Progetto di Data Mining: Analisi del Dataset Titanic

Gruppo: Canocchi Lorenzo, Donati Paolo
Istituto Tecnico Statale - G. Marconi

29 maggio 2025

Indice

1 Introduzione

L'affondamento del RMS Titanic nel 1912 è uno degli eventi storici più studiati, non solo per la portata della tragedia umana, ma anche per la quantità e la varietà di dati che ha lasciato. Il dataset dei passeggeri rappresenta una fonte ricca e diversificata di informazioni demografiche e socio-economiche, come età, sesso, classe del biglietto, tariffa, legami familiari e porto d'imbarco. Queste variabili costituiscono una base solida per l'analisi predittiva, l'apprendimento non supervisionato e l'identificazione di schemi ricorrenti.
In questo progetto applichiamo un flusso completo di data mining al dataset, che comprende preparazione dei dati, ingegnerizzazione delle caratteristiche, clustering, classificazione, pattern mining e regressione, con l'obiettivo di individuare i fattori che hanno influenzato la sopravvivenza. La nostra analisi mira non solo a identificare le caratteristiche individuali più determinanti per la sopravvivenza, ma anche a segmentare i passeggeri in gruppi omogenei sulla base di tratti comuni.
Attraverso tecniche moderne di data mining, supportate da visualizzazioni, valutazioni statistiche e strumenti di interpretazione, intendiamo rivelare sia pattern attesi sia informazioni più nascoste all'interno dei dati. Il risultato è un'analisi strutturata che collega i risultati del machine learning al contesto storico, offrendo una comprensione più profonda di come fattori demografici e sociali abbiano influito sulle sorti dei passeggeri durante il disastro.

2 Data Understanding e Preparazione dei Dati

2.1 Analisi semantica degli attributi

È stata condotta un'analisi semantica approfondita al fine di caratterizzare e classificare la natura dei dati da elaborare, con l'obiettivo di definire in modo accurato le strutture informative e le relazioni concettuali presenti nel dataset.

Attributo	**Tipo**	**Significato**
PassengerId	Nominale	Identificatore univoco del passeggero
Survived	Binario	Sopravvivenza: 0 = No, 1 = Sì
Pclass	Ordinale	Classe del biglietto: 1 = Prima, 2 = Seconda, 3 = Terza
Name	Nominale	Nome completo del passeggero
Sex	Binario	Sesso del passeggero: maschio o femmina
Age	Numerico Discreto	Età del passeggero in anni
SibSp	Numerico Discreto	Numero di fratelli/sorelle o coniugi a bordo
Parch	Numerico Discreto	Numero di genitori o figli a bordo
Ticket	Nominale	Numero del biglietto
Fare	Numerico Continuo	Tariffa pagata per il viaggio
Cabin	Nominale	Numero di cabina assegnata
Embarked	Nominale	Porto di imbarco: C = Cherbourg, Q = Queenstown, S = Southampton

2.2 Analisi statistica degli attributi

È stata eseguita un'analisi statistica descrittiva approfondita al fine di caratterizzare la distribuzione dei dati.

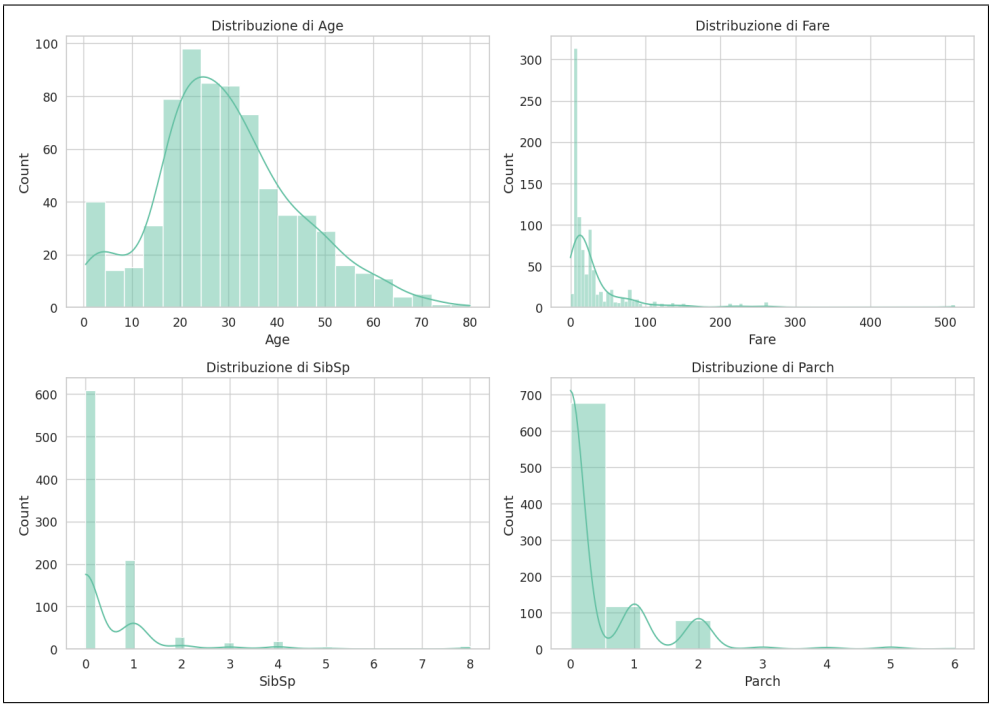
Statistica	Age	Fare	Pclass	SibSp	Parch	Survived
Conteggio	714	891.0	891.0	891.0	891.0	891.0
Media	29.7	32.2	2.31	0.52	0.38	0.38
Deviazione std	14.5	49.7	0.84	1.10	0.81	0.49
Min	0.42	0.0	1.0	0.0	0.0	0.0
25%	20.1	7.9	2.0	0.0	0.0	0.0
50% (Mediana)	28.0	14.5	3.0	0.0	0.0	0.0
75%	38.0	31.0	3.0	1.0	0.0	1.0
Max	80.0	512.3	3.0	8.0	6.0	1.0

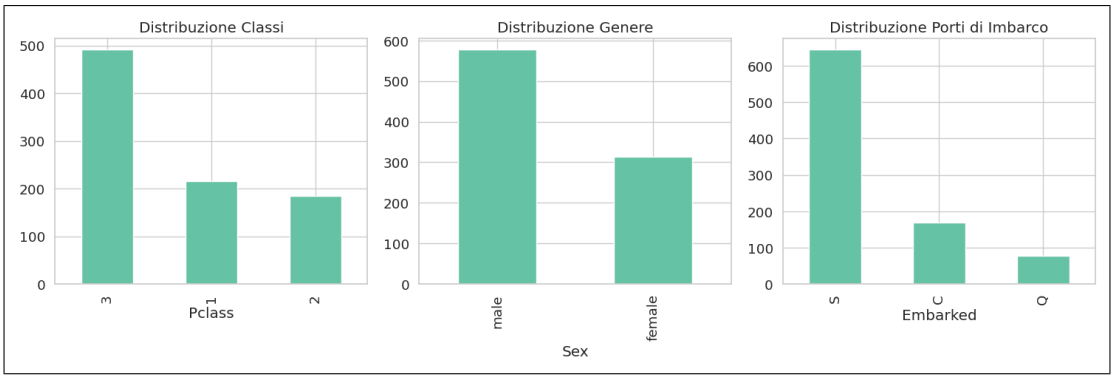
Dall'analisi statistica possiamo osservare:

- L'età media dei passeggeri è di circa 30 anni, con un range significativo da meno di un anno a 80 anni
- La tariffa media è di 32.2 sterline, con una forte asimmetria verso valori alti
- Il 38% dei passeggeri è sopravvissuto (tasso di sopravvivenza)
- La maggior parte dei passeggeri viaggiava in terza classe (mediana = 3)
- La maggior parte dei passeggeri viaggiava senza familiari (mediana di SibSp e Parch = 0)

2.3 Grafici di distribuzione

Per analizzare la distribuzione delle variabili del dataset, sono stati realizzati grafici distinti per quelle numeriche e categoriche. Le variabili numeriche sono rappresentate tramite istogrammi, utili per osservare frequenze, asimmetrie e outlier; quelle categoriche mediante diagrammi a barre, che consentono un confronto immediato tra le modalità. Queste visualizzazioni forniscono una panoramica chiara della struttura dei dati e supportano l'analisi esplorativa.





Dall'analisi dei grafici possiamo evidenziare:

- La distribuzione dell'età mostra una concentrazione nella fascia 20-40 anni
- La distribuzione della tariffa è fortemente asimmetrica (long-tailed) con pochi passeggeri che hanno pagato tariffe molto elevate
- La maggioranza dei passeggeri è imbarcata da Southampton (S)
- C'è una prevalenza di passeggeri maschi rispetto alle femmine
- La maggior parte dei passeggeri viaggiava in terza classe

2.4 Valori mancanti e duplicati

Sono stati individuati e rimossi eventuali record duplicati. Successivamente, è stata effettuata un'analisi della completezza dei dati, i cui risultati sono riportati nella tabella seguente:

Attributo	**Valori Mancanti**	**Percentuale**
Age	177	19.9%
Cabin	687	77.1%
Embarked	2	0.2%

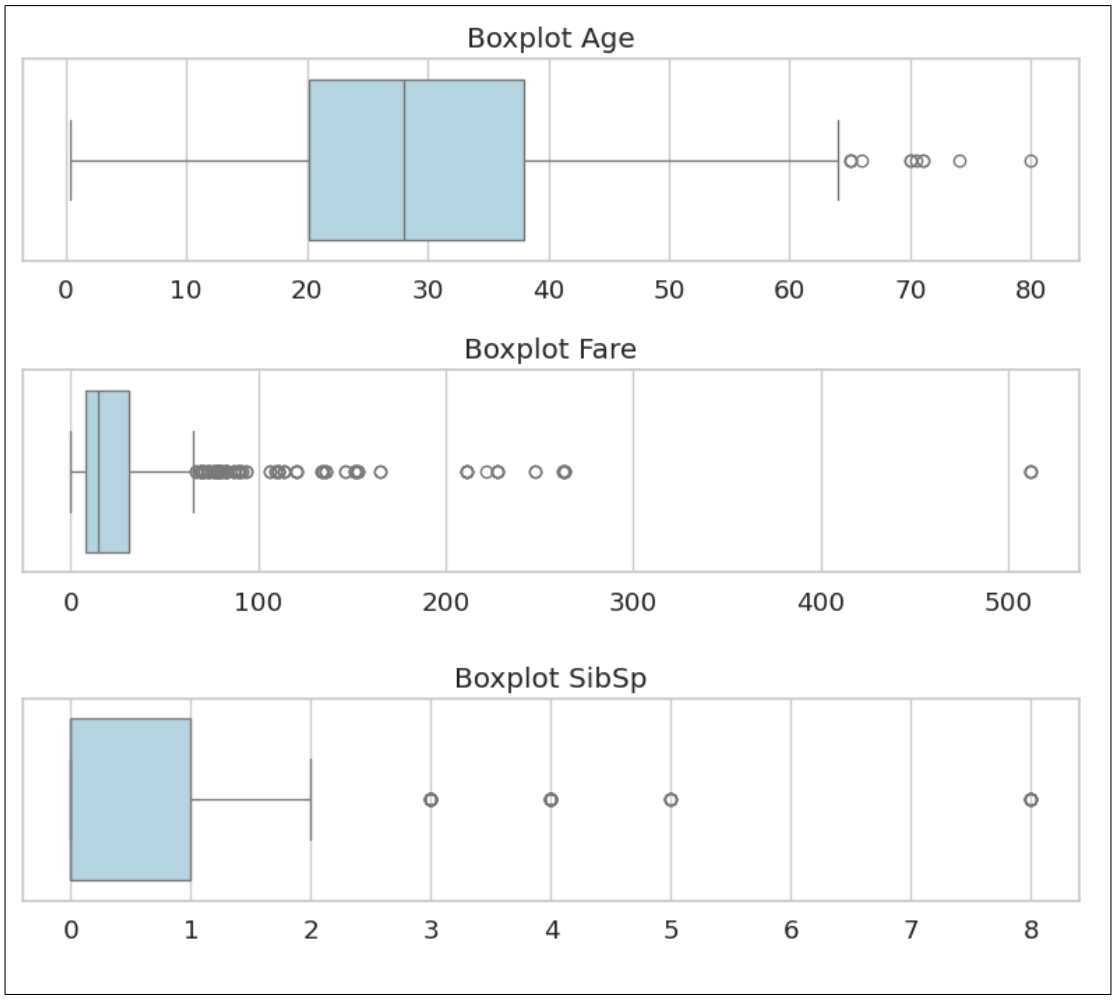
Si è scelto di non intervenire sul trattamento dei valori mancanti, al fine di preservare l'integrità e la coerenza informativa del dataset originale, evitando l'introduzione di distorsioni potenzialmente significative.

2.5 Outliers: Box Plot e IQR

Il **boxplot** è una rappresentazione grafica sintetica della distribuzione di una variabile quantitativa. La scatola centrale è delimitata dal primo quartile (Q1) e dal terzo quartile (Q3), che racchiudono il 50% centrale delle osservazioni, mentre la linea interna indica la mediana, ovvero il valore centrale della distribuzione ordinata.
Ai margini della scatola si estendono i *whiskers*, linee che raggiungono i valori estremi entro limiti definiti. I punti al di fuori di tali limiti sono classificati come *outlier*, ossia valori anomali. Per identificare gli outlier si utilizza l'intervallo interquartile (IQR), calcolato come la differenza tra il terzo e il primo quartile:

$$\text{IQR} = Q3 - Q1$$

Un valore è considerato anomalo se si discosta oltre i limiti $Q1 - 1.5 \times \text{IQR}$ e $Q3 + 1.5 \times \text{IQR}$, consentendo di rilevare valori estremi rispetto alla distribuzione complessiva.

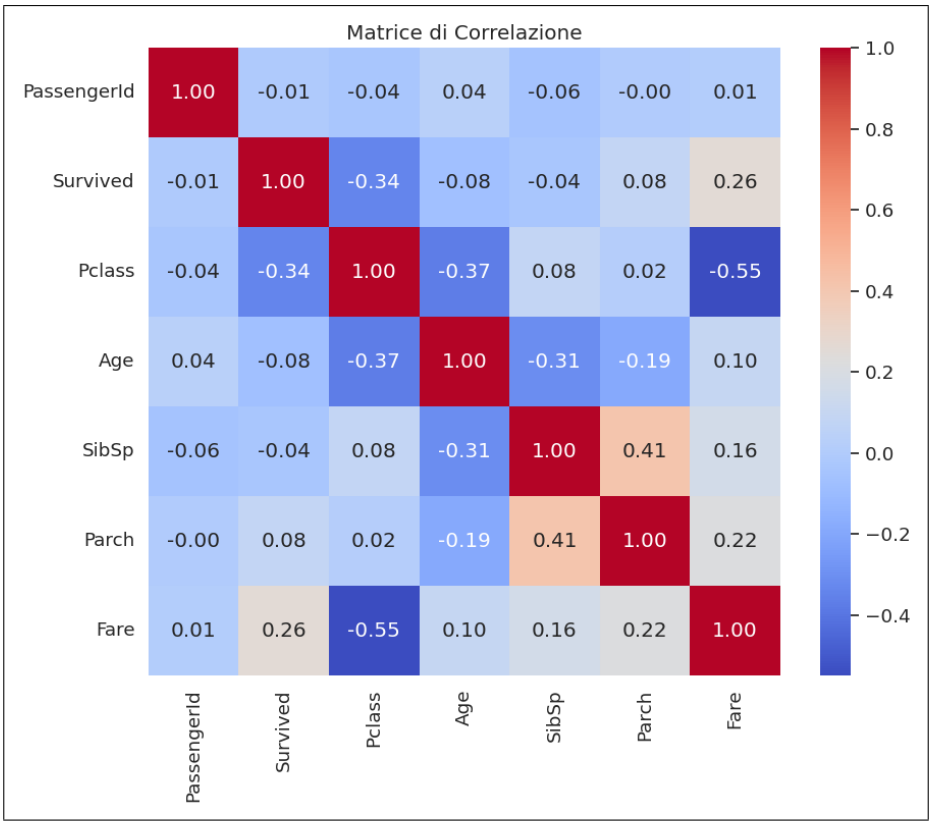


I principali outliers identificati sono i seguenti:

- *Fare*: Alcune tariffe estremamente elevate (> 250 sterline)
- *Age*: Un ridotto numero di passeggeri con età molto avanzata (> 70 anni)
- *SibSp*: Alcuni gruppi familiari particolarmente numerosi

2.6 Matrice di correlazione

La *matrice di correlazione* è una tabella che riporta i coefficienti di correlazione calcolati per ogni coppia di variabili numeriche all'interno di un dataset. Questi coefficienti misurano la forza e la direzione delle relazioni lineari tra le variabili: un valore positivo indica una relazione diretta, mentre un valore negativo segnala una relazione inversa.
L'analisi della matrice consente di identificare facilmente associazioni positive o negative significative tra variabili, nonché di individuare coppie fortemente correlate. Queste informazioni sono fondamentali per attività quali la selezione delle variabili, la riduzione della dimensionalità o la prevenzione della multicollinearità in modelli statistici e predittivi. Inoltre, la matrice favorisce la scoperta di pattern ricorrenti o gruppi di variabili con comportamenti simili, facilitando una comprensione più approfondita della struttura dei dati.



Non sono state riscontrate correlazione significative tra le diverse features analizzate.

2.7 Preparazione dei dati per il Machine Learning

2.7.1 Suddivisione del dataset (60/20/20)

Al fine di sviluppare, ottimizzare e valutare correttamente i modelli predittivi, il dataset è stato suddiviso in tre sottoinsiemi distinti secondo le proporzioni 60% / 20% / 20%. In particolare:

- Il *training set* (60%) è stato utilizzato per l'apprendimento del modello, ovvero per stimare i parametri interni sulla base dei dati disponibili.
- Il *validation set* (20%) è servito per la fase di ottimizzazione, ad esempio per la selezione degli iperparametri o per prevenire fenomeni di overfitting.
- Il *test set* (20%) è stato infine impiegato per la valutazione finale delle prestazioni del modello su dati mai visti in precedenza, fornendo una stima oggettiva della sua capacità di generalizzazione.

Questa strategia di suddivisione consente di garantire una valutazione robusta e affidabile delle performance del modello, preservando l'indipendenza tra fase di addestramento e fase di test.

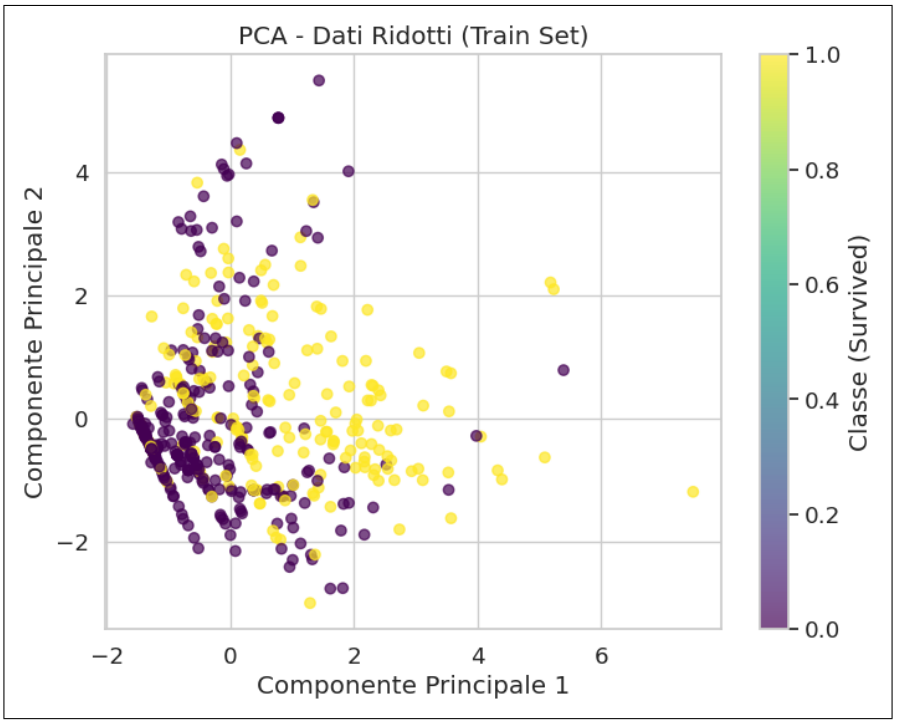
2.7.2 Normalizzazione

Per garantire che tutte le variabili numeriche contribuiscano equamente ai modelli predittivi e per migliorare la stabilità numerica degli algoritmi, è stata eseguita una fase di normalizzazione dei dati. In particolare, sono state applicate due tecniche di scaling: *StandardScaler* e *MinMaxScaler*, selezionate in base alle caratteristiche specifiche dei dati e ai requisiti degli algoritmi utilizzati.

- *StandardScaler*: questa tecnica standardizza ogni feature sottraendo la media e dividendo per la deviazione standard, producendo una distribuzione con media zero e varianza unitaria. È particolarmente indicata quando i dati seguono una distribuzione normale o quando si utilizzano algoritmi che assumono tale distribuzione, come la regressione lineare o le reti neurali. Tuttavia, è sensibile alla presenza di outlier, poiché questi influenzano significativamente la media e la deviazione standard del dataset.
- *MinMaxScaler*: questa tecnica normalizza i dati scalando ogni feature in un intervallo specificato, tipicamente [0, 1]. È utile quando si desidera mantenere la forma della distribuzione originale dei dati e quando si utilizzano algoritmi che non assumono una distribuzione normale, come k-Nearest Neighbors o algoritmi basati su distanze. Tuttavia, è anch'essa sensibile agli outlier, poiché questi determinano i valori minimi e massimi utilizzati per la scalatura .

2.7.3 Riduzione dimensionale con PCA

Per ridurre la dimensionalità del dataset e migliorare l'efficienza computazionale dei modelli, è stata applicata la *Principal Component Analysis* (PCA), una tecnica di proiezione lineare che consente di trasformare le variabili originali in un nuovo insieme di variabili ortogonali dette *componenti principali*.
La PCA è stata eseguita successivamente alla normalizzazione dei dati, condizione necessaria affinché le componenti non siano dominate da feature con varianze più elevate. L'obiettivo era mantenere la massima quantità di varianza possibile utilizzando un numero ridotto di componenti, selezionati in modo da conservare una quota significativa dell'informazione originaria.

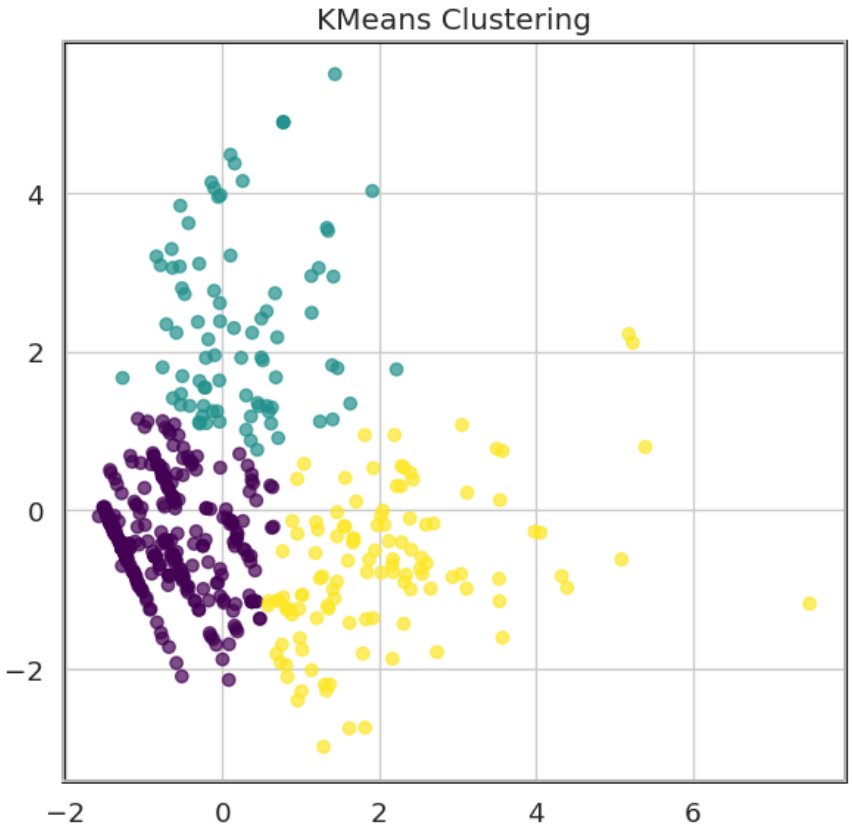


3 Clustering

Per analizzare la struttura latente dei dati e rilevare la presenza di gruppi omogenei, è stata applicata una procedura di *clustering*, una tecnica di apprendimento non supervisionato che consente di suddividere un insieme di osservazioni in sottoinsiemi coerenti, detti *cluster*, sulla base della similarità tra le istanze.
Il clustering ha lo scopo di raggruppare i dati in modo tale che le osservazioni appartenenti allo stesso gruppo risultino il più possibile simili tra loro, mentre quelle appartenenti a gruppi diversi siano dissimili. Questo approccio permette di evidenziare pattern nascosti, strutture ricorrenti o segmentazioni naturali all'interno del dataset, senza fare ricorso a etichette predefinite.
L'analisi è stata condotta su dati precedentemente normalizzati e trasformati tramite PCA, per ridurre la dimensionalità e migliorare la qualità dei risultati, sia in termini di efficienza computazionale che di interpretabilità dei cluster.

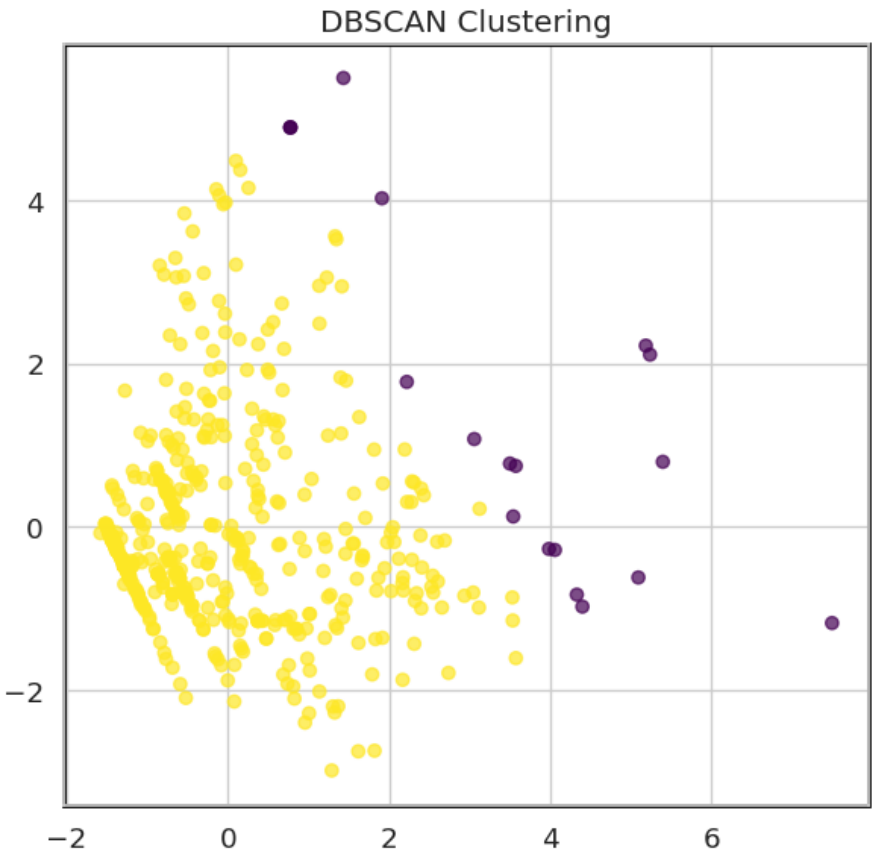
3.1 K-Means

Tra i metodi di clustering più utilizzati, il *K-Means* è un algoritmo iterativo che suddivide i dati in un numero predefinito di cluster k.
Il procedimento si basa sulla minimizzazione della distanza intra-cluster, calcolata generalmente rispetto ai centroidi, che rappresentano il centro geometrico di ciascun gruppo.
Il funzionamento dell'algoritmo prevede un'assegnazione iniziale casuale dei centroidi, seguita da una fase iterativa in cui i punti vengono assegnati al centroide più vicino, e i centroidi vengono successivamente aggiornati sulla base delle nuove assegnazioni. Il processo si arresta al raggiungimento della convergenza, ossia quando le assegnazioni non cambiano più in modo significativo.
La scelta del numero ottimale di cluster è stata guidata tramite metodi quantitativi come il *Silhouette Score* e l'analisi del *dentro-cluster sum of squares* (elbow method).



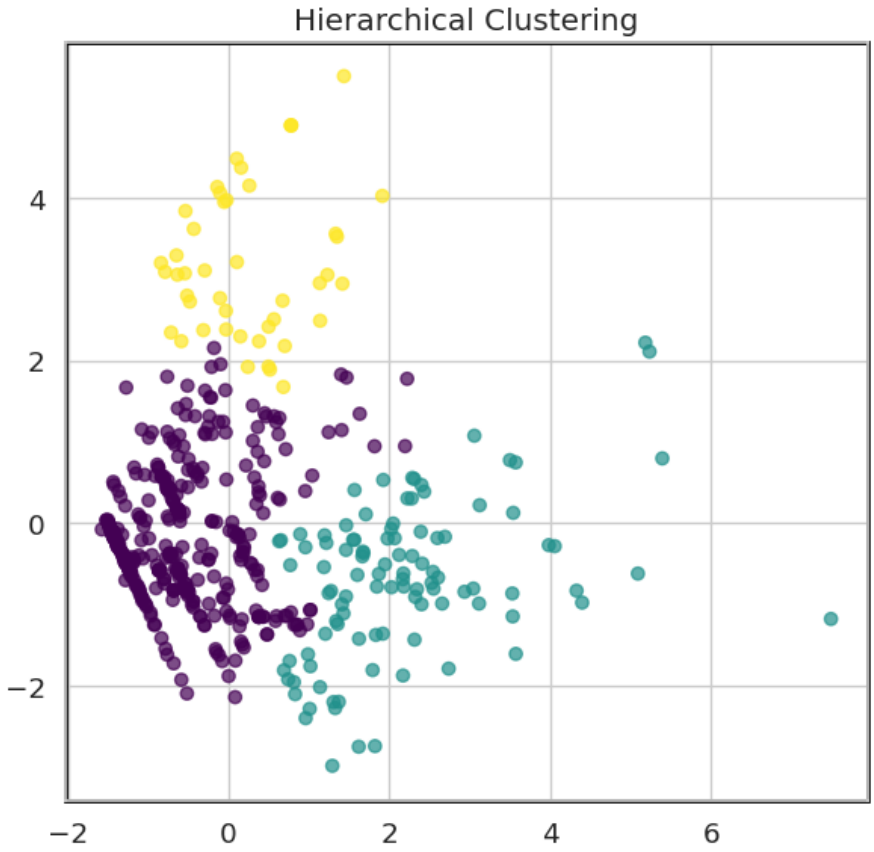
3.2 DBSCAN

Il metodo *DBSCAN* (Density-Based Spatial Clustering of Applications with Noise) è un algoritmo di clustering basato sulla densità, che permette di individuare gruppi di punti densamente connessi, separandoli da regioni a bassa densità, considerate come rumore o outlier.
DBSCAN non richiede la specifica a priori del numero di cluster e risulta particolarmente efficace nell'identificare strutture di forma arbitraria. L'algoritmo si basa su due parametri fondamentali: la distanza massima ε per definire il vicinato di un punto, e il numero minimo di punti (`minPts`) richiesti per formare una regione densa.
Questa tecnica è particolarmente utile in presenza di dataset rumorosi o con densità non uniformi, ed è robusta nei confronti di outlier e anomalie



3.3 Clustering Gerarchico

Il *clustering gerarchico* è una tecnica che costruisce una struttura ad albero (dendrogramma) che rappresenta le relazioni di inclusione tra i diversi gruppi di dati. Si basa sull'unione o sulla divisione successiva dei cluster, in modo agglomerativo (bottom-up) o divisivo (top-down).
Nel caso agglomerativo, inizialmente ogni osservazione è considerata come un cluster a sé stante, e i cluster vengono fusi iterativamente sulla base di una misura di similarità (es. linkage: single, complete, average o Ward). Il risultato è una gerarchia di cluster che può essere tagliata a un certo livello per ottenere una partizione desiderata.
Questa tecnica è particolarmente indicata per l'analisi esplorativa, in quanto non richiede una scelta preliminare del numero di cluster, ed è utile per evidenziare la struttura gerarchica presente nei dati.

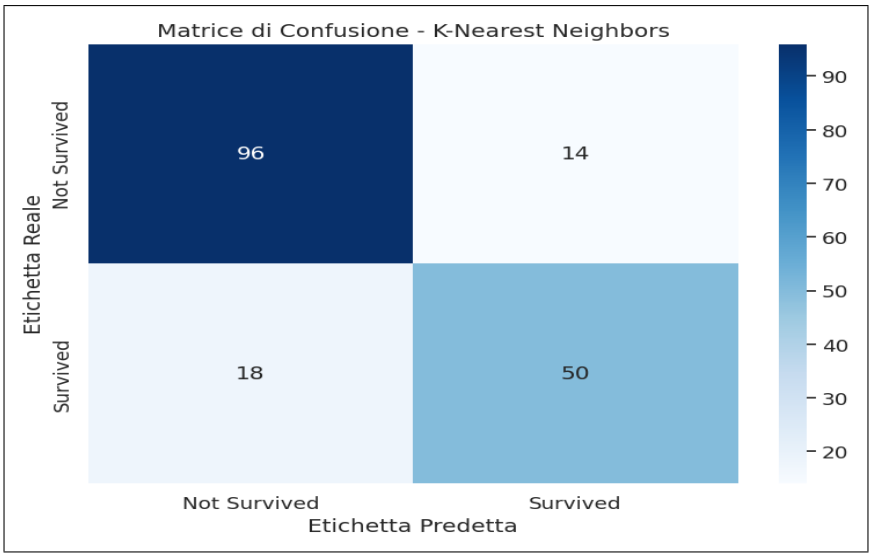


4 Classificazione con Modelli di Machine Learning

La *classificazione* è una tecnica di apprendimento supervisionato finalizzata ad assegnare ad ogni osservazione l'appartenenza ad una o più categorie predefinite (etichette). L'obiettivo principale consiste nel costruire un modello in grado di apprendere dalla relazione tra le feature (variabili indipendenti) e le classi target (variabili dipendenti) presenti nel dataset di addestramento, per poi applicare tale conoscenza a dati non visti.
Il processo di classificazione si articola in diverse fasi. In primo luogo, il modello viene addestrato su un insieme di dati etichettati, in cui ogni esempio è associato alla sua rispettiva classe. Durante questa fase, il modello apprende le regole e le caratteristiche distintive che differenziano le diverse classi. Successivamente, il modello viene validato e testato per verificarne la capacità di generalizzazione e per quantificare le sue prestazioni tramite metriche specifiche (accuratezza, precisione, recall, F1-score e support).

4.1 K-Nearest Neighbors (KNN)

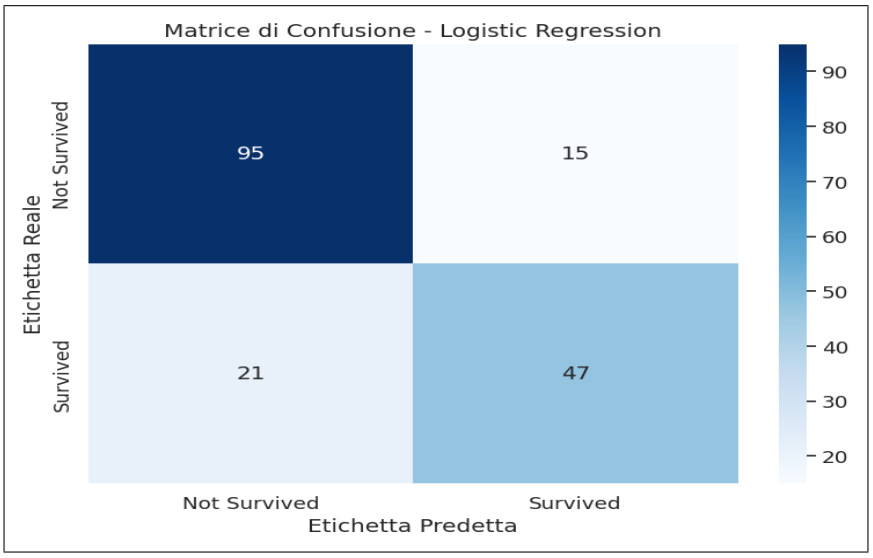
Il *K-Nearest Neighbors* è un algoritmo di classificazione basato sulla distanza. Classifica un'osservazione assegnandola alla classe più rappresentata tra i suoi k vicini più prossimi nel set di addestramento, secondo una metrica di distanza (solitamente euclidea). È un metodo non parametrico e particolarmente sensibile alla scelta del valore di k, nel nostro caso 5, e alla scala delle variabili.
Sotto è riportata la matrice di confusione del modello e una tabella contenente le prestazioni i esso sul validation set.



Classe	Precision	Recall	F1-Score	Support
Not Survived	0.84	0.87	0.86	110
Survived	0.78	0.74	0.76	68
Accuracy			0.82	178
Macro avg	0.81	0.80	0.81	178
Weighted avg	0.82	0.82	0.82	178

4.2 Logistic Regression

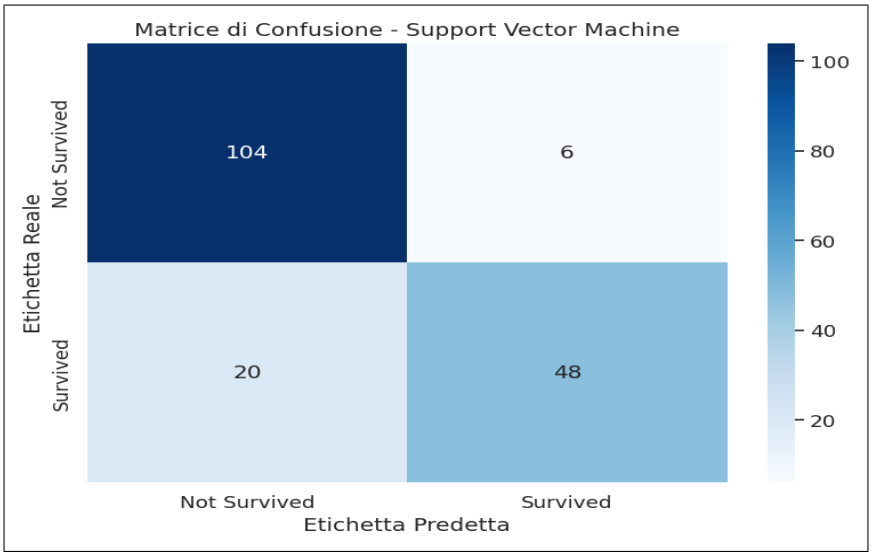
La *regressione logistica* è un modello lineare per la classificazione binaria. Stima la probabilità che un'osservazione appartenga a una determinata classe tramite la funzione logistica (sigmoide). L'output è una probabilità che può essere soggetta a una soglia per la classificazione finale. È interpretabile e adatta a problemi lineari.
Sotto è riportata la matrice di confusione del modello e una tabella contenente le prestazioni i esso sul validation set.



Classe	Precision	Recall	F1-Score	Support
Not Survived	0.82	0.86	0.84	110
Survived	0.76	0.69	0.72	68
Accuracy			0.80	178
Macro avg	0.79	0.78	0.78	178
Weighted avg	0.80	0.80	0.80	178

4.3 Support Vector Machine (SVM)

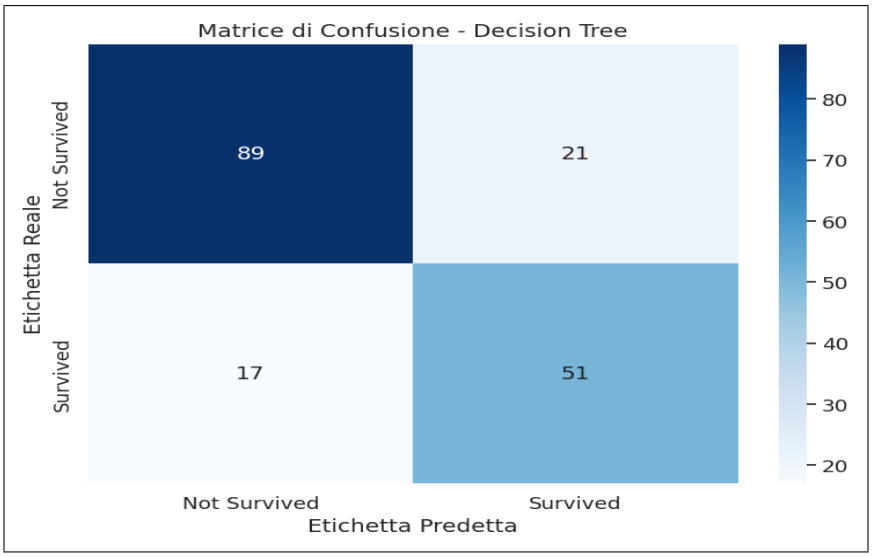
La *Support Vector Machine* cerca l'iperpiano che separa le classi massimizzando il margine tra i punti più vicini, detti *support vectors*. Può essere estesa a problemi non lineari mediante kernel, che trasformano lo spazio delle caratteristiche. È robusta e spesso efficace anche in spazi ad alta dimensionalità.
Sotto è riportata la matrice di confusione del modello e una tabella contenente le prestazioni i esso sul validation set.



Classe	Precision	Recall	F1-Score	Support
Not Survived	0.84	0.95	0.89	110
Survived	0.89	0.71	0.79	68
Accuracy			0.85	178
Macro avg	0.86	0.83	0.84	178
Weighted avg	0.86	0.85	0.85	178

4.4 Decision Tree

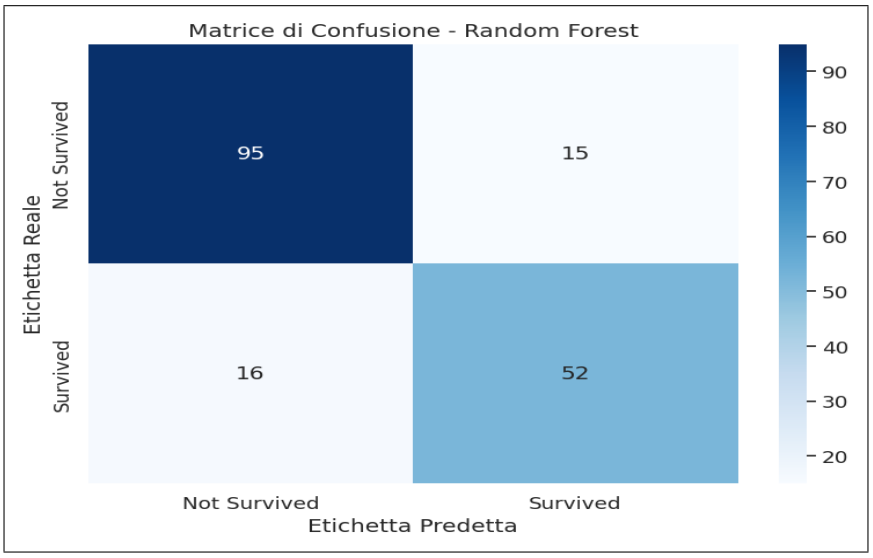
I *decision tree* sono modelli predittivi basati su una struttura ad albero, in cui ogni nodo interno rappresenta una condizione su una variabile, e ogni ramo una conseguente suddivisione del dataset. Il modello apprende partizionando ricorsivamente i dati fino a raggiungere omogeneità nei nodi foglia. È interpretabile ma soggetto a overfitting.
Sotto è riportata la matrice di confusione del modello e una tabella contenente le prestazioni in esso sul validation set.



Classe	Precision	Recall	F1-Score	Support
Not Survived	0.84	0.81	0.82	110
Survived	0.71	0.75	0.73	68
Accuracy			0.79	178
Macro avg	0.77	0.78	0.78	178
Weighted avg	0.79	0.79	0.79	178

4.5 Random Forest

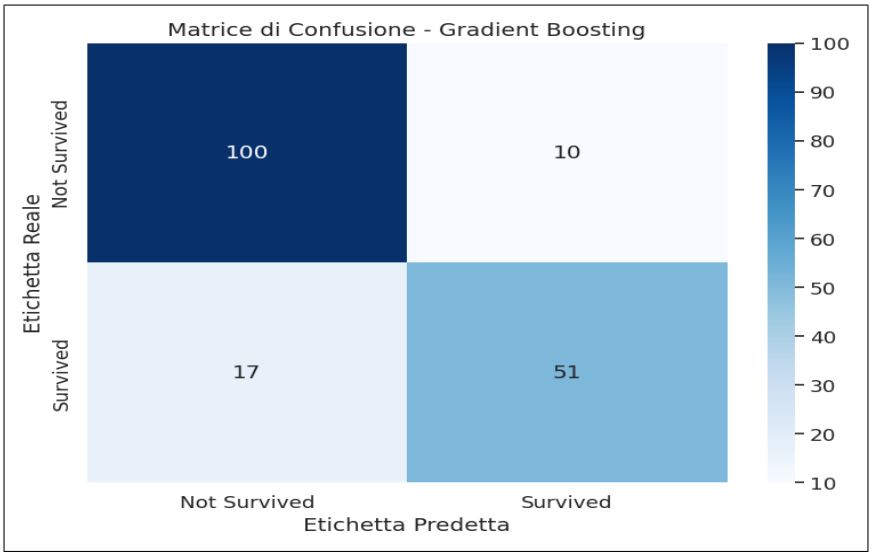
La *Random Forest* è un metodo di ensemble che costruisce una moltitudine di alberi decisionali su sottocampioni del dataset e aggrega le loro previsioni (voto di maggioranza). Riduce la varianza dei singoli alberi e migliora la generalizzazione, risultando più stabile e accurato rispetto a un singolo decision tree.
Sotto è riportata la matrice di confusione del modello e una tabella contenente le prestazioni i esso sul validation set.



Classe	Precision	Recall	F1-Score	Support
Not Survived	0.86	0.86	0.86	110
Survived	0.78	0.76	0.77	68
Accuracy			0.83	178
Macro avg	0.82	0.81	0.82	178
Weighted avg	0.83	0.83	0.83	178

4.6 Gradient Boosting

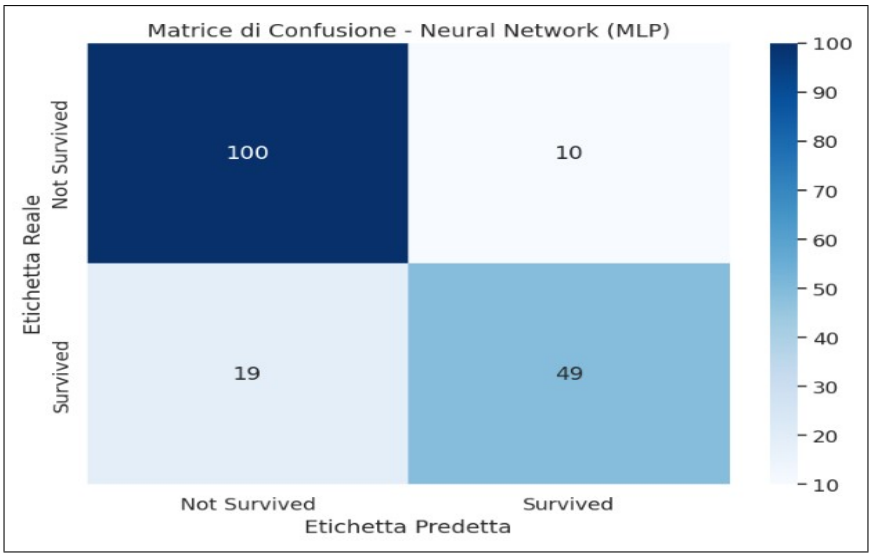
Il *Gradient Boosting* è una tecnica di ensemble che costruisce sequenzialmente modelli deboli (solitamente alberi decisionali), dove ogni nuovo modello cerca di correggere gli errori residui del precedente. L'ottimizzazione avviene tramite discesa del gradiente su una funzione di perdita. È potente ma può richiedere tuning attento.
Sotto è riportata la matrice di confusione del modello e una tabella contenente le prestazioni i esso sul validation set.



Classe	Precision	Recall	F1-Score	Support
Not Survived	0.85	0.91	0.88	110
Survived	0.84	0.75	0.79	68
Accuracy			0.85	178
Macro avg	0.85	0.83	0.84	178
Weighted avg	0.85	0.85	0.85	178

4.7 Multi-layer Perceptron (MLP)

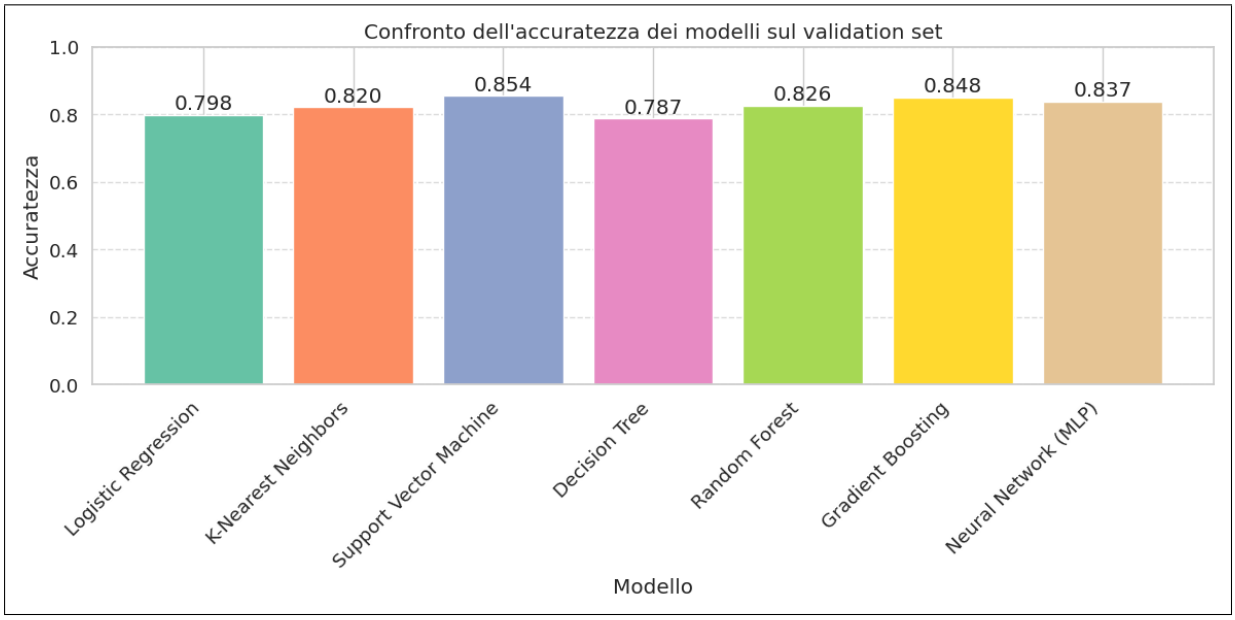
Il *Multi-layer Perceptron* è un tipo di rete neurale artificiale feedforward composta da uno o più livelli nascosti (hidden layer), ciascuno costituito da neuroni completamente connessi. Utilizza funzioni di attivazione non lineari (es. ReLU) e viene addestrato tramite backpropagation. È adatto a catturare relazioni complesse e non lineari.
Sotto è riportata la matrice di confusione del modello e una tabella contenente le prestazioni i esso sul validation set.



Classe	Precision	Recall	F1-Score	Support
Not Survived	0.84	0.91	0.87	110
Survived	0.83	0.72	0.77	68
Accuracy			0.84	178
Macro avg	0.84	0.81	0.82	178
Weighted avg	0.84	0.84	0.83	178

5 Valutazione e Ottimizzazione

5.1 Confronto tra i modelli



Nel confronto tra modelli applicati al dataset Titanic, quelli più complessi come Support Vector Machine, Gradient Boosting e reti neurali mostrano le migliori prestazioni, grazie alla loro capacità di catturare relazioni non lineari tra le variabili. Al contrario, modelli più semplici come Logistic Regression e Decision Tree risultano meno efficaci. Questo suggerisce che per problemi con dinamiche complesse, come la previsione della sopravvivenza nel Titanic, è preferibile utilizzare algoritmi avanzati capaci di modellare interazioni più articolate.

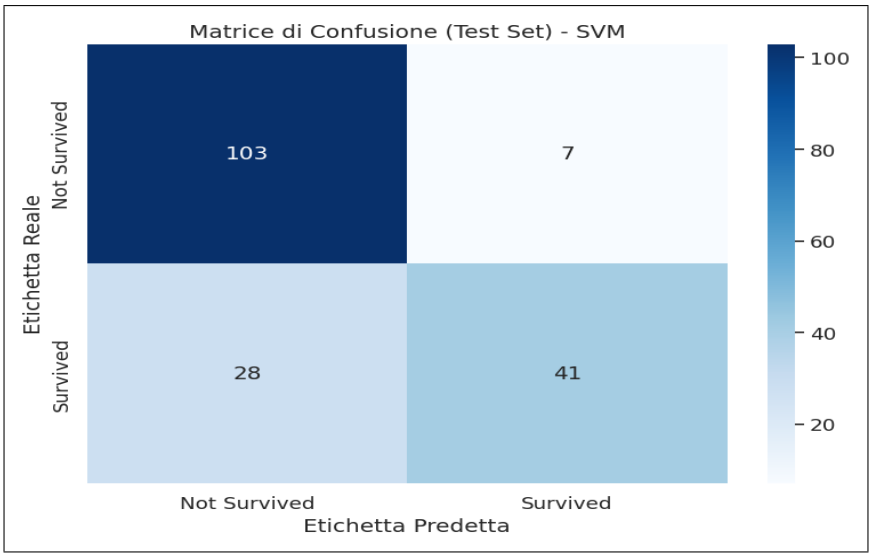
5.2 Ottimizzazione con Grid Search dei migliori 3 modelli

Per ottenere un modello più accurato applichiamo l'ottimizzazione iperparametrica sui tre modelli di machine learning: Support Vector Machine (SVM), Random Forest e Gradient Boosting, applicati al dataset Titanic. L'obiettivo dell'ottimizzazione è massimizzare la capacità predittiva dei modelli, migliorando le prestazioni attraverso la scelta accurata dei valori degli iperparametri.
Per ciascun modello è stata utilizzata la tecnica della Grid Search, un metodo esaustivo che esplora sistematicamente tutte le combinazioni possibili di un insieme predefinito di iperparametri. Ad esempio, per il modello SVM sono stati testati diversi valori di C, gamma e kernel, mentre per Random Forest e Gradient Boosting sono stati variati parametri come n_estimators, max_depth, learning_rate, min_samples_split e altri.
La valutazione di ogni combinazione di iperparametri è stata eseguita tramite validazione incrociata (cross-validation) a 3 fold, dove il training set viene suddiviso in tre sottoinsiemi: due vengono utilizzati per l'addestramento e uno per la validazione, a rotazione. Questo consente una stima più robusta delle performance e riduce il rischio di overfitting (L'overfitting si verifica quando un modello impara troppo bene i dati di training,perdendo così la capacità di generalizzare su dati nuovi).
La metrica utilizzata per confrontare le configurazioni è stata l'accuratezza, ovvero la percentuale di classificazioni corrette. Una volta identificata la configurazione ottimale per ciascun modello, esso è stato valutato su un validation set separato, per ottenere una stima più realistica delle sue prestazioni generali.

Modello	**Accuratezza CV**	**Accuratezza Validation**
SVM	0.8109	0.8596
Random Forest	0.8146	0.8146
Gradient Boosting	0.8258	0.8371
Miglior Modello		**SVM (0.8596)**

5.3 Prestazioni del miglior modello sul test set

Il modello SVM viene valutato sul test set, che è stato mantenuto separato per tutta la durata del processo di addestramento e ottimizzazione. Questa valutazione fornisce una stima imparziale della capacità del modello di generalizzare su dati completamente nuov.
Viene calcolata l'accuratezza sul test set, seguita da un classification report dettagliato che include precision, recall e F1-score per ciascuna classe. Inoltre, viene visualizzata la matrice di confusione, utile per analizzare visivamente gli errori di classificazione.



Classe	Precision	Recall	F1-Score	Support
Not Survived	0.79	0.94	0.85	110
Survived	0.85	0.59	0.70	69
Accuracy			0.80	179
Macro avg	0.82	0.77	0.78	179
Weighted avg	0.81	0.80	0.80	179

6 Conclusioni e Riflessioni Finali

L'analisi condotta sul dataset Titanic ha dimostrato come un flusso completo di data mining—dalla preparazione dei dati alla modellazione supervisionata—possa mettere in luce sia i pattern più evidenti sia quelli meno intuitivi. Dopo aver gestito in modo rigoroso i valori mancanti e normalizzato le variabili mediante StandardScaler e MinMaxScaler, l'applicazione della PCA ha consentito di ridurre significativamente la dimensionalità mantenendo la maggior parte della varianza informativa, semplificando così lo spazio delle features.
Nella fase di clustering, K-Means ha segmentato i passeggeri in gruppi omogenei sulla base di metriche quantitative, mentre DBSCAN e clustering gerarchico hanno ulteriormente arricchito l'interpretazione esplorativa identificando outlier e strutture gerarchiche nei dati.
L'approccio di classificazione supervisionata ha evidenziato come modelli complessi, quali Support Vector Machine, Gradient Boosting e reti neurali multilayer perceptron, siano in grado di captare relazioni non lineari tra le variabili, superando ampiamente gli algoritmi più semplici (ad es. logistic regression e decision tree) in termini di accuratezza e F1-score sul validation set. In particolare, l'SVM ottimizzato tramite grid search ha raggiunto un'accuratezza di validazione pari a 0.8596 e si è confermato il miglior modello, mantenendo un buon equilibrio tra precision e recall anche nel test finale (accuratezza 0.80, recall della classe "Survived" 0.59).
Questa sperimentazione valida l'importanza di una selezione calibrata degli iperparametri. L'arricchimento del dataset con variabili estratte da nomi o cabine, l'adozione di ensemble stacking tra i modelli top-performer e l'integrazione di informazioni temporali relative alle dinamiche di abbandono della nave potrebbero ulteriormente potenziare le capacità predittive. In ultima analisi, l'approccio organico qui descritto getta le basi per una comprensione sempre più profonda dei fattori che hanno determinato la sopravvivenza nel disastro del Titanic.