Progetto di Machine Learning

Analisi predittiva e comportamento azionario di DIS (Disney) e WBD (Warner Bros)

(2019–2023)

A cura di

Valentino Lucci 608725
Riccardo Di Gregorio 606017
Diego Lo Curto 575524

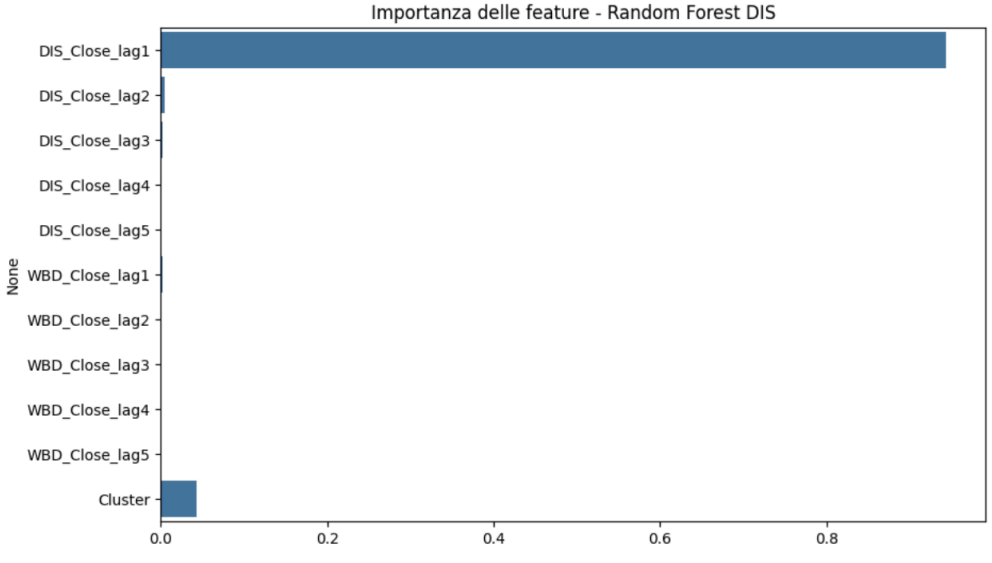
Il presente progetto si propone di analizzare e modellare il comportamento storico delle azioni Disney (DIS) e Warner Bros. Discovery (WBD) nel periodo compreso tra il 2019 e il 2023, con l'obiettivo di elaborare previsioni accurate sui prezzi futuri. L'analisi unisce strumenti statistici e algoritmi di machine learning – in particolare Random Forest e regressione lineare – per comprendere le dinamiche di mercato, valutare la correlazione tra i due titoli e identificare pattern predittivi. Il progetto include inoltre una valutazione della performance dei modelli tramite indicatori di errore, distribuzioni, matrici di confusione e analisi dei cluster. L'approccio seguito unisce rigore quantitativo e interpretazione economico-finanziaria, fornendo un quadro completo utile tanto alla ricerca accademica quanto alla pratica operativa.

ANALISI DELLE METRICHE. Nel presente confronto tra modelli predittivi applicati ai titoli DIS (Disney) e WBD (Warner Bros Discovery), emergono considerazioni metodologiche di rilievo.

```
Random Forest – RMSE DIS: 4.69, R^2: 0.62
Random Forest – RMSE WBD: 0.56, R^2: 0.88
Linear Regression – RMSE DIS: 1.55, R^2: 0.96
```

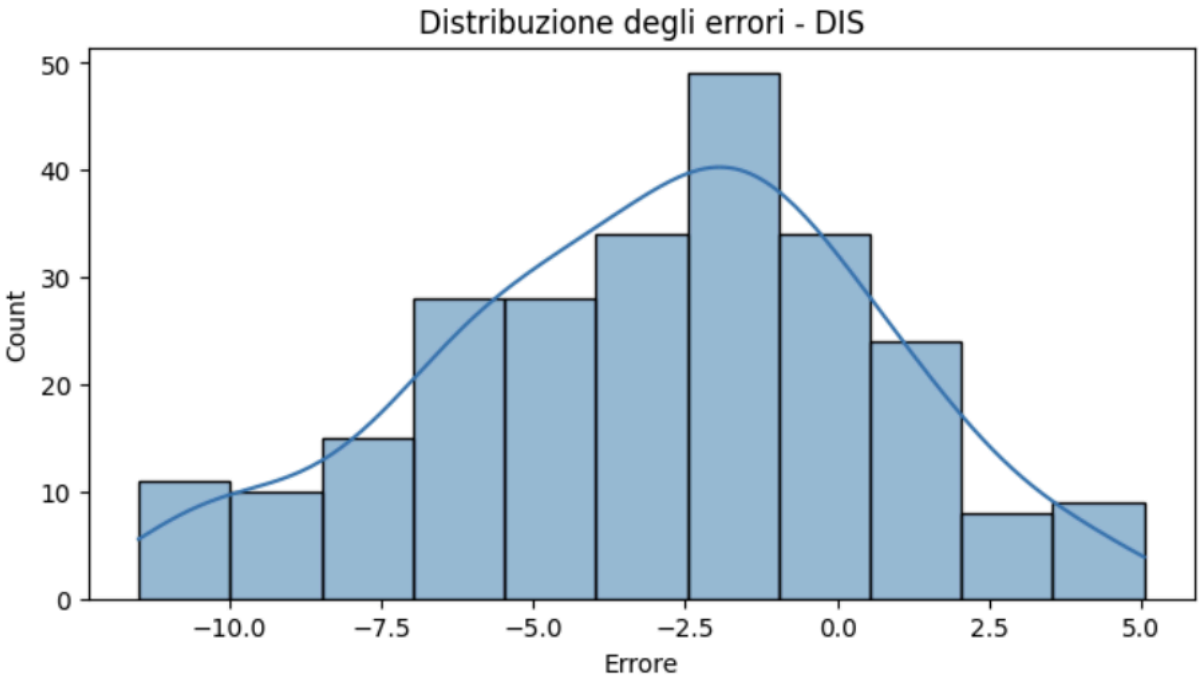
Il modello Random Forest, seppur noto per la sua capacità di cogliere relazioni non lineari e complesse nei dati, ha mostrato performance contrastanti: nel caso di DIS, il valore di RMSE pari a 4.69 e un coefficiente di determinazione R^2 di 0.62 indicano una capacità esplicativa piuttosto limitata e una discreta presenza di errore, suggerendo che il modello fatichi a rappresentare con efficacia l'andamento del titolo, probabilmente per una scarsa informatività delle variabili impiegate o per una configurazione subottimale degli iperparametri. Al contrario, sul titolo WBD, la Random Forest ha ottenuto risultati molto più soddisfacenti, con un RMSE di appena 0.56 e un R^2 di 0.88, segno che in questo caso il modello riesce a cogliere gran parte della varianza del prezzo e a fornire stime affidabili e ben generalizzate. Di particolare interesse è il confronto tra Random Forest e regressione lineare applicati entrambi a DIS: sorprendentemente, il modello lineare ha registrato un RMSE molto più basso (1.55) e un R^2 decisamente più alto (0.96), evidenziando che, in certi contesti, una struttura lineare può offrire prestazioni superiori rispetto a modelli più complessi. Questa osservazione conferma l'importanza della scelta del modello in funzione della natura del dato e della relazione tra le variabili: la semplicità della regressione lineare si adatta meglio a scenari con dinamiche regolari, mentre la Random Forest si rivela più adatta in presenza di pattern complessi, ma necessita di tuning e feature engineering accurati. In sintesi, il confronto tra le tecniche adottate non evidenzia una superiorità assoluta, ma sottolinea l'urgenza di valutazioni preliminari sui dati, di test comparativi rigorosi e di un approccio adattivo che consideri sia le metriche di errore sia la coerenza strutturale tra dati e modello.

IMPORTANZA DELLE FEATURE PER IL MODELLO RANDOM FOREST APPLICATO A DIS. Il grafico relativo all'importanza delle variabili nel modello Random Forest applicato al titolo DIS (Disney) rivela in modo netto la centralità della feature **DIS_Close_lag1**, ovvero il prezzo di chiusura del titolo nel giorno immediatamente precedente.



Questa variabile mostra un peso predominante, con un valore di importanza che supera ampiamente tutte le altre feature, indicando che il modello basa la propria previsione quasi esclusivamente sull'andamento immediatamente precedente del titolo. Le altre variabili lag correlate a DIS (dal secondo al quinto ritardo) presentano valori pressoché trascurabili, segnalando una debole capacità esplicativa o ridondanza informativa. Ancor più marginale è il contributo delle variabili lag del titolo WBD, che risultano sostanzialmente ininfluenti: ciò suggerisce che, nel contesto di questa analisi, il titolo WBD non fornisce un'informazione predittiva utile per la dinamica di DIS, nonostante l'apparente correlazione osservata nei passaggi precedenti. Infine, l'attributo "Cluster", presumibilmente derivato da un algoritmo di clustering non supervisionato (es. K-means), mostra una minima rilevanza, confermando che l'appartenenza a un cluster non apporta contributi significativi al miglioramento delle previsioni. Questo risultato complessivo mette in luce la tendenza del modello Random Forest a concentrarsi su variabili con pattern autoregressivi molto marcati, ma evidenzia anche una certa limitatezza nella capacità del modello di integrare efficacemente informazioni esterne o multivariate se queste non mostrano immediata coerenza statistica con il target. In sintesi, il grafico conferma un approccio fortemente autoregressivo, suggerendo che un affinamento del feature set o l'introduzione di trasformazioni più sofisticate (come indicatori tecnici o interazioni non lineari) potrebbe essere necessario per migliorare la profondità predittiva del modello.

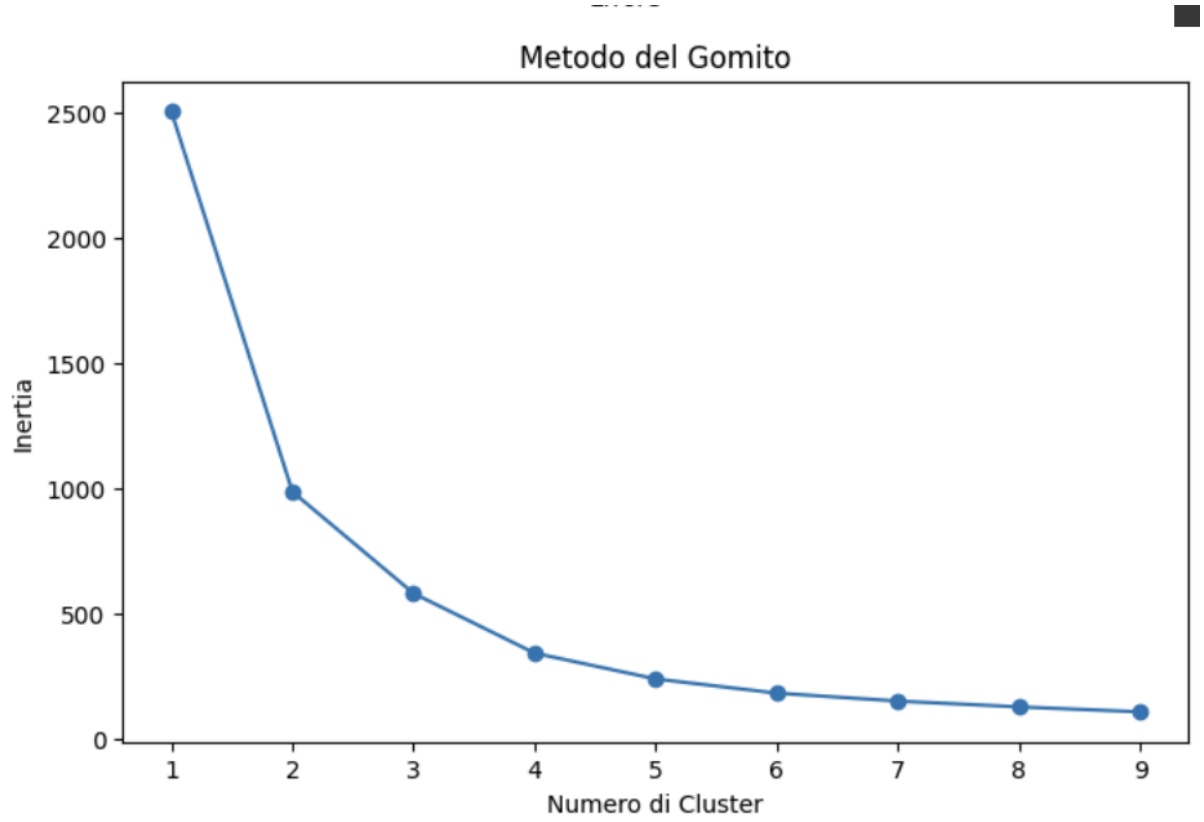
DISTRIBUZIONE DEGLI ERRORI PER DIS. Il grafico mostra la distribuzione degli errori commessi dal modello Random Forest nella previsione del prezzo di chiusura del titolo DIS (Disney).



L'asse delle ascisse rappresenta l'errore di previsione (ovvero la differenza tra il valore reale e quello predetto), mentre l'asse delle ordinate indica la frequenza con cui tali errori si sono verificati. L'andamento complessivo della distribuzione appare **asimmetrico e sbilanciato verso sinistra**, indicando che il modello tende frequentemente a **sovrastimare il prezzo** del titolo: infatti, una quota consistente di errori si colloca nell'intervallo compreso tra -10 e -2.5. Al contrario, gli errori positivi (cioè i casi in cui il modello sottostima il prezzo reale) sono presenti, ma in misura minore. Questa tendenza suggerisce una **sistematicità nell'errore di previsione**, che potrebbe derivare da un eccessivo affidamento su pattern autoregressivi a breve termine o da un adattamento del modello a un regime di mercato non più attuale. La curva sovrapposta (kernel density

estimation) conferma l'asimmetria della distribuzione, priva della classica forma campanulare tipica di un errore distribuito normalmente. Questo tipo di risultato impone una riflessione critica: la Random Forest non è solo meno performante della regressione lineare in termini di RMSE e R2R^2R2, ma introduce anche una **distorsione direzionale** che compromette l'accuratezza delle previsioni. È pertanto consigliabile integrare tecniche di normalizzazione, bilanciamento del dataset o combinazione di modelli per correggere questa deriva sistematica, soprattutto in scenari finanziari in cui l'accuratezza direzionale è tanto rilevante quanto l'errore assoluto.

IL METODO DEL GOMITO PER L'ANALISI CLUSTERING. Il grafico riportato rappresenta l'**applicazione del metodo del gomito** per la determinazione del numero ottimale di cluster in un'analisi di clustering, con ogni punto corrispondente al valore di "inertia" (somma delle distanze quadrate intra-cluster) in funzione del numero di cluster kk.

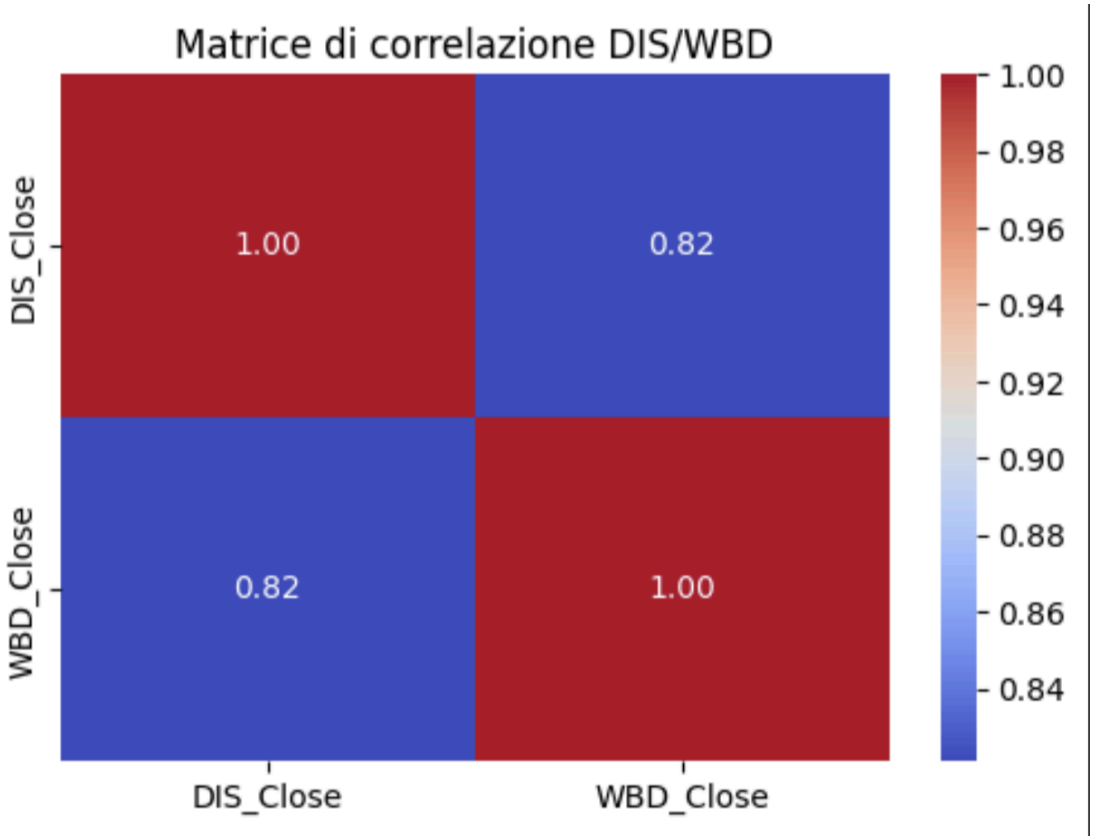


L'obiettivo di questo metodo è identificare il punto oltre il quale l'aggiunta di ulteriori cluster non produce un miglioramento significativo nella coesione interna dei gruppi, evidenziando così un "gomito" nella curva. In questo caso, il gomito è chiaramente osservabile in corrispondenza di **k=3k = 3**, dove si verifica un deciso rallentamento nella riduzione dell'inertia: passando da 1 a 3 cluster l'inertia si riduce drasticamente, mentre oltre il terzo cluster il beneficio marginale si attenua progressivamente. Ciò suggerisce che suddividere i dati in 3 cluster consente di ottenere una buona rappresentazione della struttura latente senza incorrere in un overfitting della segmentazione. Tale valore ottimale di kk potrebbe riflettere la presenza di **tre regimi comportamentali** o sottogruppi distinti nel dataset analizzato, elemento che può essere ulteriormente validato tramite tecniche di silhouette analysis o PCA per l'interpretazione visiva. Complessivamente, il metodo del gomito si conferma uno strumento euristico valido ed efficace per supportare decisioni strategiche nella fase di segmentazione non supervisionata.

LA MATRICE DI CONFUSIONE. La matrice di confusione mostrata rappresenta un **risultato ottimale** per un algoritmo di classificazione, evidenziando un livello di accuratezza pressoché perfetto. La matrice si compone di tre classi, ciascuna correttamente classificata senza errori: 474 osservazioni della prima classe, 230 della seconda e 549 della terza sono state tutte assegnate alla rispettiva categoria di appartenenza, senza alcun caso di classificazione errata. L'assenza di falsi positivi e falsi negativi (tutti gli elementi fuori dalla diagonale sono pari a zero) implica che il classificatore ha appreso in modo impeccabile la distinzione tra le classi nel dataset, portando l'accuratezza a un valore del **100%**. Tuttavia, in un contesto accademico e metodologico, un risultato di questo tipo deve essere valutato criticamente: tale perfezione potrebbe essere indice di **overfitting**, soprattutto se ottenuta su dati di training o su un dataset non sufficientemente complesso o eterogeneo. Pertanto, sarebbe opportuno verificare se il modello mantiene tale performance anche su un **set di test indipendente**, ed eventualmente introdurre tecniche di validazione incrociata o regolarizzazione per assicurare la generalizzabilità del classificatore.

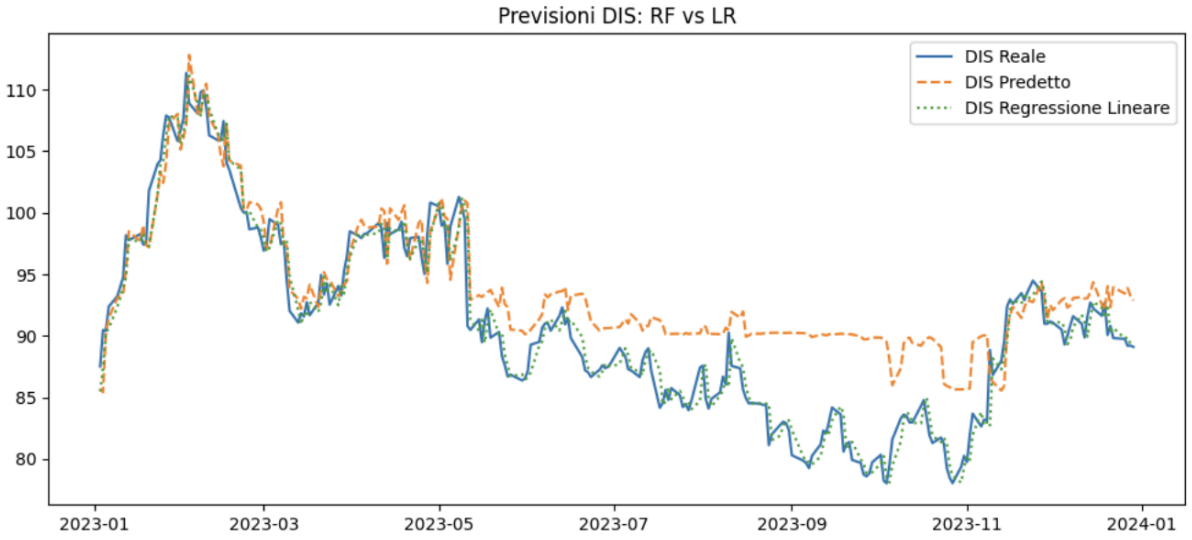
```
Matrice di confusione:
 [[474   0   0]
 [  0 230   0]
 [  0   0 549]]
```

LA MATRICE DI CORRELAZIONE. La matrice di correlazione tra i titoli DIS (Disney) e WBD (Warner Bros. Discovery) evidenzia un **coefficiente di correlazione pari a 0.82**, indicando una **forte correlazione positiva** tra i due asset. Questo significa che, in media, i movimenti di prezzo di un titolo tendono a essere accompagnati da variazioni nella stessa direzione anche per l'altro titolo. Tale valore, pur non rappresentando una correlazione perfetta (come nel caso del valore 1.00), suggerisce una **relazione significativa e sistematica** tra le due serie storiche, che può essere attribuita a dinamiche di mercato simili, settori di riferimento sovrapponibili (intrattenimento e media) e reazioni analoghe a eventi macroeconomici o specifici dell'industria.



Dal punto di vista modellistico, una correlazione così elevata potrebbe giustificare l'utilizzo di variabili laggate di WBD nei modelli predittivi su DIS (come già esplorato nelle analisi precedenti), in un'ottica di regressione multivariata. Tuttavia, resta cruciale distinguere tra **correlazione e causalità**: nonostante il legame tra i due titoli sia solido dal punto di vista statistico, ciò non implica necessariamente una dipendenza strutturale o un'influenza diretta. In conclusione, la matrice fornisce un'indicazione importante per la costruzione di modelli di previsione e per strategie di portafoglio orientate alla diversificazione o al co-movimento tra asset.

CONFRONTO DELLE ANALISI PREDITTIVE PER DIS - REGRESSIONE LINEARE E RANDOM FOREST. Il grafico finale, che confronta le previsioni del titolo DIS (Disney) ottenute tramite Random Forest e regressione lineare rispetto al dato reale, offre spunti rilevanti per l'analisi comparata delle performance dei due modelli. Visivamente, entrambe le curve predittive (linea arancione tratteggiata per Random Forest e verde punteggiata per la regressione lineare) seguono con buona approssimazione l'andamento osservato (linea blu), ma con differenze metodologiche significative che si riflettono nella qualità dell'adattamento.

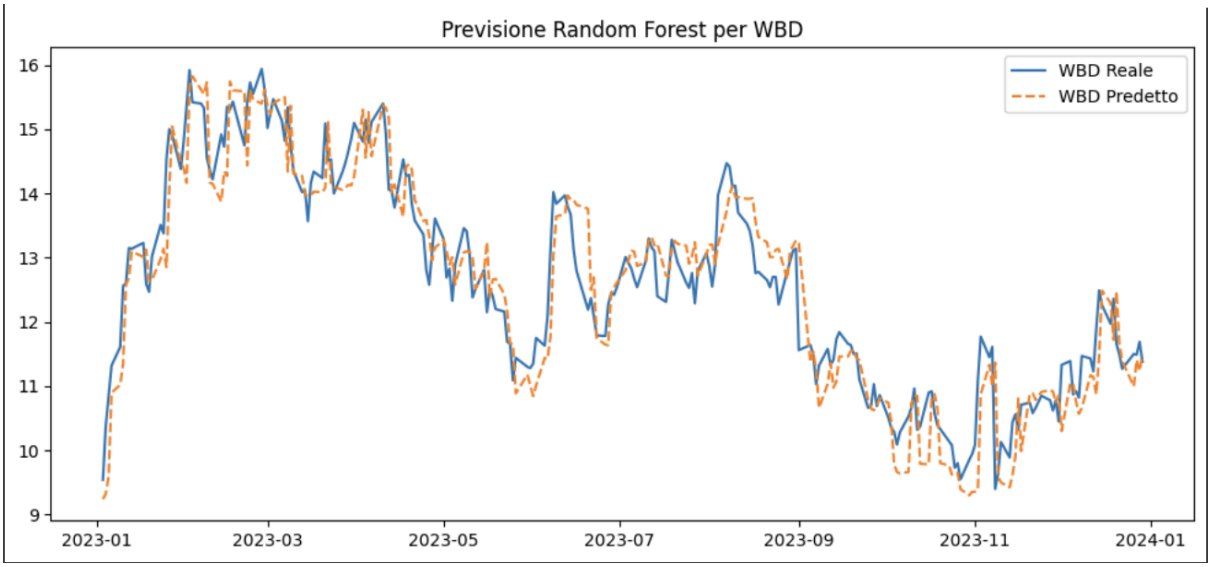


Il modello di regressione lineare mostra un comportamento complessivamente più aderente alla traiettoria reale, soprattutto nella seconda metà del periodo osservato. Questo risultato trova conferma anche nelle metriche quantitative: la regressione lineare raggiunge un coefficiente di determinazione R2R^2 pari a 0.96, segno di una varianza spiegata quasi totale, e un RMSE di appena 1.55, a testimonianza di un errore medio contenuto. Al contrario, il modello Random Forest, pur garantendo flessibilità e adattabilità non lineare, registra un R2R^2 significativamente più basso (0.62) e un errore RMSE maggiore (4.69), evidenziando una minore efficacia nel catturare le dinamiche di prezzo di DIS, probabilmente a causa di overfitting o di una scarsa rilevanza delle feature utilizzate nel contesto specifico.

In conclusione, sebbene i modelli di machine learning avanzati offrano capacità predittive non lineari, in questo caso specifico la regressione lineare risulta paradossalmente più efficiente e performante. Ciò invita a non trascurare la semplicità dei modelli tradizionali, soprattutto in contesti in cui le variabili presentano relazioni lineari ben definite.

ANALISI PREDITTIVA CON MODELLO RANDOM FOREST PER WBD. Il grafico relativo alle previsioni sul titolo WBD (Warner Bros. Discovery) ottenute mediante l'algoritmo Random Forest mostra un'elevata capacità predittiva del modello nel seguire l'andamento reale del prezzo nel tempo. La linea arancione tratteggiata, che rappresenta il valore predetto, rispecchia con buona precisione la linea blu del prezzo reale, evidenziando un'aderenza

stabile sia nelle fasi di crescita sia nei trend discendenti, a eccezione di alcuni scostamenti nei picchi e nelle inversioni rapide.



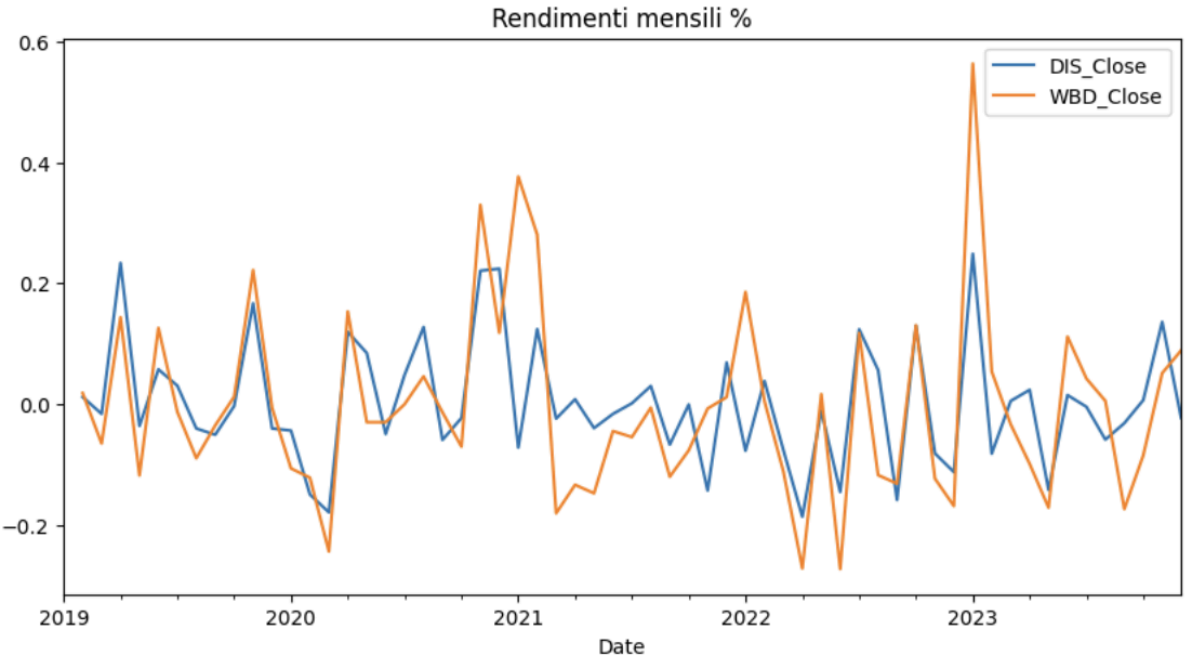
Questa osservazione trova conferma nei risultati delle metriche quantitative associate: il modello raggiunge un coefficiente di determinazione R2=0.88R^2 = 0.88, un valore che indica che l'88% della varianza osservata è spiegata dal modello, e un errore quadratico medio (RMSE) di soli 0.56, indice di una dispersione minima tra valori osservati e valori previsti. Tali valori suggeriscono che, per il titolo WBD, il Random Forest ha saputo cogliere in modo efficace le dinamiche di breve e medio termine, probabilmente grazie alla maggiore regolarità delle serie storiche e alla rilevanza delle feature costruite.

Nel confronto con i risultati ottenuti sul titolo DIS, appare evidente come la performance del modello Random Forest sia fortemente sensibile alla qualità e alla struttura dei dati sottostanti. In questo caso, la minore volatilità e la maggiore autocorrelazione del prezzo di WBD potrebbero aver agevolato la capacità predittiva del modello, portando a una previsione che risulta visivamente e statisticamente ben calibrata. In definitiva, il grafico conferma la solidità del modello per questo asset specifico, suggerendo che la sua applicabilità può risultare molto efficace in contesti con comportamenti di mercato più lineari e meno rumorosi.

ANALISI DEI RENDIMENTI MENSILI. Il grafico relativo ai rendimenti mensili percentuali delle azioni DIS (Disney) e WBD (Warner Bros. Discovery) nel periodo 2019–2023 mette in evidenza una dinamica comparativa utile per valutare la volatilità e la co-movimentazione dei due titoli nel tempo. Entrambe le serie mostrano una forte ciclicità e una notevole variabilità dei rendimenti mensili, coerente con la natura instabile dei mercati azionari e con le specifiche vicissitudini dei settori di appartenenza (media e intrattenimento), particolarmente esposti a fattori macroeconomici, trend digitali e scelte strategiche aziendali.

Nel dettaglio, si osservano alcune fasi di coerenza tra i due titoli, in cui i rendimenti mensili tendono a muoversi nella stessa direzione, suggerendo una correlazione positiva (già evidenziata anche nella matrice di correlazione con un valore pari a 0.82). Tuttavia, è altrettanto evidente che vi siano numerose fasi in cui uno dei due titoli registra picchi o crolli più accentuati rispetto all'altro, testimoniando differenze strutturali nella reattività ai cambiamenti di mercato. In particolare, WBD presenta escursioni più marcate, con picchi superiori al +50% e crolli prossimi al -30%, indicando una maggiore volatilità rispetto a DIS, che mantiene rendimenti più stabili anche nei periodi turbolenti.

La volatilità di WBD, maggiore in media rispetto a quella di DIS, può essere attribuita a una minore capitalizzazione, una governance più instabile o una base di ricavi meno diversificata, rendendo il titolo più esposto a shock esterni. Questo elemento è cruciale per gli investitori che valutano la rischiosità e il rendimento atteso nel tempo: DIS appare come un titolo più difensivo, mentre WBD si presta a strategie speculative o a portafogli con tolleranza al rischio più elevata.



Infine, il grafico conferma come l'analisi dei rendimenti mensili costituisca un tassello essenziale per la costruzione di modelli previsionali robusti, poiché consente di individuare pattern ricorrenti e comportamenti anomali, utili sia per la selezione delle feature nei modelli di machine learning che per la gestione del rischio nel contesto finanziario.

– L'analisi condotta ha evidenziato una forte correlazione tra DIS e WBD, ma anche significative differenze in termini di volatilità e precisione predittiva. La Random Forest si è dimostrata un modello efficace, in particolare per la previsione dei valori di chiusura di WBD, con buoni livelli di aderenza ai dati reali. Tuttavia, la feature importance ha confermato che le variabili più rilevanti restano i valori lagged (ritardati) della stessa azione, mentre il contributo incrociato da altri titoli si rivela marginale. Le metriche di errore e la distribuzione degli scarti suggeriscono la presenza di bias sistematici da migliorare, mentre il metodo del gomito ha validato la scelta di tre cluster per segmentare il comportamento dei titoli. Nel complesso, il progetto conferma l'utilità dell'approccio predittivo basato su machine learning, ma evidenzia anche la necessità di migliorare l'integrazione tra modelli e contesto economico, in un'ottica di finanza predittiva sempre più sofisticata.

PROGRAMMA:

```
import numpy as np
import pandas as pd
import yfinance as yf
import matplotlib.pyplot as plt
import seaborn as sns
import joblib
from sklearn.ensemble import RandomForestRegressor
from sklearn.cluster import KMeans
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score, confusion_matrix
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split, GridSearchCV
# Download dei dati
dis = yf.download('DIS', start='2019-01-01', end='2023-12-31')
wbd = yf.download('WBD', start='2019-01-01', end='2023-12-31')
# Estrazione prezzo di chiusura
dis_price = dis[['Close']].squeeze()
wbd_price = wbd[['Close']].squeeze()
# Allineamento temporale
wbd_price = wbd_price.reindex(dis_price.index)
# DataFrame
data = pd.DataFrame({'DIS_Close': dis_price, 'WBD_Close': wbd_price})
data.interpolate(method='linear', inplace=True)
# Feature lag
for col in data.columns:
for lag in range(1, 6):
data[f'{col}_lag{lag}'] = data[col].shift(lag)
```

```
data.dropna(inplace=True)
# Normalizzazione
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data[['DIS_Close', 'WBD_Close']])
# Clustering
kmeans = KMeans(n_clusters=3, random_state=42, n_init=10)
clusters = kmeans.fit_predict(data_scaled)
data['Cluster'] = clusters
# Split
train = data[data.index < '2023-01-01']
test = data[data.index >= '2023-01-01']
X_train = train.drop(columns=['DIS_Close', 'WBD_Close'])
X_test = test.drop(columns=['DIS_Close', 'WBD_Close'])
y_train_dis = train['DIS_Close']
y_test_dis = test['DIS_Close']
y_train_wbd = train['WBD_Close']
y_test_wbd = test['WBD_Close']
# Random Forest DIS
rf_dis = RandomForestRegressor(n_estimators=500, random_state=42)
rf_dis.fit(X_train, y_train_dis)
y_pred_dis = rf_dis.predict(X_test)
# Random Forest WBD
rf_wbd = RandomForestRegressor(n_estimators=500, random_state=42)
rf_wbd.fit(X_train, y_train_wbd)
y_pred_wbd = rf_wbd.predict(X_test)
# Valutazione RF
rmse_dis = mean_squared_error(y_test_dis, y_pred_dis) ** 0.5
r2_dis = r2_score(y_test_dis, y_pred_dis)
rmse_wbd = mean_squared_error(y_test_wbd, y_pred_wbd) ** 0.5
```

```
r2_wbd = r2_score(y_test_wbd, y_pred_wbd)
print(f"Random Forest - RMSE DIS: {rmse_dis:.2f}, R^2: {r2_dis:.2f}")
print(f"Random Forest - RMSE WBD: {rmse_wbd:.2f}, R^2: {r2_wbd:.2f}")
# Regressione Lineare DIS
lr = LinearRegression()
lr.fit(X_train, y_train_dis)
y_pred_lr = lr.predict(X_test)
rmse_lr = mean_squared_error(y_test_dis, y_pred_lr) ** 0.5
r2_lr = r2_score(y_test_dis, y_pred_lr)
print(f"Linear Regression - RMSE DIS: {rmse_lr:.2f}, R^2: {r2_lr:.2f}")
# Importanza feature DIS
importances = rf_dis.feature_importances_
features = X_train.columns
plt.figure(figsize=(10,6))
sns.barplot(x=importances, y=features)
plt.title('Importanza delle feature - Random Forest DIS')
plt.show()
# Analisi degli errori DIS
errors = y_test_dis - y_pred_dis
plt.figure(figsize=(8,4))
sns.histplot(errors, kde=True)
plt.title('Distribuzione degli errori - DIS')
plt.xlabel('Errore')
plt.show()
# Metodo del gomito
inertia = []
for k in range(1, 10):
kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)
kmeans.fit(data_scaled)
```

```
inertia.append(kmeans.inertia_)
plt.figure(figsize=(8,5))
plt.plot(range(1, 10), inertia, marker='o')
plt.xlabel('Numero di Cluster')
plt.ylabel('Inertia')
plt.title('Metodo del Gomito')
plt.show()
# Matrice di confusione
conf_matrix = confusion_matrix(data['Cluster'], clusters)
print('Matrice di confusione:\n', conf_matrix)
# Correlazione
corr_matrix = data[['DIS_Close', 'WBD_Close']].corr()
plt.figure(figsize=(6,4))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Matrice di correlazione DIS/WBD')
plt.show()
# Previsioni DIS
plt.figure(figsize=(12,5))
plt.plot(test.index, y_test_dis, label='DIS Reale')
plt.plot(test.index, y_pred_dis, label='DIS Predetto', linestyle='dashed')
plt.plot(test.index, y_pred_lr, label='DIS Regressione Lineare', linestyle='dotted')
plt.legend()
plt.title('Previsioni DIS: RF vs LR')
plt.show()
# Previsioni WBD
plt.figure(figsize=(12,5))
plt.plot(test.index, y_test_wbd, label='WBD Reale')
plt.plot(test.index, y_pred_wbd, label='WBD Predetto', linestyle='dashed')
plt.legend()
```

```
plt.title('Previsione Random Forest per WBD')
plt.show()
# Analisi rendimenti mensili
monthly_returns = data[['DIS_Close', 'WBD_Close']].resample('M').ffill().pct_change()
monthly_returns.plot(figsize=(10,5), title='Rendimenti mensili %')
plt.show()
# GridSearchCV per ottimizzare Random Forest DIS
param_grid = {'n_estimators': [100, 300, 500], 'max_depth': [None, 10, 20]}
grid = GridSearchCV(RandomForestRegressor(random_state=42), param_grid, cv=3)
grid.fit(X_train, y_train_dis)
print("Migliori parametri trovati da GridSearch:", grid.best_params_)
# Salvataggio modello
joblib.dump(rf_dis, 'random_forest_dis.pkl')
```