MODELLO PREDITTIVO DATABASE TITANIC

Giuseppe Russo

Consegna del progetto

L'obiettivo del progetto del modulo di Machine Learning Base è **creare un modello in grado di prevedere se un passeggero del Titanic sarebbe sopravvissuto o meno al tragico naufragio**, in base a diverse caratteristiche e variabili disponibili nel famoso dataset Titanic.

Le colonne di questo dataset, che corrispondono alle informazioni dei passeggeri, sono: genere, età, classe di viaggio, luogo di imbarco e sopravvivenza o meno.

Schema – Il processo di Machine Learning

1. Caricamento e Analisi dei dati
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 che verranno mostrate in seguito.

2. Suddivisione del dataset
Successivamente, il dataset è stato suddiviso in train e test set per garantire una valutazione affidabile del modello. Il train test poi ulteriormente diviso in un nuovo train e in validation set.

3. Gestione dei valori mancanti nella feature «Age»
Fase cruciale. Riscontrati questi valori mancanti, per la variabile Age sono state confrontate diverse strategie: eliminazione dei record con valore mancante, imputazione con media, mediana e con valore arbitrario. Queste strategie sono state valutate attraverso la Logistic Regressione per definire una baseline.

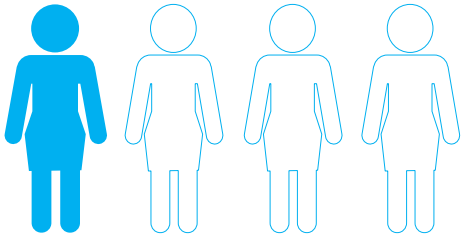
Schema – Il processo di Machine Learning

4. Sostituzione dei valori mancanti
In base ai risultati della Logistic Regression sono stati sostituiti i valori mancanti nella colonna con il valore che corrisponde all'età media nel dataset.

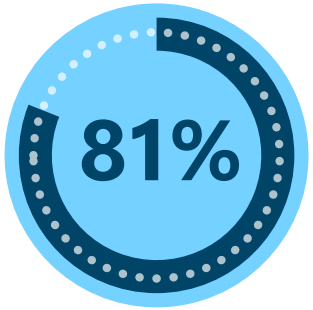
5. Encoding variabili categoriche con One-Hot
Le variabili categoriche sono state poi codificate tramite One-Hot Encoding, rendendo il dataset pronto per l'addestramento del modello. Questo passaggio ha garantito che tutte le informazioni rilevanti fossero utilizzabili in modo coerente dall'algoritmo.

6. Implementazione Decision Tree Classifier
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 sulle capacità del modello di predire la sopravvivenza dei passeggeri.

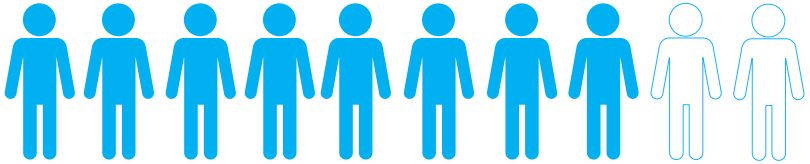
Rapida analisi del Dataset Titanic: genere



1 donna su **4**



8 uomini su **10**



Una prima analisi sui deceduti

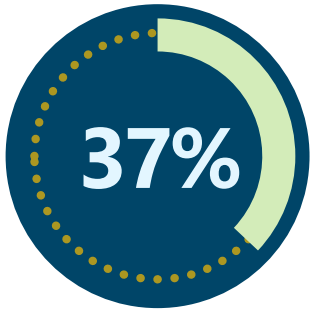
Dal conteggio delle singole istanze si può notare come sia altissima la percentuale sul totale maschile degli uomini deceduti.

Più clemente è stata la tragedia ai danni delle donne, che comunque erano in quantità inferiore sulla nave rispetto agli uomini: 314 donne e 577 uomini.

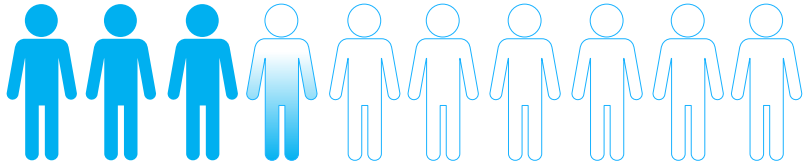
Il totale

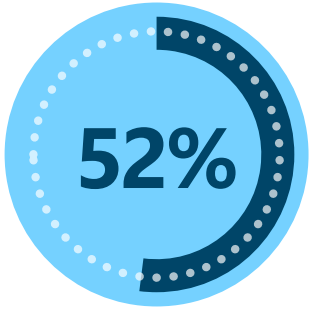
Rapportando le percentuali ai numeri assoluti, la quantità di persone decedute supera il 60%. 3 persone su 5.

Rapida analisi del Dataset Titanic: classe di viaggio

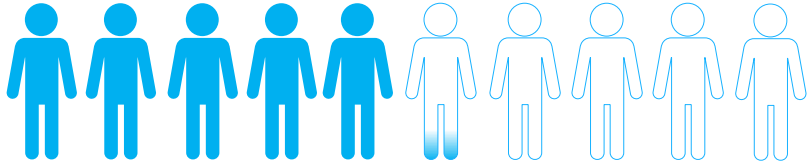


Prima classe: Quasi **2** persone su **5**

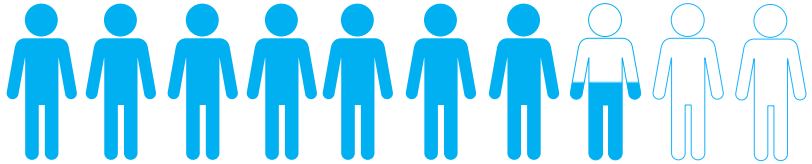




Seconda classe: Poco più di **1** persona su **2**



Terza classe: 3 persone su **4**

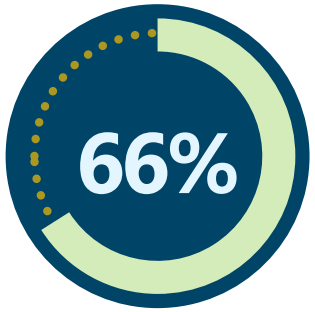


Analisi sui deceduti

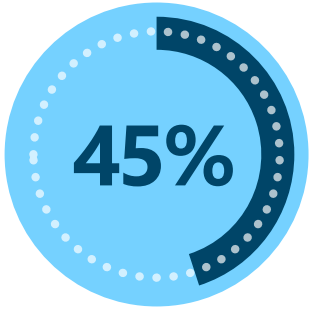
Dal conteggio delle singole istanze si può notare come ci sia una differenza nella percentuale di deceduti per ogni classe di viaggio.

È netta la progressione del numero di deceduti all'abbassarsi della classe, rendendo chiaro come lo stato sociale abbia influito sulla sopravvivenza dell'individuo.

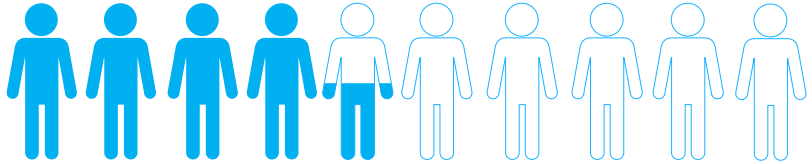
Rapida analisi del Dataset Titanic: imbarco



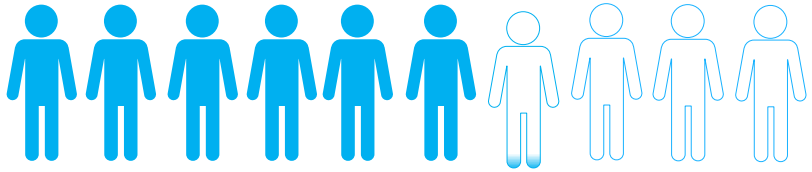
Da Southampton: 2 persone su **3**



Da Cherbourg: Poco meno di **1** persona su **2**



Da Queenstown: Poco più di **3** persone su **5**



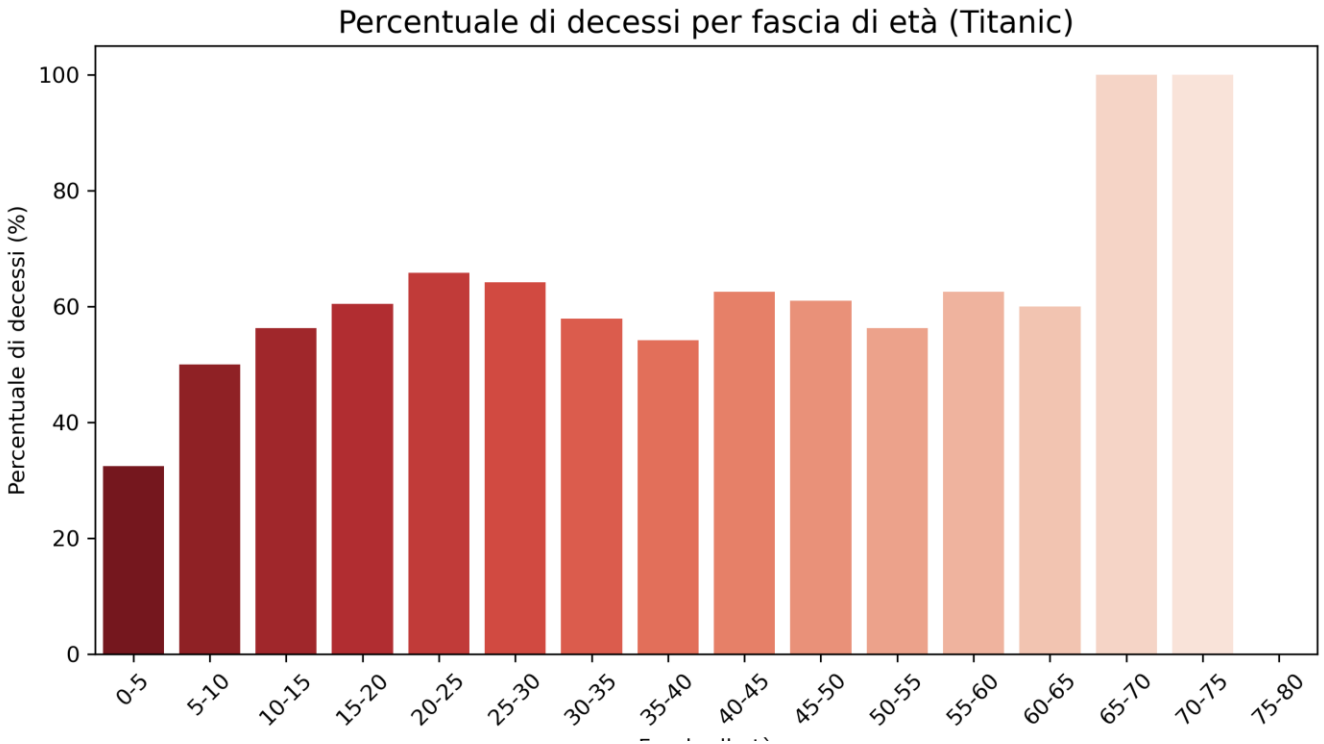
Analisi sui deceduti

Il Titanic ha fatto salire i passeggeri da tre attracchi differenti: la partenza Southampton, Cherbourg in Francia e Queenstown in Irlanda.

La stragrande maggioranza delle persone è salita alla partenza a Sothampton.

Si può notare come da Cherbourg la percentuale di deceduti sia più bassa, forse per motivi logistici.

Rapida analisi del Dataset Titanic: fasce d'età



Analisi sui deceduti

Da questo istogramma appare chiaro come l'età incida sui decessi solo agli estremi:

Solo i bambini sotto ai 5 anni hanno una percentuale di decesso più bassa degli altri gruppi, con un campione di 40 individui.

Gli over 65 anni hanno il 100% di decesso, ma con un campione molto ridotto, che rende decisamente meno affidabile questo numero.

Esclusi questi gruppi, tutte le altre fasce hanno una percentuale di decessi tra il 50% e il 65%, con delle variazioni non progressive in base all'età, ma alternate.

Svolgimento e specifiche

Quindi, analizzati questi dati, come si può prevedere se un singolo passeggero con le proprie istanze sia sopravvissuto o meno?

Quali sono le combinazioni di questi dati che portano ad una previsione affidabile?

La consegna chiede di addestrare un modello, precisamente un modello di albero di decisione, con il 75% degli individui di questo database.

Successivamente, di separare ulteriormente il training set in due sottoinsiemi, sempre secondo la proporzione 75-25. Il secondo sottoinsieme deve essere il validation set.

Il notebook

Il mio notebook è stato scritto ed è a questo link:

https://drive.google.com/file/d/1AediqwIV1Tp-TBC49WnSVOEjMGkESJT-/view?usp=sharing

Gestione dei valori mancanti nella feature «Age»

La Logistic Regression è stata utilizzata su diverse strategie per valutare quale impatto, modificando il dataset, ognuna di esse avesse sull'accuratezza del modello .

	Accuracy
Eliminazione righe	0.793
Imputazione con media	0.794
Imputazione con mediana	0.793
Imputazione con -1	0.793

Risultati praticamente uguali, ad eccezion fatta della media che viene scelta come soluzione per riempire i dati mancanti nella colonna «Age», anche per coerenza della struttura del dataset.

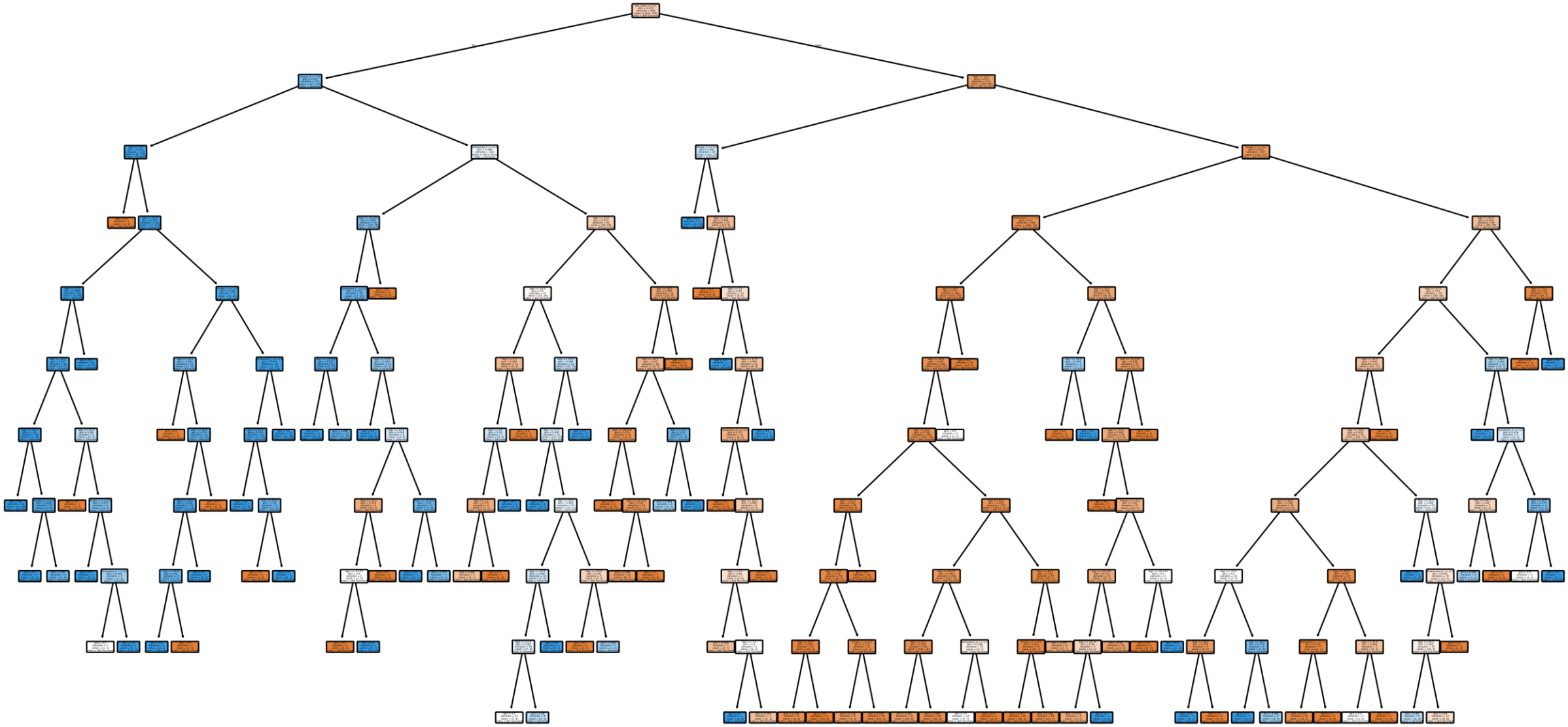
La profondità dell'albero

Il validation set ha lo scopo di validare la sua profondità, provando i seguenti valori: 2, 5, 10, 25, profondità max, mentre la metrica di valutazione deve essere l'accuratezza.

	Accuracy
Profondità 2	0.7904
Profondità 5	0.8024
Profondità 10	0.8144
Profondità 25	0.7964
Profondità max	0.7964

Scelta la profondità 10, l'accuratezza sul test set è di 0.7892.

L’albero decisionale



Conclusione

Infine, l'accuratezza per la profondità 10 sul test set è inferiore a quella del validation set. Questo modello non garantisce una precisione estrema.
Con ulteriori informazioni sarebbe possibile avere un albero più profondo con meno possibilità di andare in overfitting.

Grazie per l'attenzione!