Analisi sui fattori di sopravvivenza dei passeggeri del Titanic

Loris Umberto Pennucci
863003783

Misael Monaco
863002365

***Sommario*—Questo studio analizza i fattori che hanno influenzato la sopravvivenza dei passeggeri del Titanic utilizzando tecniche di machine learning e analisi dei dati. Attraverso l'analisi esplorativa del dataset Kaggle, sono stati identificati i principali predittori di sopravvivenza: sesso, classe sociale, età, dimensione della famiglia e posizione della cabina. Il modello di regressione logistica ottimizzato ha raggiunto un'accuratezza del 78% nel predire la sopravvivenza; Dal risultato delle analizi si può evidenziare il costrutto sociale di mettere in salvo prima donne e bambini, nonchè delle disparità socio-economiche nell'accesso alle scialuppe di salvataggio.**

I. Introduzione

Il naufragio del RMS Titanic rappresenta una delle tragedie marittime più studiate della storia moderna. Nel presente lavoro, analizziamo i fattori che hanno determinato la sopravvivenza dei passeggeri, specialmente l'effetto delle norme sociali dell'epoca. Il disastro della fregata Birkenhead[1], nel 1845, popolarizzò la pratica di "prima donne e bambini" nell'evenenienza di disastri nautici, rendendola un de-facto protocollo inufficciale utilizzato nelle marine civili delle principali potenze marittime. In quanto una mera formalità, tale pratica fu fortemente influenzata anche da fattori esterni, come, nel caso del Titanic, dalla situazione socio-economica dei passeggeri.

Il dataset utilizzato è stato fornito da Kaggle nell'ambito di una competizione[2] allo scopo di addestrare un modello di Machine Learning in grado di predirre la sopravvienza dei passeggeri. Il dataset è suddiviso in due parti: un set di training contenente 891 osservazioni con la variabile target `Survived`, e un set di test con 418 osservazioni prive della variabile target, utilizzato per la valutazione finale del modello predittivo.

A. Descrizione delle Features

Le caratteristiche del dataset includono:

- **PassengerId**: Identificativo univoco
- **Survived**: Variabile binaria target (0=morto, 1=sopravvissuto)
- **Pclass**: Classe del biglietto (1=prima, 2=seconda, 3=terza)
- **Name**: Nome completo e titolo del passeggero
- **Sex**: Sesso (male/female)
- **Age**: Età in anni
- **SibSp**: Numero di fratelli/coniugi a bordo
- **Parch**: Numero di genitori/figli a bordo
- **Ticket**: Numero del biglietto
- **Fare**: Prezzo del biglietto
- **Cabin**: Numero della cabina
- **Embarked**: Porto d'imbarco (C=Cherbourg, Q=Queenstown, S=Southampton)

[1]https://wshc.org.uk/the-story-of-the-birkenhead
[2]https://www.kaggle.com/competitions/titanic

II. Analisi Esplorativa dei Dati

Prima di procedere al preprocessing, è stata condotta un'analisi esplorativa per identificare pattern e correlazioni nei dati. Si noti, per quanto il dataset di testing non si possa utilizzare per l'addestramento del modello, può comunque risultare utile per fornire un'immagine più chiara della popolazione collettiva. Naturalmente, calcoli riguardo i tassi di sopravvivenza non includono i passeggeeri il cui esito è ignoto.

A. Analisi sesso, classe ed età

Figura 1: Distribuzioni delle variabili principali: sesso, classe ed età dei passeggeri

L'analisi rivela una prevalenza maschile, con la terza classe che rappresenta quasi metà dei passeggeri. La distribuzione dell'età mostra una concentrazione tra i 20-40 anni, suggerendo che i passeggeri maschili di terza classe in questa fascia d'età costituiscano una componente significativa della popolazione.

Per evidenziare meglio tale relazione, mostriamo il rapporto tra le tre proprietà

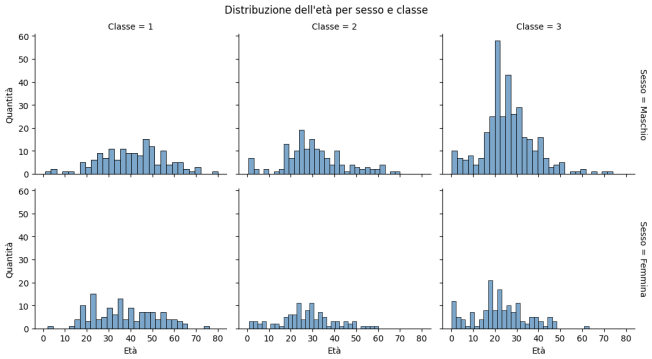


Figura 2: Distribuzione della popolazione in base al sesso, classe ed età

B. Analisi della Sopravvivenza

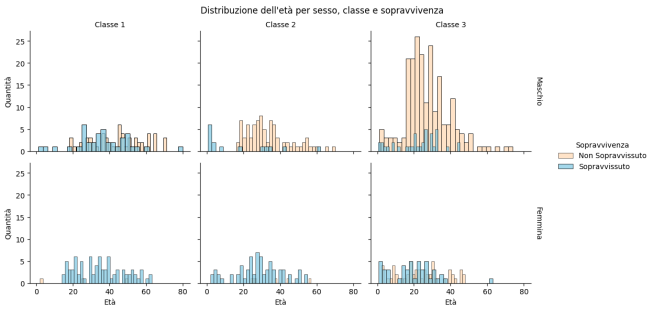


Figura 3: Superstiti in base a classe, sesso ed età

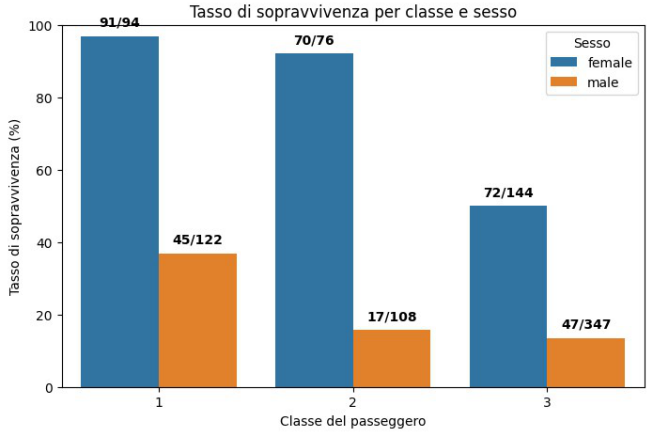


Figura 4: Tasso di sopravvivenza per classe e sesso

I risultati evidenziano disparità drammatiche nei tassi di sopravvivenza:

- Donne prima classe: 96%
- Donne seconda classe: 92%
- Donne terza classe: 50%
- Uomini prima classe: 36%
- Uomini seconda classe: 15%
- Uomini terza classe: 13%

Emerge chiaramente che la popolazione femminile presenta un tasso di sopravvivenza maggiore indipendentemente dalla classe, mentre la popolazione maschile di terza classe ha il tasso più basso. In base all'età, invece, si nota una maggiore sopravvivenza per passeggeri di giovane età, particolarmente all'interno della popolazione maschile di seconda classe. Si può notare come, nonostante la pratica del "prima donne e bambini", il tasso di sopravvivenza delle donne di terza classe sia virtualmente dimezzato rispetto le altre, dimostrando come anche la situazione socio-economica dei passeggeri abbia avuto incidenza sulla disponibilità di scialuppe.

C. Analisi del Prezzo di Biglietto

L'analisi della feature `Fare` rivela, come prevedibile, una forte correlazione con la classe di appartenenza.

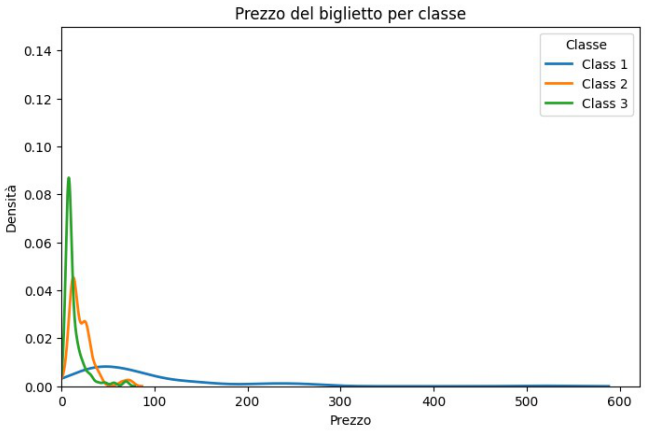


Figura 5: Distribuzione del prezzo del biglietto per classe

La terza classe presenta un picco negli intorni di 5-10 unità monetarie, mentre i biglietti di prima e seconda classe mostrano prezzi progressivamente maggiori.

D. Analisi della Dimensione della Famiglia

È stata introdotta una nuova feature `FamilySize` per analizzare l'impatto della dimensione familiare, definita come la somma di `SibSp` e `ParCh` più uno, in modo da includere anche il passeggero analizzato.

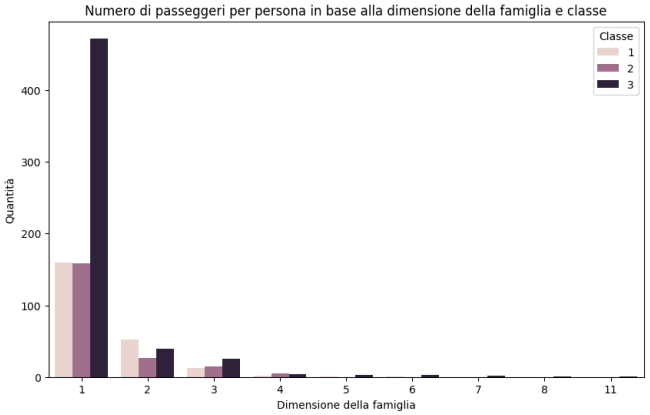


Figura 6: Relazione tra dimensione famiglia e classe

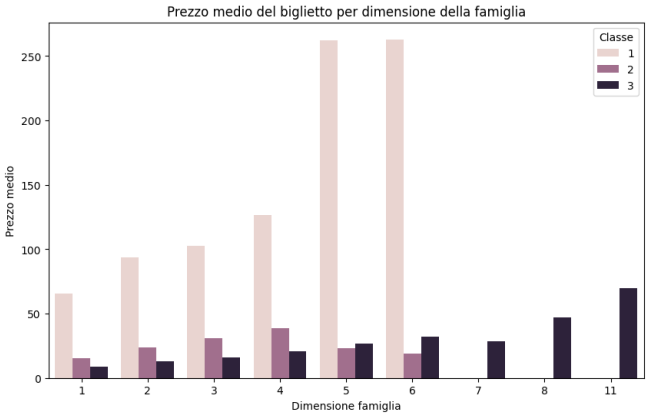


Figura 7: Relazione tra dimensione famiglia, classe prezzo

Come prevedibile, il prezzo medio di un biglietto di prima classe è maggiore rispetto alle altre due classi, ed il prezzo del biglietto è proporzionale alla dimensione della famiglia. Vista la correlazione tra tasso di sopravvivenza e classe, è probabile che anche il campo `FamilySize` possa essere correlato con la sopravvivenza.

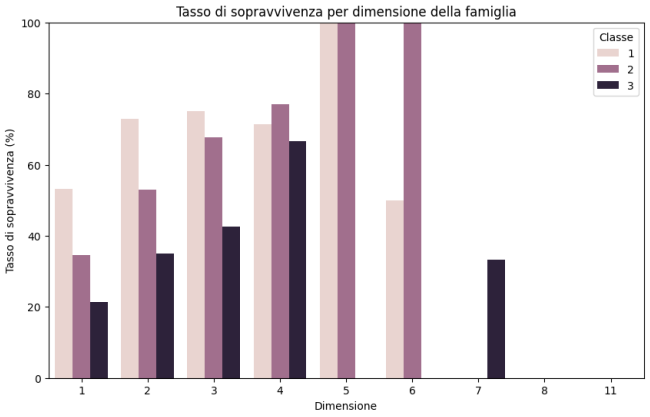


Figura 8: Tasso di sopravvivenza per dimensione di famiglia

Si noti che, all'aumentare della dimensione di una famiglia, è molto probabile aumenti il numero di donne e bambini: visti i grafici precedenti, queste due categorie hanno un tasso di sopravvivenza maggiore rispetto ai maschi, quindi, in una famiglia popolosa, il numero di donne e bambini aumenteranno la percentuale di superstiti. In altri termini, la dimnesione della famiglia non è direttamente responsabile del maggior tasso di sopravvivenza.

E. Analisi del Porto di Imbarco

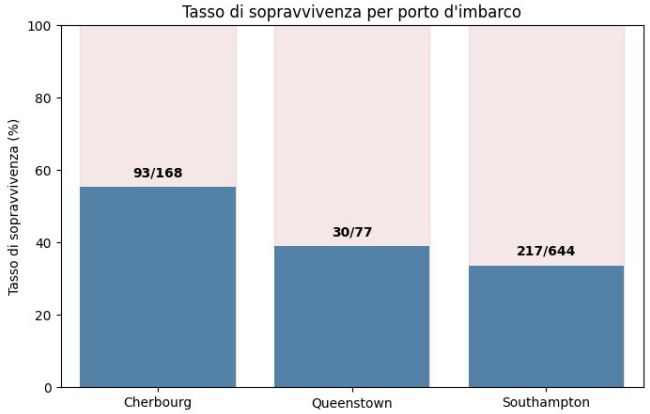


Figura 9: Distribuzione dei passeggeri per porto di imbarco e classe

La maggior parte dei passeggeri si è imbarcata a Southampton, mentre circa la metà dei passeggeri di prima classe si è imbarcata a Cherbourg. Questo porto presenterà il tasso di sopravvivenza più elevato, non per fattori geografici, ma perchè "composto" principalmente da passeggeri di prima classe.

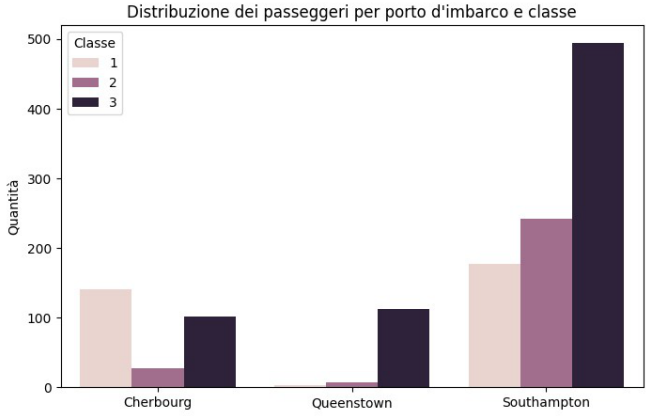


Figura 10: Tasso di sopravvivenza per porto di imbarco

F. Analisi delle Cabine e dei Ponti

L'analisi delle cabine risulta complessa per l'elevato numero di valori mancanti (77%). Secondo la documentazione storica[3], il Titanic presentava 8 ponti sopra la linea di

[3]https://www.encyclopedia-titanica.org/titanic-deckplans

galleggiamento (A-G), più un ponte che ospitava le scialuppe di salvataggio, comunemente definito come ”Boad Deck”.

La distribuzione delle cabine per classe era:

- Ponti A-B-C: esclusivamente prima classe
- Ponti D-E: tutte le classi
- Ponti F-G: seconda e terza classe
- Boat Deck: 6 cabine di prima classe (T,U,W,X,Y,Z)

Le cabine erano numerate in base al ponte della nave[4], quindi in base al valore di Cabin possiamo ricavare su quale ponte si trovasse il passeggero. Si noti che, durante il viaggio, solo la cabina T era occupata sul Boad Deck; per semplificare il modello, il passeggero appartenente a quella cabina verrà considerato residente del ponte A, vista la sua prossimità.

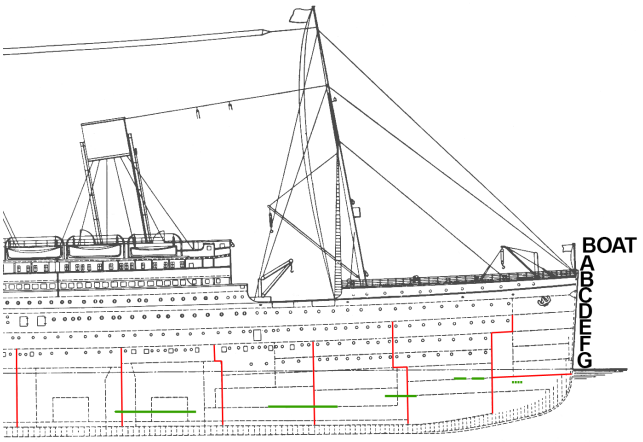


Figura 11: Ponti del Titanic

Visto che non ci è possibile ricavare la cabina (e conseguentemente il ponte) di tutti i passeggeri, inseriremo tutti i passeggeri la cui cabina è ignota in un fittizio ponte ”X”.

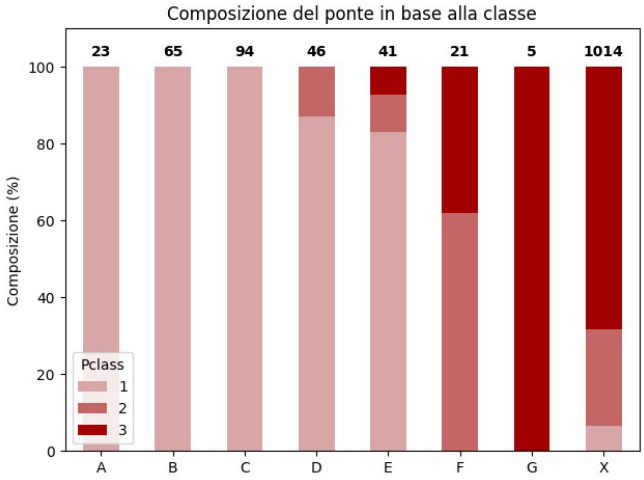


Figura 12: Composizione dei vari ponti

[4]https://www.ggarchives.com/OceanTravel/Titanic/01-PlanningBuildingLaunching/Decks-ComprehensiveDetails.html

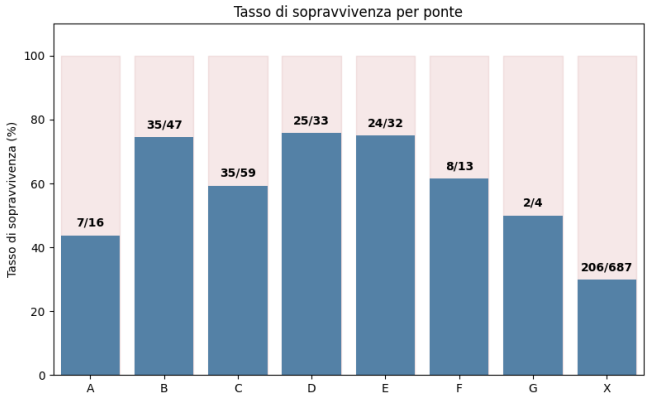


Figura 13: Tasso di sopravvivenza per ponte della nave

Si noti che il tasso di sopravvivenza, in quanto percentuale, risulta essere misleading: il ponte G, nonostante composto da soli passeggeri di terza classe, presenta una sopravvivenza maggiore rispetto il ponte A, composto da passeggeri di prima classe. Ciò è dovuto al ridotto numero di passeggeri di cui abbiamo la certezza fossero situati sul ponte G. Se andiamo ad analizzare un ponte composto da più classi, noteremo che la terza classe presenta lo stesso un survival rate minore delle altre classi. In seguito, è rappresentato il ponte F.

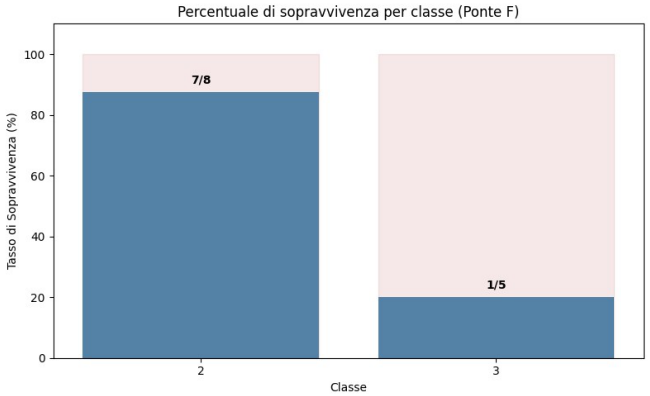


Figura 14: Dettaglio sui superstiti del ponte F

Il ponte X è quello con tasso di sopravvivenza minore rispetto agli altri. Visto che è occupato da passeggeri di tutte le classi, è opportuno verificare il tasso di sopravvivenza di ognuna.

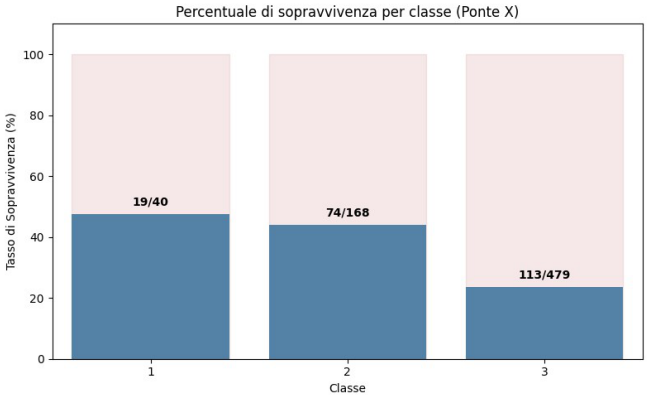


Figura 15: Dettaglio sui superstiti del ponte X

In linea con ciò che abbiamo verificato in precedenza, i passeggeri di terza classe presentano sempre un tasso di sopravvivenza minore rispetto le altre.

G. Analisi dei Titoli

Per quanto sia evidente che il nome di un passeggero non sia influente sulla sua sopravvivenza, possiamo estrarre il titolo di ogni passeggero, tramite un'espressione regex, fornendoci un indicatore sullo stato sociale dei vari passeggeri tramite il titolo

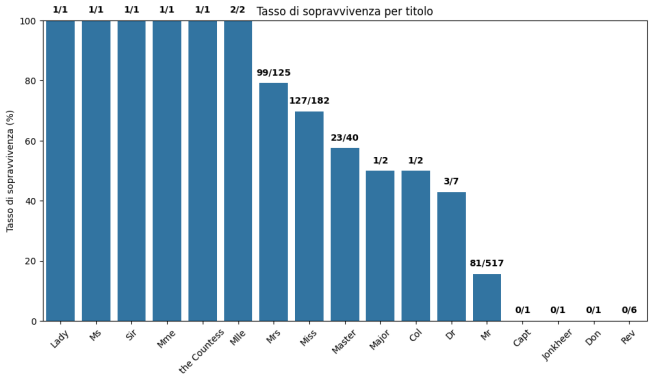


Figura 16: Tasso di sopravvivenza per titolo

L'analisi rivela che titoli nobiliari e femminili presentano tassi di sopravvivenza superiori, confermando i pattern identificati nelle analisi precedenti.

III. Preprocessing dei Dati

A. Gestione dei Valori Mancanti

Trascurando i valori mancanti di `Survived` nel dataset di test, si procede alla gestione degli altri parametri mancanti.

1) Porto d'Imbarco: Due passeggere, Amelie Icard e Martha Evelyn Stone, presentano valori mancanti per `Embarked`. Le due condividevano cabina, identificativo e prezzo del biglietto, suggerendo che viaggassero insieme. Consultando archivi storici[5], risulta che Amelie Icard fosse la domestica di Martha Stone, entrambe imbarcate al porto di Southampton:

2) Età: Per la gestione delle età mancanti, in linea con quello che abbiamo riscontrato nell'analisi precedente, bisogna prestare attenzione alla classe ed al sesso del passeggero. Assegneremo a tutti i passeggeri con età mancante l'età mediana in base al suo sesso e classe.

```
median_ages = df.groupby(['Sex', 'Pclass'])['Age'].median
    ()
df['Age'] = df.apply(
    lambda row: median_ages[row['Sex'], row['Pclass']]
    if pd.isna(row['Age']) else row['Age'], axis=1
)
```

Listing 1: Imputazione età

3) Prezzo del Biglietto: Un singolo passeggero ha il valore mancante, ovvero Thomas Storey. Egli viaggiava in terza classe senza nessun parente. Per assegnargli il fare più opportuno, possiamo ricavare il prezzo mediano di tutti i maschi nella sue stesse circostanze, ovvero passeggeri di terza classe solitari. Questo valore corrisponde a 7.85£; secondo archivi storici[6], il prezzo effettivo del suo biglietto era di 7.2£, sufficentemente vicino alla nostra stima.

B. Preprocessing Finale

Le feature numeriche sono state discretizzate tramite `bin` di dimensioni uguali. Soltanto il campo `FamilySize` è stato discretizzato manualmente.

```
le = LabelEncoder()

# Binning delle feature numeriche
df['Fare'] = pd.qcut(df['Fare'], 10)
df['Age'] = pd.qcut(df['Age'], 10)

# Encoding
df['Fare'] = le.fit_transform(df['Fare'].astype(str))
df['Age'] = le.fit_transform(df['Age'].astype(str))

# Categorizzazione famiglia
def bin_family_size(size):
    if size == 1: return 0      # solo
    elif 2 <= size <= 4: return 1  # piccola
    elif 5 <= size <= 8: return 2  # media
    else: return 3              # grande

df['FamilySize'] = df['FamilySize'].apply(bin_family_size)
df = pd.get_dummies(df, columns=['FamilySize'], prefix='
    FamilySize')
```

Listing 2: Encoding e discretizzazione

IV. Addestramento e Valutazione dei Modelli

A. Matrice di Correlazione

La matrice di correlazione fornisce una visione d'insieme delle relazioni lineari tra le variabili numeriche del dataset. Questo ci permette di identificare le feature migliori da utilizzare durante la fase di addestramento.

[5]https://www.encyclopedia-titanica.org/titanic-survivor/martha-evelyn-stone.html

[6]https://www.encyclopedia-titanica.org/titanic-victim/thomas-storey.html

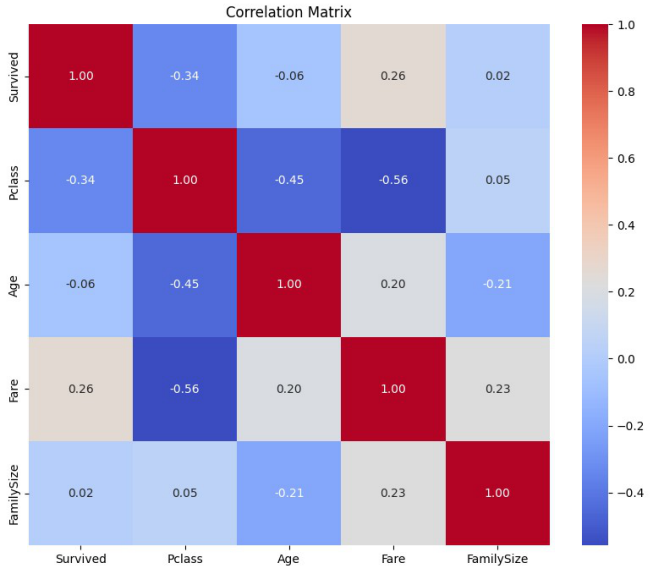


Figura 17: Matrice di Correlazione tra le variabili del dataset

In linea con quello che abbbiamo osservato finora, il parametro `Survived` è fortemente correlato a `pClass` e `Fare`.

B. Confronto tra Algoritmi

Sono stati testati diversi algoritmi di classificazione per identificare quello con le migliori performance:

```
models = {
    'Random Forest': RandomForestClassifier(random_state
        =42),
    'Logistic Regression': LogisticRegression(max_iter
        =1000, random_state=42),
    'Support Vector Machine': SVC(random_state=42),
    'Decision Tree': DecisionTreeClassifier(random_state
        =42),
    'K-Nearest Neighbors': KNeighborsClassifier(),
    'Gradient Boosting': GradientBoostingClassifier(
        random_state=42),
    'XGBoost': XGBClassifier(eval_metric='logloss',
        random_state=42)
}

for name, model in models.items():
    print(f"\n{name}")
    model.fit(X_train, Y_train)
    y_pred = model.predict(X_test)
    print("Confusion Matrix:")
    print(confusion_matrix(Y_test, y_pred))
    print("Classification Report:")
    print(classification_report(Y_test, y_pred))
```

Listing 3: Addestramento modelli multipli

C. Risultati dei Modelli

Il confronto tra i diversi algoritmi ha evidenziato che la Regressione Logistica presenta una precisione marginalmente superiore (82%) rispetto agli altri modelli. Questo risultato è stato ulteriormente migliorato tramite ottimizzazione degli iperparametri.

Di seguito sono rappresentate le `Learning Curve` di ogni modello.

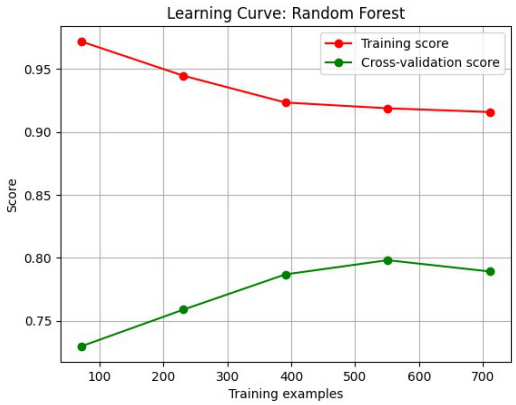


Figura 18: Learning Curve Random Forest

Figura 19: Learning Curve Logistic Regression

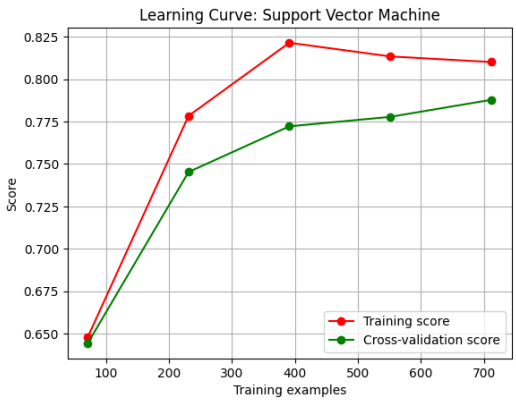


Figura 20: Learning Curve SVM

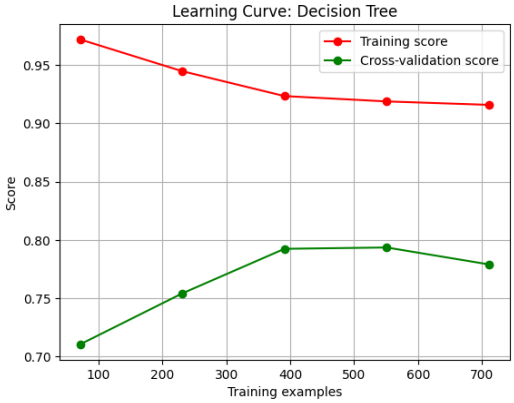


Figura 21: Learning Curve Decision Tree

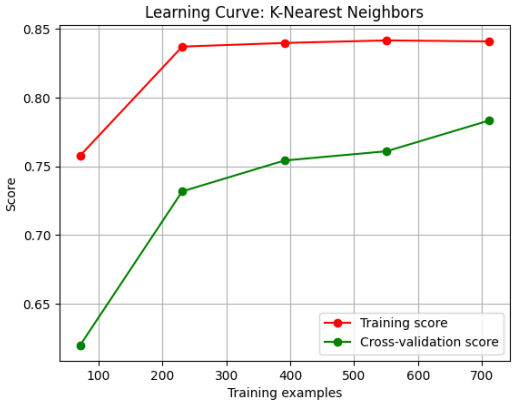


Figura 22: Learning Curve K-Nearest Neighbors

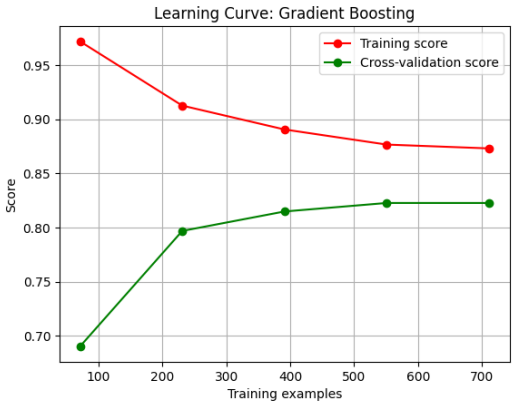


Figura 23: Learning Curve Gradient Boosting

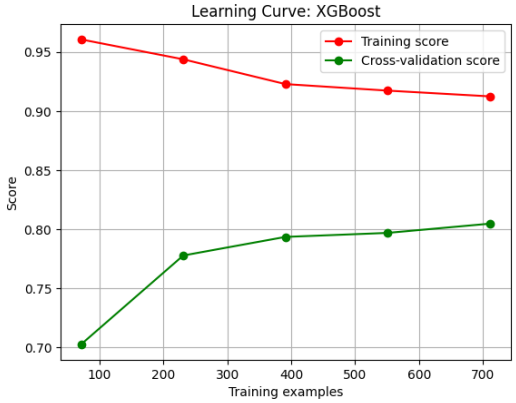


Figura 24: Learning Curve XGBoost

D. Ottimizzazione degli Iperparametri

Per migliorare ulteriormente le performance del modello di Regressione Logistica, è stata applicata l'ottimizzazione bayesiana degli iperparametri:

```
search_space = {
    'C': (1e-3, 100.0, 'log-uniform'),
    'penalty': ['l1', 'l2'],
    'solver': ['liblinear', 'saga']
}

opt = BayesSearchCV(
    estimator=LogisticRegression(max_iter=1000,
        random_state=42),
    search_spaces=search_space,
    n_iter=32, cv=5, scoring='f1', random_state=42, n_jobs
        =-1
)

opt.fit(X_train, Y_train)
```

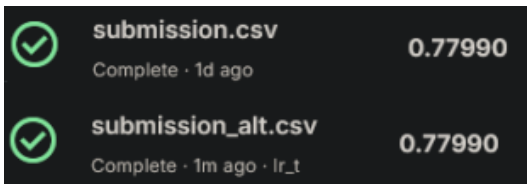
Listing 4: Hyperparameter tuning con Bayesian Optimization

L'ottimizzazione ha prodotto un miglioramento marginale, portando l'accuratezza all'83% sul validation set.

E. Valutazione Finale

Il modello ottimizzato è stato testato sul dataset di test della competizione Kaggle, ottenendo un'accuratezza del 78%. Questo risultato conferma la validità dell'approccio di feature engineering adottato.

Un confronto è stato effettuato anche addestrando il modello con l'inclusione dei titoli: le performance sono risultate sostanzialmente identiche, suggerendo che l'informazione contenuta nei titoli è già catturata da altre feature del modello.



V. Discussione

A. Fattori Chiave di Sopravvivenza

L'analisi ha identificato cinque fattori principali:

1. Sesso: Il protocollo ”donne e bambini prima” è chiaramente evidenziato dai dati, con le donne che mostrano tassi di sopravvivenza 3-4 volte superiori agli uomini.

2. Classe: La stratificazione sociale si riflette drammaticamente nei tassi di sopravvivenza, con i passeggeri di prima classe che avevano accesso prioritario alle scialuppe.

3. Età: I bambini beneficiavano del protocollo di evacuazione prioritaria.

4. Dimensione della Famiglia: Le famiglie di dimensioni medie (2-4 membri) mostravano migliori tassi di sopravvivenza, probabilmente per la presenza di donne e bambini.

5. Posizione sulla Nave: I ponti superiori (A, B, C) offrivano migliori opportunità di evacuazione rispetto ai ponti inferiori.

VI. Conclusioni

L’analisi quantitativa del disastro del Titanic evidenzia gli effetti dei costrutti sociali dell’epoca e rivela l’impatto significativo delle disparità socio-economiche sulla sopravvivenza. Il modello di machine learning sviluppato dimostra che è possibile predire la sopravvivenza con discreta accuratezza (78%) utilizzando caratteristiche demografiche e socio-economiche.

I risultati hanno implicazioni che vanno oltre l’interesse storico, fornendo insights sulla gestione delle emergenze e sull’importanza dell’equità nell’accesso alle risorse di salvataggio durante le crisi.

Riferimenti bibliografici

[1] The Wiltshire and Swindon History Centre, The Story of the Birkenhead” https://wshc.org.uk/the-story-of-the-birkenhead/

[2] Kaggle, ”Titanic - Machine Learning from Disaster,” https://www.kaggle.com/competitions/titanic/overview

[3] Encyclopedia Titanica, ”Titanic Deck Plans,” https://www.encyclopedia-titanica.org/titanic-deckplans

[4] GG Archives, ”Titanic Deck Details,” https://www.ggarchives.com/OceanTravel/Titanic/01-PlanningBuildingLaunching/Decks-ComprehensiveDetails.html

[5] Encyclopedia Titanica, ”Martha Evelyn Stone,” https://www.encyclopedia-titanica.org/titanic-survivor/martha-evelyn-stone.html

[6] Encyclopedia Titanica, ”Thomas Storey,” https://www.encyclopedia-titanica.org/titanic-victim/thomas-storey.html