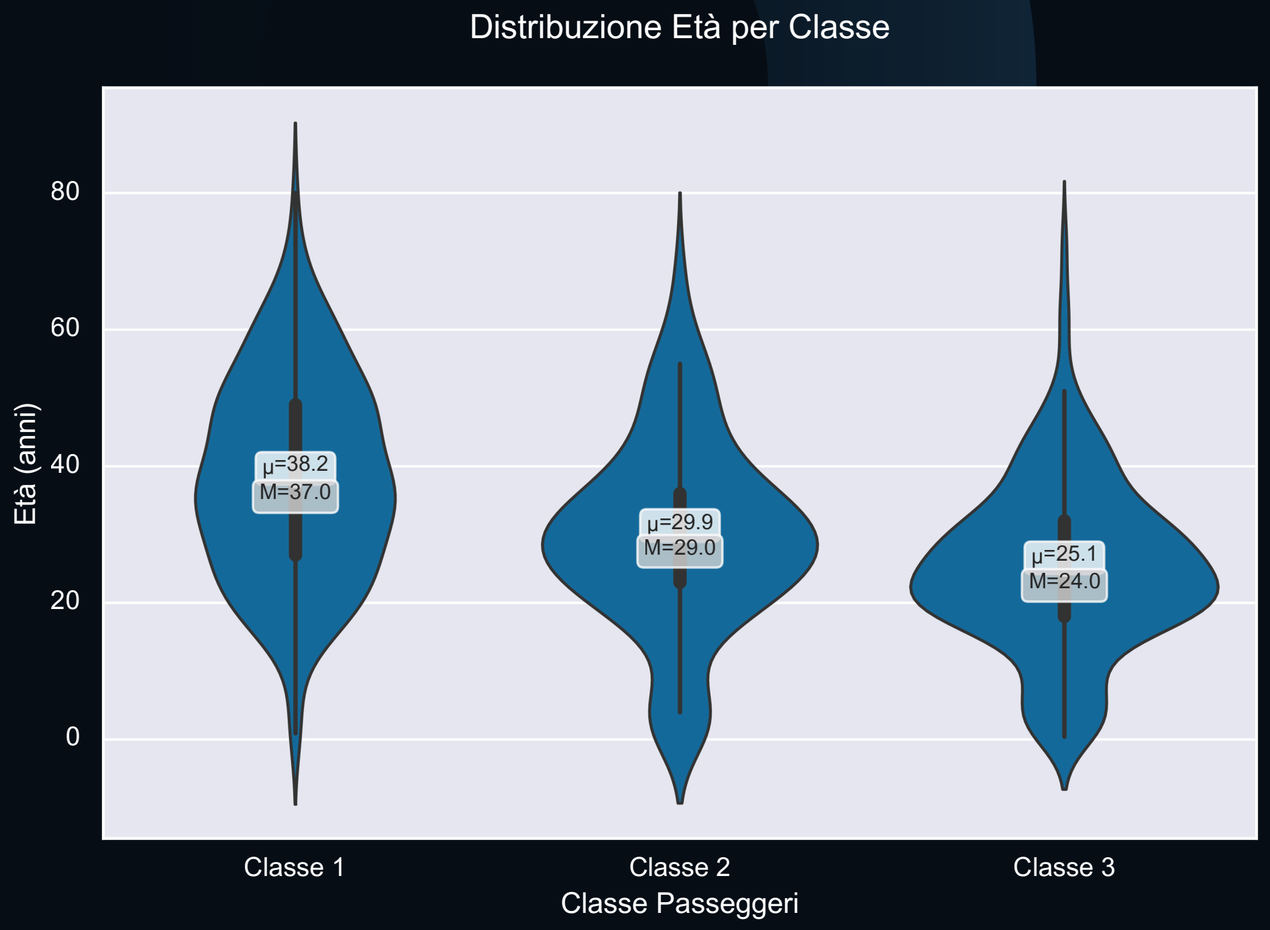
Sopravvivenza e Distribuzione Classe per Porto d'Imbarco

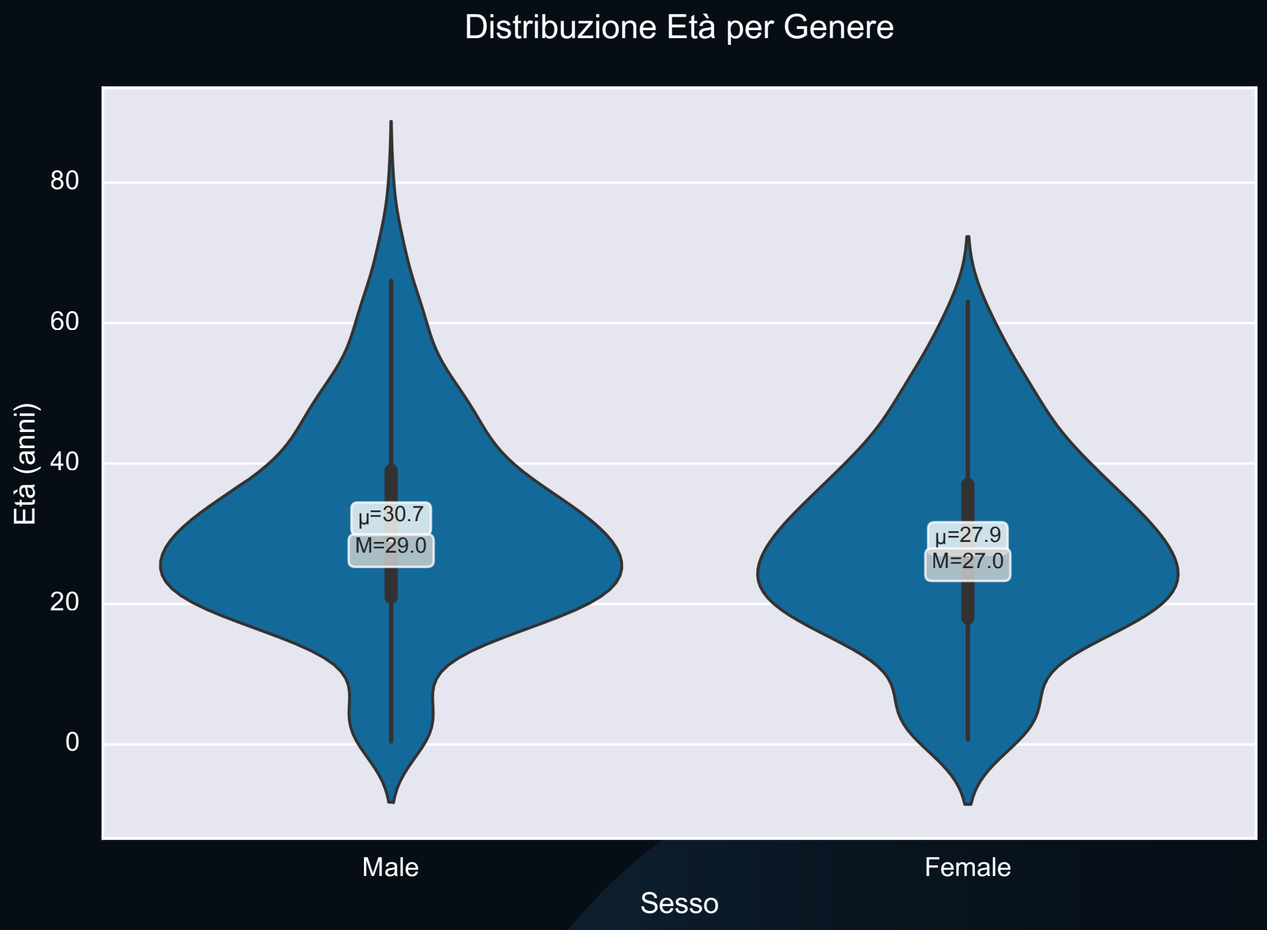




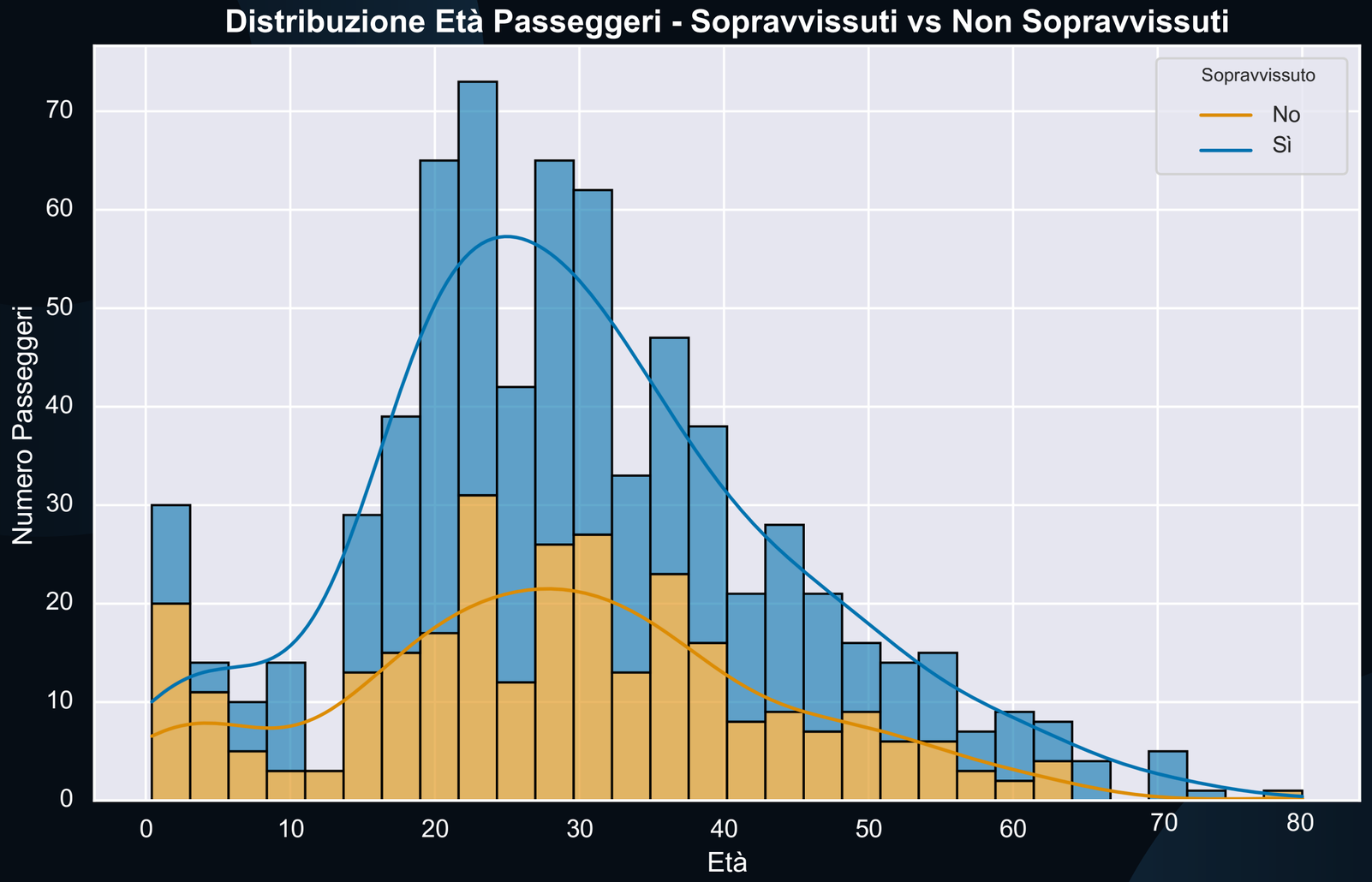
L’analisi per porto d’imbarco mostra differenze nella sopravvivenza: Cherbourg registra il tasso più alto (55,4%), seguito da Queenstown (39,0%) e Southampton (33,7%). Queste differenze riflettono la composizione sociale dei passeggeri: Cherbourg aveva molti viaggiatori di prima classe, Queenstown prevalentemente di terza classe, e Southampton una distribuzione intermedia. Il porto d’imbarco agisce quindi come indicatore indiretto dello status socio-economico, che ha influenzato significativamente le probabilità di sopravvivenza.

Distribuzione dell'Età per Categorie Socio-demografiche



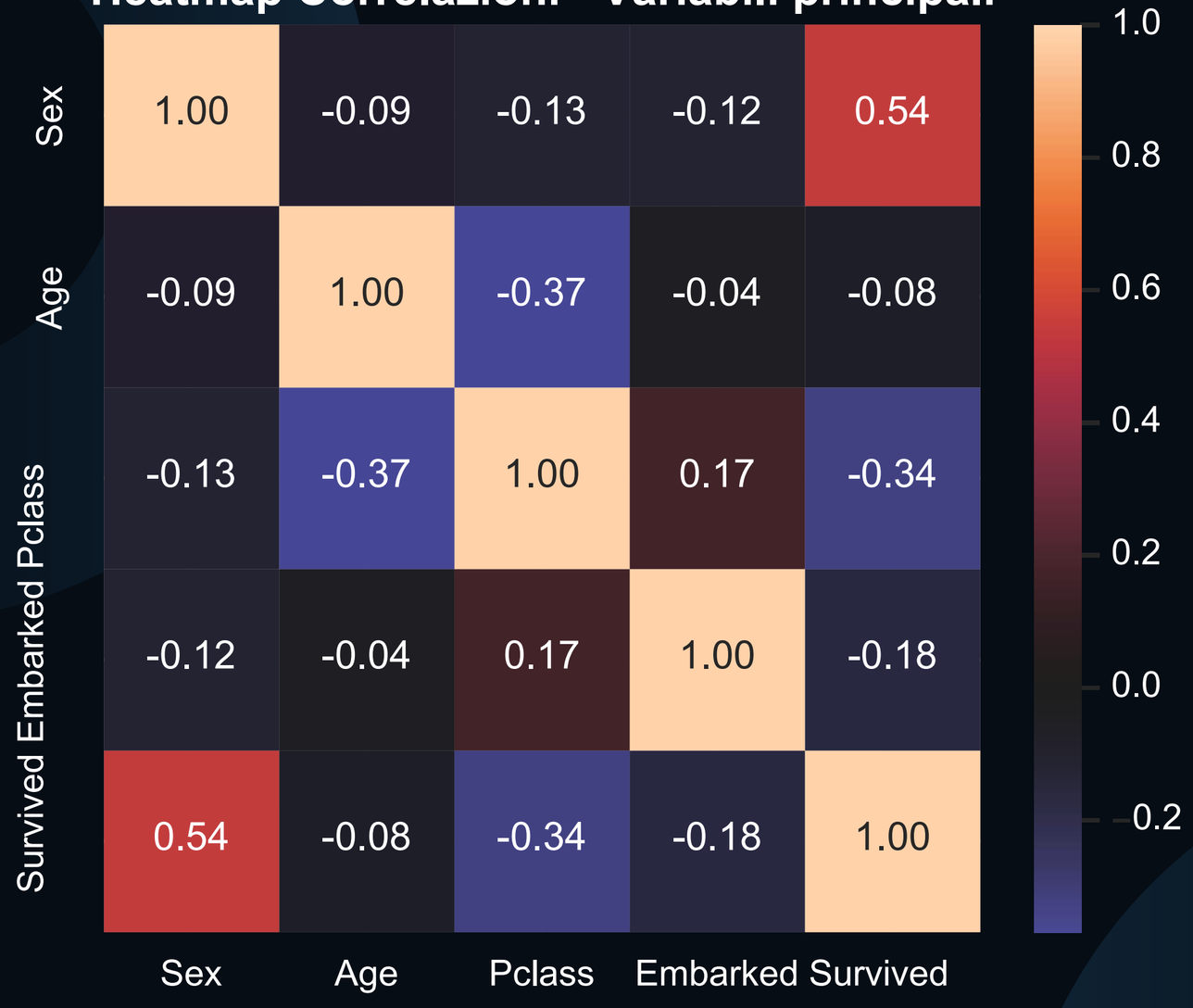


L’analisi dell’età evidenzia correlazioni con lo status socio-economico. La mediana decresce con la classe: prima classe 37 anni, seconda 29, terza 24, riflettendo la prevalenza di adulti nelle classi superiori e giovani emigranti in terza classe. Le differenze di genere sono modeste (uomini 29 anni, donne 27), ma combinando classe e genere emergono dettagli interessanti: in prima classe gli uomini sono più anziani (40 vs 35 anni) e in terza classe le donne sono particolarmente giovani (21,5 anni). La distribuzione conferma che l’età era un indicatore dello status sociale a bordo.



L’istogramma sovrapposto mostra che la maggior parte dei passeggeri aveva tra 20 e 40 anni, con un picco intorno ai 20-25 anni. I sopravvissuti sono più equilibrati nelle fasce centrali di età, mentre i non sopravvissuti predominano in quasi tutte le fasce, riflettendo il tasso di mortalità del 62%. La probabilità di sopravvivere risulta relativamente più alta tra i giovani e gli adulti di mezza età (15-45 anni).

Heatmap Correlazioni - Variabili principali



La matrice di correlazione evidenzia che il genere è il fattore più determinante per la sopravvivenza (correlazione 0,54). L'età mostra correlazioni moderate, negativa con la classe (-0,37), mentre la classe presenta una correlazione negativa con la sopravvivenza (-0,34) e positiva con il porto d'imbarco (0,17). Le altre variabili hanno correlazioni più deboli, confermando che genere e classe sociale erano i principali predittori di salvezza, con età e porto d'imbarco come fattori complementari.

CONCLUSIONI DELL'ANALISI ESPLORATIVA SUL TITANIC

L'analisi dei dati mostra che la sopravvivenza non fu casuale, ma determinata da fattori socio-demografici e strutturali.

Il genere è il fattore più influente: il 74,2% delle donne è sopravvissuto, contro il 18,9% degli uomini. Anche la classe di viaggio mostra un chiaro gradiente sociale, con tassi di sopravvivenza del 63% in prima classe, 47,3% in seconda e solo 24,2% in terza. L'età influisce in modo moderato: i giovani adulti (20-40 anni) avevano maggiori possibilità di salvarsi, mentre gli anziani erano più penalizzati. Il porto di imbarco riflette lo status socio-economico: Cherbourg registra il tasso più alto di sopravvivenza (55,4%), Southampton il più basso (33,7%).

Le differenze sociali si riflettono nella struttura demografica: la prima classe è composta principalmente da adulti maturi (mediana 37 anni), mentre la terza classe da giovani migranti (mediana 24 anni). Anche il porto di partenza accentua la disparità: da Cherbourg proviene il 50,6% della prima classe, da Queenstown il 93,5% della terza.

Per la costruzione dei modelli predittivi, le variabili principali sono state selezionate in base alla capacità discriminante: Sex fornisce il primo punto di separazione più netto, Pclass aggiunge la stratificazione sociale, Age introduce una soglia quantitativa e Embarked intercetta variazioni socio-geografiche residue.
In sintesi, genere e classe sono i predittori principali, mentre età e porto di imbarco completano il quadro, confermando l'importanza di fattori sociali e strutturali nella sopravvivenza dei passeggeri.

GRID SEARCH CV

Per valutare l’impatto delle strategie di imputazione sulla colonna ‘Age’ e sull’accuratezza del modello, sono state testate diverse soluzioni:

- Eliminazione record incompleti: semplice ma riduce il campione e può far perdere informazioni.
- Media globale: mantiene i dati ma appiattisce la distribuzione.
- Mediana: più robusta agli outlier e rappresenta meglio l’età centrale.
- Mediana per sottogruppi (Pclass + Sex): rispetta la struttura socio-demografica, fornendo imputazioni più realistiche.

Ogni metodo è stato valutato con Decision Tree + Grid Search per garantire un confronto equo.
Il confronto delle performance consente di identificare la strategia più adatta, guidando la scelta finale con basi oggettive.

RISULTATI ANALISI COMPARATIVA

Dall'analisi dei risultati emerge che:

- La semplice eliminazione dei valori nulli riduce la quantità di dati e produce la performance più bassa sul test set.
- L'imputazione con media o mediana migliora l'accuratezza, garantendo maggiore completezza dei dati.
- L'imputazione basata sulla mediana raggruppata per classe di viaggio e genere (Test 4) mantiene la performance elevata e, allo stesso tempo, rispetta la struttura socio-demografica dei passeggeri, preservando pattern importanti legati a Pclass e Sex.

Motivazione per la scelta del Test 4:
Il Test 4 viene selezionato per le fasi successive perché combina:

- Accuratezza ottimale pari agli altri approcci basati sull'imputazione.
- Maggiore coerenza con la struttura dei dati, riducendo il rischio di introdurre bias dovuti a valori medi generici.
- Capacità di preservare differenze significative tra gruppi, fondamentali per la costruzione di modelli predittivi più realistici e robusti.

IMPLEMENTATIZIONE DECISION TREE CLASSIFIER

In questa fase è stata condotta la validazione della profondità dell'albero di decisione. L'obiettivo era individuare il livello di complessità ottimale del modello per bilanciare bias e varianza.
Sono stati scelti diversi valori di profondità massima da testare, nello specifico [2, 5, 10, 25, None]. Per ciascun valore, è stato addestrato un classificatore ad albero di decisione, fissando il seme casuale a 0 per garantire la riproducibilità dei risultati. Ogni modello è stato valutato sul validation set e le accuratezze ottenute sono state registrate.

Dall'analisi è emerso che la profondità ottimale corrisponde a 5, con un'accuratezza sul validation set pari a 0.8263. Sulla base di questo risultato, è stato addestrato un modello finale sfruttando l'intero training set, cioè unendo training set e validation set, per massimizzare i dati disponibili.

Il modello così costruito è stato infine valutato sul test set, ottenendo un'accuratezza di 0.8117. Questo valore rappresenta una stima affidabile della capacità del modello di generalizzare su dati non visti.

In conclusione, la scelta di limitare la profondità dell'albero a 5 ha permesso di ottenere un modello stabile ed efficace. L'accuratezza raggiunta conferma la bontà della configurazione selezionata, rendendo l'albero di decisione uno strumento adeguato per predire la sopravvivenza dei passeggeri del Titanic.

ACCURATEZZA FINALE DEL MODELLO

81.17%

Questa è la precisione del nostro albero decisionale nel prevedere la sopravvivenza dei passeggeri sul set di dati di test, dimostrando l'efficacia del modello.

Notebook

https://drive.google.com/file/d/13Dc6thlK_H10X7wN7mQ-eWg_f_tOZ-GL/view?usp=sharing

THANK YOU