|  |
| --- |
|  |
| BUSINESS INTELLIGENCE PER I SERVIZI FINANZIARI |
| ANNO SCOLASTICO 2024-2025 |
|  |
|  |

Nome e Cognome: Alessandro Rocchi

Matricola: 899810

Anno Accademico: 2024/2025

SOMMARIO

[BUSINESS INTELLIGENCE PER I SERVIZI FINANZIARI 1](#_Toc199270030)

[SOMMARIO 2](#_Toc199270032)

[executive summary 3](#_Toc199270033)

[Introduzione al problema 4](#_Toc199270034)

[Descrizione dati 4](#_Toc199270035)

[Filtraggio dataset 6](#_Toc199270036)

[Analisi esplorativa 8](#_Toc199270037)

[Risultati del progetto 12](#_Toc199270038)

[IDENTIFICAZIONE DELLA CONFIGURAZIONE OTTIMALE DELL’ALGORITMO K-NN 12](#_Toc199270039)

[IMPORTAZIONE DEI MODULI NECESSARI 12](#_Toc199270040)

[PREPARAZIONE DEI DATI E SUDDIVISIONE DEL DATASET 12](#_Toc199270041)

[FASE DI ADDESTRAMENTO, PREDIZIONE E VALUTAZIONE DEL MODELLO 13](#_Toc199270042)

[OTTIMIZZAZIONE DEL PARAMETRO K CON CROSS-VALIDATION 14](#_Toc199270043)

[RISULTATI DEI VALORI DI MSE E RMSE: CROSS-VALIDATION 15](#_Toc199270044)

[VALORE DI K OTTIMALE: ELBOW METHOD 15](#_Toc199270045)

[CONFIGURAZIONE OTTIMALE: METODO GRID SEARCH 16](#_Toc199270046)

[FILLING DELLA MATRICE DI RATING CON LA CONFIGURAZIONE OTTIMALE 17](#_Toc199270047)

[ALLENAMENTO DELL’ALGORITMO SULL’INTERO DATASET CON LA CONFIGURAZIONE OTTIMALE 17](#_Toc199270048)

[CREAZIONE DEL RATING\_MATRIX\_ORIGINALE 18](#_Toc199270049)

[PREDIZIONE DEL RATING DI TUTTI I PRODOTTI PER UN USER SPECIFICO 19](#_Toc199270050)

[TOP 10 PRODOTTI CONSIGLIATI PER L’UTENTE “AGTMZCWIWBH43TCW7UKG2YV2EKKA” 20](#_Toc199270051)

[FILLING DELLA MATRICE DEI RATING CON LA CONFIGURAZIONE OTTIMALE 21](#_Toc199270052)

[SEGMENTAZIONE DEGLI UTENTI IN BASE ALLE PREFERENZE: ALGORITMO DI CLUSTERING K-MEANS CON COSINE SIMILARITY 23](#_Toc199270053)

[CREAZIONE DELLA LISTA PERSONALIZZATA DI RACCOMANDAZIONI PER CIASCUN UTENTE 26](#_Toc199270054)

[FILLING MATRICE DEI RATING: MATRIX FACTORIZATION 27](#_Toc199270055)

[CONFIGURAZIONE OTTIMALE DELL’ALGORITMO SVD: METODO GRID SEARCH 28](#_Toc199270056)

[CONFRONTO DEGLI ERRORI MSE E RMSE: MODELLI OTTIMIZZATI K-NN E SVD 29](#_Toc199270057)

[TOP 10 PRODOTTI CONSIGLIATI PER L’UTENTE “AGTMZCWIWBH43TCW7UKG2YV2EKKA” 30](#_Toc199270058)

[FILLING DELLA MATRICE DI RATING: MATRIX FACTORIZATION 30](#_Toc199270059)

[Conclusioni 32](#_Toc199270060)

# sommario dei dati utilizzati

Questo capitolo è mirato a presentare i dati finanziari utilizzati per svolgere il progetto, motivando la loro scelta, le metodologie di acquisizione e preparazione dei dati.

TITOLI FINANZIARI

Sono stati selezionati sei titoli azionari facenti parte dell’indice S&P500, distribuiti in tre settori economici distinti: Tecnologico, Automobilistico e Finanziario.

1. Settore Tecnologico

* Apple Inc. (AAPL)
  + Descrizione: Gigante tecnologico globale, leader nella produzione di elettronica di consumo, software e servizi.
  + Motivazione della Scelta: La decisione di analizzare AAPL è strategica per esaminare un’azienda “blue chip” ben affermata. Nonostante la sua solidità, AAPL è costantemente soggetta a fattori globali come le tensioni commerciali (ad esempio, i recenti ritardi sui dazi imposti dall’UE e che hanno avuto un effetto positivo sul valore del titolo) e le vendite in mercati chiave (come il rallentamento in Cina e il declino nel settore dei “Micro Computers” messo in luce da recenti rapporti). La sua performance mostra un equilibrio tra innovazione e la gestione di un ecosistema di prodotti e servizi consolidato. La sua inclusione offre l’opportunità di analizzare come il mercato reagisce a notizie sia di settore che macroeconomiche per un gigante con una forte capitalizzazione e lealtà del marchio.
* NVIDIA Corporation (NVDA)
  + Descrizione: Leader mondiale nella progettazione di processori grafici (GPU) e chip per l’intelligenza artificiale, i data center e il gaming.
  + Motivazione della Scelta: NVDA rappresenta il dinamismo e l’alta crescita del settore tecnologico, trovandosi al centro dell’avanzamento dell’intelligenza Artificiale. Negli ultimi anni, il suo valore azionario ha subito un aumento notevole, evidenziando una forte reattività alla richiesta di chip AI. Tuttavia, l’azienda si trova anche ad affrontare rischi legati alla geopolitica, come le limitazioni alle esportazioni verso la Cina, che possono avere un impatto significativo sulle sue previsioni di fatturato. La sua forte volatilità e la sua funzione cruciale nell’ambito dell’AI la rendono un ottimo esempio per l’analisi predittiva in un contesto di rapido cambiamento tecnologico e speculazione di mercato.

1. Settore Automobilistico

* Tesla Inc. (TSLA)
  + Descrizione: Produttore leader di veicoli elettrici (EV), batterie, soluzioni per l’energia e tecnologie di guida autonoma.
  + Motivazione della Scelta: Tesla è stata selezionata come simbolo di innovazione e crescita nel settore automobilistico, guidando il mercato delle auto elettriche e spingendo oltre i limiti della tecnologia. Le sue prestazioni sono caratterizzate da notevole instabilità, soggetta a fattori come l’aumento della concorrenza nel mercato delle auto elettriche, le cifre di produzione e consegna, le politiche di prezzo aggressive (riduzioni di prezzo) e i progressi delle sue tecnologie come il Full Self-Driving. Il valore delle azioni è anche molto influenzato dall’umore del mercato e dall’impatto di Elon Musk. Questa instabilità la rende particolarmente adatta per testare modelli predittivi in un contesto di mercato estremamente dinamico e in continua trasformazione.
* General Motors Company (GM)
  + Descrizione: Tradizionale casa automobilistica statunitense, produttrice di veicoli a combustione interna e con una crescente enfasi sui veicoli elettrici e la guida autonoma.
  + Motivazione della Scelta: GM è stata inclusa per offrire una visione alternata rispetto a Tesla. Essa è una delle case automobilistiche tradizionali che si trova a fronteggiare una complicata transizione verso i veicoli elettrici, necessitando di notevoli investimenti e di gestire contemporaneamente l’attività classica. I suoi risultati sono influenzati da elementi come le trattative sindacali, le vendite di auto con motori a combustione interna e le ripercussioni delle politiche commerciali, come i possibili dazi. Anche se non ha registrato la stessa crescita rapida di Tesla in termini di valore assoluto, la sua relativa stabilità e le difficoltà di adattamento del settore la rendono un esempio significativo per comprendere le dinamiche di un’industria in notevole cambiamento.

1. Settore Finanziario

* JPMorgan Chase & Co. (JPM)
  + Descrizione: Una delle più grandi e diversificate istituzioni bancarie e di servizi finanziari al mondo, con operazioni che spaziano dal costumer banking all’investment banking.
  + Motivazione della Scelta: JPMorgan Chase è stata scelta per la sua solidità e per la sua rappresentatività nel panorama bancario tradizionale. Questa banca ha costantemente mostrato ottimi risultati finanziari (ad esempio, ha superato le previsioni nel primo trimestre del 2025 con un incremento dei ricavi e dell’utile per azione), anche in scenari macroeconomici poco chiari, grazie alla sua abilità di diversificazione e a una gestione oculata. Le sue performance sono fortemente influenzate dall’andamento dei tassi d’interesse, dalla salute dell’economia mondiale e dalle normative nel settore bancario. L’analisi di JPM offre l’opportunità di esaminare la resilienza e le dinamiche di una banca “troppo grande per fallire” in relazione a fattori economici generali.
* Wells Fargo & Company (WFC)
  + Descrizione: Un’altra delle principali banche statunitensi, con una forte enfasi sul consumer banking e sul mercato immobiliare.
  + Motivazione della Scelta: Wells Fargo presenta un confronto interessante con JPMorgan Chase. Dopo aver superato un periodo di controversie e importanti ristrutturazioni, l’azienda sta ora vivendo una fase di recupero. Le sue prestazioni sono fortemente condizionate dal mercato dei prestiti e dai tassi d’interesse, ma anche dalla fiducia degli investitori e dalle ripercussioni delle passate questioni legali. L’ultimo aumento del valore del titolo riflette un moderato ottimismo per il suo processo di recupero e il contesto dei tassi. La sua presenza consente di esaminare come eventi specifici (come le controversie e le riforme seguenti) influenzino la volatilità e i profitti di un’azione finanziaria, offrendo anche una visione alternativa sulla dipendenza dai tassi d’interesse.

FUNZIONI UTILIZZATE PER SCARICARE I DATI

I dati storici sui diversi tipi di prezzi (Open, Close, Low, High, Volume) per ciascun titolo sono stati scaricati utilizzando la libreria Python “**yfinance**”, che permette un accesso semplice ed efficiente ai dati di Yahoo! Finance.

Inoltre, abbiamo impostato le date d’inizio e fine delle serie storiche, che comprendono le date dal 31.5.2014 – 31.5.2024 (Periodo di 10 anni).

Una volta scaricati i dati storici di ciascun’azione, sono stati filtrati per mantenere solo i **prezzi di chiusura giornalieri**, che è di nostro particolare interesse per lo svolgimento dell’intero progetto. Tutti gli altri prezzi sono stati droppati.

Visto che tutti i Dataframe presentavano **colonne multi-indice** (Price e Ticker), per una migliore manipolazione dei dati abbiamo deciso di rimuovere l’indice di colonna “Ticker”, che non aggiunge informazione utile.

FUSIONE DEI DATI

Una volta preparati i Dataframe di tutti i titoli abbiamo:

Innanzitutto, abbiamo creato un dizionario contenente ogni azione per un accesso più veloce e agevolato di tutte le serie di prezzi per ogni titolo, rispettando la struttura Key-Value: **Ticker – Serie di prezzi di chiusura.**

È stata utilizzata la funzione “*concat*” della libreria di Pandas per mettere assieme tutte le serie in un unico DataFrame, iterando su ogni coppia di Ticker-Serie di prezzi di chiusura del dizionario sopra menzionato. In questo modo, abbiamo creato un unico DataFrame contenente ogni serie di prezzi di chiusura per ogni azione.

Immagine che contiene testo, schermata, Carattere, numero

Il contenuto generato dall'IA potrebbe non essere corretto.

PRESENTAZIONE DEI DATI

Dopo il download e la fusione dei prezzi di chiusura giornalieri di ogni titolo in un unico DataFrame, abbiamo plottato tutti i prezzi di chiusura, in modo tale da mostrare graficamente l’andamento durante l’intero periodo di analisi.

Immagine che contiene testo, Diagramma, linea, diagramma

Il contenuto generato dall'IA potrebbe non essere corretto.

# STATISTICHE DESCRITTIVE

Questo capitolo si concentra sull’analisi delle principali caratteristiche statistiche delle serie temporali relative ai prezzi di chiusura e ai rendimenti di titoli scelti. Attraverso l’analisi di misure sia univariate che multivariate, si punta ad acquisire una comprensione dettagliata dei profili di rischio e rendimento di ogni asset e delle loro interrelazioni, creando una base empirica robusta per i passi successivi nella modellazione predittiva.

CALCOLO RENDIMENTO CUMULATO E COMPOSTO ANNUO

Per valutare la performance complessiva di ciascun titolo sull’intero periodo dell’analisi, sono stati calcolati i rendimenti cumulati e i rendimenti composti annuali.

* RENDIMENTO CUMULATO

Il rendimento cumulato indica quanto è cresciuto in totale un investimento durante tutto il periodo considerato. Non tiene conto della durata.

Abbiamo ottenuto i seguenti risultati:

AAPL: 863.2216% 🡪

NVDA: 24413.0871%

TSLA: 1210.1367%

GM: 62.6196%

JPM: 375.6964%

WFC: 54.5285%

Possiamo osservare che al primo posto c’è NVDA con un rendimento cumulato stratosferico del 24413%. Questo evidenzia il boom dell’intelligenza Artificiale e del ruolo dominante di NVDA come fornitore chiave di GPU. Tale rendimento sottolinea il potenziale esplosivo di aziende che operano in settori ad altissima crescita e innovazione.

Ciò vuol dire che, se avessimo investito solo 100 euro in NVDA, al giorno d’oggi avremmo avuto 24513.0871 euro: Esempio lampante del potere della crescita composta su titoli ad alta performance in periodi prolungati.

Dallo stesso modo, abbiamo TSLA e AAPL con rendimenti cumulati anch’essi straordinari. Nel caso di TSLA riflette l’evoluzione dei veicoli elettrici e la sua rapida espansione in un settore automobilistico in trasformazione. Nel caso di AAPL, tale rendimento sottolinea la capacità di innovazione continua e la fedeltà del brand.

Dall’altro canto, abbiamo rendimenti cumulati da parte di GM e WFC che in confronto con altri titoli dello stesso settore hanno avuto un rendimento molto modesto, non così esplosivi. Questo evidenzia le sfide del settore automobilistico tradizionale, che sta affrontando una transizione costosa e complessa verso i veicoli elettrici. La crescita di GM è meno speculativa e più legata alla performance e alle dinamiche industriali tradizionali.

Dello stesso modo, il rendimento cumulato di WFC nettamente inferiore a JPM riflette le sfide specifiche affrontate da Wells Fargo, in particolare il lungo percorso di recupero da scandali passati e le restrizioni normative che hanno rallentato la sua crescita e il suo riprezzamento sul mercato.

* RENDIMENTO COMPOSTO ANNUO (CAGR)

Il rendimento CAGR misura il rendimento medio annuo composto: Indicatore di quanto è cresciuto l’investimento ogni anno, in media.

Il CAGR annualizza il rendimento cumulato sull’interno periodo. Questo rende il confronto tra investimenti piu’ intuitivo e comparabile con altri tassi di interesse o di crescita economica.

Sono stati ottenuti i seguenti risultati

AAPL: 25.42%

NVDA: 73.34%

TSLA: 29.34%

GM: 4.982%

JPM: 16.88%

WFC: 4.448%

I CAGR’s di NVDA, TSLA e AAPL continuano a spiccare in modo eccezionale. I risultati continuano a confermare la loro capacità di creare valore a un ritmo senza pari.

Per quanto riguarda il settore bancario, JPM mostra un CAGR con crescita solida, evidenziando la capacità di generare rendimenti significativi in un settore piu’ maturo. WFC, in contrapposizione, mostra un CAGR nettamente inferiore. Questo riconferma un’altra volta l’impatto che hanno avuto gli scandali in passato.

CALCOLO RENDIMENTI SEMPLICI E LOGARITMICI

* Immagine che contiene testo, Diagramma, linea, Carattere

  Il contenuto generato dall'IA potrebbe non essere corretto.Grafico Rendimenti AAPL

Possiamo osservare momenti nei quali i rendimenti hanno avuto una volatilità molto variabile:

Innanzitutto, possiamo evidenziare la presenza di picchi alti e bassi nei rendimenti.

* Picchi bassi:
  + *Inizio 2019: Calo pronunciato*
    - Evento: Il 2 gennaio 2019, Apple ha emesso un avviso sugli utili, riducendo le previsioni di fatturato di 9 miliardi di dollari, principalmente a causa della debole domanda di iPhone in Cina.
    - Motivo: Questo è stato il primo taglio delle previsioni di fatturato da parte di Apple dal 2007, portando a un calo del 10% del prezzo delle azioni e a una perdita di oltre 70 miliardi di dollari in valore di mercato.
  + *Marzo 2020: Picco più basso in assoluto*
    - Evento: Il 16 marzo 2020, le azioni di Apple sono scese del 13%, con una perdita di 156 miliardi di dollari in capitalizzazione di mercato.
    - Motivo: Questo calo è stato innescato dal crollo del mercato globale a causa della **pandemia di COVID-19**, con Apple che ha annunciato la chiusura di tutti i suoi negozi al di fuori della Cina.
  + *Settembre 2020: Calo del 8%*
    - Evento: Dopo aver raggiunto il traguardo di $2 trilioni di capitalizzazione di mercato ad agosto 2020, Apple ha subito un significativo calo all'inizio di settembre. Il 3 settembre, le azioni sono scese di circa l'8% in un solo giorno, segnando la peggiore giornata di trading per la società da marzo 2020.
    - Motivi: **Controversia con Epic Games**: Apple ha rimosso Fortnite dall'App Store dopo che Epic Games aveva implementato un sistema di pagamento alternativo, violando le politiche di Apple. Questa disputa ha sollevato preoccupazioni sulla posizione dominante di Apple nel mercato delle app. Inoltre, **Evento di lancio poco incisivo**: Il 15 settembre, Apple ha presentato nuovi prodotti, tra cui Apple Watch e iPad, ma l'assenza dell'iPhone 12 ha deluso le aspettative del mercato, contribuendo al calo delle azioni.
* Picchi alti:

#### Agosto 2020: Picco più alto in assoluto

* + - Evento: Il 19 agosto 2020, Apple è diventata la prima azienda statunitense a raggiungere una capitalizzazione di mercato di 2 trilioni di dollari.
    - Motivo: Questo traguardo è stato raggiunto grazie a una forte performance finanziaria, con le azioni che sono aumentate di oltre il 50% nel 2020, superando le aspettative di Wall Street nonostante la pandemia.

* Grafico Rendimenti NVDA

# Immagine che contiene testo, Diagramma, linea, Carattere Il contenuto generato dall'IA potrebbe non essere corretto.

* Picco alto:
  + *Fine 2016: Picco piu’ alto in assoluto*
    - Evento: Annuncio degli utili del terzo trimestre 2016
    - Motivo: NVDA ha mostrato un aumento notevole nei guadagni e nei profitti per azione, superando le previsioni degli esperi. Tale risultato è stato stimolato dalla crescente richiesta delle sue schede grafiche nel campo dell’AI e nei data center, oltre all’affermazione nel settore dei videogiochi.
* Picchi bassi:
  + *Fine 2018: Picco piu’ basso in assoluto*
    - Evento: Crollo del mercato delle criptovalute e tensioni commerciali USA-Cina.
    - Motivo: La richiesta di GPU per il mining di criptovalute ha subito un forte calo, creando un surplus di magazzino. Inoltre, le tensioni commerciali tra gli Stati Uniti e la Cina hanno aggravato l’incertezza nel campo della tecnologia, portando a un notevole ribasso del valore delle azioni di NVDA.
  + *Inizio 2020: Calo del 20% del valore delle azioni.*
    - Evento: Crollo del mercato azionario dovuto alla pandemia di Covid-19.
    - Motivo: La diffusione mondiale del COVID-19 ha provocato un crollo significativo nei mercati finanziari. NVDA ha visto un forte abbassamento del valore delle proprie azioni a causa dell’incertezza economica.
* Grafico Rendimenti TSLA

Immagine che contiene testo, Diagramma, linea, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

* Picchi alti:
  + *Dicembre 2019*
    - Evento: Nel dicembre 2019, le azioni Tesla hanno superato per la prima volta i $400, chiudendo a $404,40.
    - Motivo: Questo aumento è stato alimentato da una sorpresa positiva nei risultati del terzo trimestre, in cui Tesla ha riportato un profitto inaspettato, aumentando la fiducia degli investitori nella redditività dell'azienda.
  + *Gennaio – febbraio 2020*
    - Evento: All'inizio di febbraio 2020, le azioni Tesla sono aumentate del 36% in due giorni, raggiungendo quasi $969 per azione.
    - Motivo: Questo rally è stato attribuito a un "short squeeze", in cui gli investitori che avevano scommesso contro il titolo hanno dovuto ricomprare le azioni per coprire le loro posizioni, alimentando ulteriormente l'aumento del prezzo.
  + *Gennaio – aprile 2021*
    - Evento: Nel primo trimestre del 2021, Tesla ha consegnato 184.800 veicoli, superando le aspettative del mercato.
    - Motivo: Questo record di consegne ha portato a un aumento del prezzo delle azioni del 7% in un solo giorno, riflettendo la forte domanda per i veicoli elettrici di Tesla.
* Picchi bassi:
  + *Settembre 2018*
    - Evento: Nel settembre 2018, la SEC ha citato in giudizio Elon Musk per le sue dichiarazioni su Twitter riguardo alla privatizzazione di Tesla.
    - Motivo: La causa ha causato un calo significativo del prezzo delle azioni, poiché gli investitori temevano le implicazioni legali e la potenziale rimozione di Musk dalla leadership dell'azienda.
  + *Febbraio 2020*
    - Evento: A febbraio 2020, le azioni Tesla sono diminuite di quasi il 20% in un solo giorno.
    - Motivo: Il calo è stato causato dai ritardi nelle consegne del Model 3 in Cina a causa dell'epidemia di COVID-19, sollevando preoccupazioni sull'impatto della pandemia sulle operazioni globali di Tesla.
  + *Settembre 2020*
    - Evento: Nel settembre 2020, Tesla è stata esclusa dall'inclusione nell'indice S&P 500.
    - Motivo: L'esclusione ha deluso gli investitori, portando a un calo del 21,1% delle azioni in un solo giorno, segnando la più grande perdita giornaliera nella storia dell'azienda fino a quel momento.
* Grafico Rendimenti GM

Immagine che contiene testo, Diagramma, linea, schermata

Il contenuto generato dall'IA potrebbe non essere corretto.

* Picco più basso:
  + *Marzo 2020*
    - Evento: Crollo del mercato azionario dovuto alla pandemia di COVID-19.
    - Motivo: La diffusione globale del COVID-19 ha innescato una vendita massiccia nei mercati finanziari. GM ha subito un forte calo del prezzo delle azioni a causa dell’incertezza economica, della chiusura temporanea degli impianti di produzione e della diminuzione della domanda di veicoli.
* Picco piu’ alto:
  + *Novembre 2020*
    - Evento: Annuncio di un investimento significativo nei veicoli elettrici
    - Motivo: GM ha annunciato un aumento dell’investimento nei veicoli elettrici e autonomia a 27 miliardi di dollari entro il 2025, con l’obiettivo di lanciare 30 nuovi modelli di veicoli elettrici. Questo annuncio ha rafforzato la fiducia degli investitori nella strategia futura dell’azienda, portando a un aumento significativo del prezzo delle azioni.
* Grafico Rendimenti JPM e WFC

Immagine che contiene testo, Diagramma, linea, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.Immagine che contiene testo, schermata, Diagramma, linea

Il contenuto generato dall'IA potrebbe non essere corretto.

* Picco piu’ basso:
  + *Marzo 2020*
    - Evento: Crollo del mercato dovuto alla pandemia di COVID-19
    - Motivo: Sempre dovuto alla diffusione globale della pandemia anche JPMorgan ha subito un forte calo del prezzo delle azioni, raggiungendo un minimo di $76,91 il 19 marzo 2020, a causa dell’incertezza economica.
* Picco piu’ alto:
  + *Novembre 2020:*
    - Evento: Annuncio dell’efficacia del vaccino Pfizer-BioNTech contro il COVID-19.
    - Motivo: L’annuncio che il vaccino Pfizer-BioNTech era efficace al 90% ha portato a un’ondata di ottimismo nei mercati finanziari. Le azioni di JPMorgan hanno beneficiato di questa notizia, poiché’ gli investitori prevedevano una ripresa economica piu’ rapida e un miglioramento delle prospettive per il settore bancario.

1. Che cosa hanno in comune le serie storiche?

Le serie storiche dei rendimenti semplici e rendimenti logaritmici:

* **Mostrano un comportamento molto simile nel tempo**, con picchi e cadute quasi sovrapposti: Entrambi misurano il ritorno giornaliero e seguono lo stesso andamento dei prezzi.
* **Sono entrambi stazionari**, ovvero, **oscillano attorno a una media** (prossima a **zero**), senza tendenza crescente o decrescente chiara.
* La **volatilità non è costante**: Ci sono periodi in cui i rendimenti oscillano molto piu’ del normale.
* I rendimenti logaritmici tendono ad essere **piu’ simmetrici e matematicamente piu’ trattabili**, mentre quelli semplici sono piu’ intuitivi.

1. C’è una correlazione positiva fra società dello stesso settore?

Per calcolare la correlazione tra società dello stesso settore si hanno utilizzato i **rendimenti logaritmici giornalieri** in quanto tendono ad avere proprietà statistiche piu’ desiderabili.

Inoltre, i rendimenti dello stesso settore si sono uniti in un solo DataFrame per il calcolo della correlazione tramite la funzione **“.corr()”.**

Immagine che contiene testo, schermata, Carattere, numero

Il contenuto generato dall'IA potrebbe non essere corretto.Immagine che contiene testo, schermata, Carattere, numero

Il contenuto generato dall'IA potrebbe non essere corretto.Immagine che contiene testo, schermata, Carattere, numero

Il contenuto generato dall'IA potrebbe non essere corretto.

Immagine che contiene testo, Carattere, schermata, bianco

Il contenuto generato dall'IA potrebbe non essere corretto.Immagine che contiene testo, schermata, Carattere, bianco

Il contenuto generato dall'IA potrebbe non essere corretto.Immagine che contiene testo, Carattere, schermata, bianco

Il contenuto generato dall'IA potrebbe non essere corretto.

Possiamo osservare che tra AAPL e NVDA c’è una correlazione positiva moderata, il che ci fa capire che i loro rendimenti tendono a muoversi nella stessa direzione piu’ della metà delle volte, anche se non in perfetta sincronia. Era ragionevole aspettarsi una correlazione positiva tra le due perche’ sono entrambe due giganti del settore tecnologico, fortemente influenzate da tendenze macroeconomiche simili. Tuttavia, il fatto che la correlazione non sia estremamente alta indica che esistono anche fattori di crescita unici per ciascuna azienda. NVDA è fortemente legata al boom dell’AI, mentre AAPL, ha una struttura aziendale piu’ consolidata, incentrata sull’ecosistema di prodotti e servizi. Questa diversità nei fattori specifici impedisce una correlazione perfetta.

Per quanto riguarda la correlazione tra TSLA e GM, possiamo dire che è una correlazione positiva debole (0.33465). Indica che, anche se tendano a muoversi nella stessa direzione, il legame lineare tra di loro è relativamente debole. Questa correlazione conferma la diversità di queste due aziende nel settore automobilistico. Da un lato, TSLA si presenta come un vero e proprio innovatore nel mercato dei veicoli elettrici, con una valutazione alimentata dalla crescita e dall’innovazione, spesso influenzata da aspetti legati alla figura di Elon Musk. Dall’altro lato, invece, GM si presenta come un’azienda automobilistica classica che sta attraversando un passaggio massiccio verso l’elettrico, ma continua a dipendere enormemente dalle vendite di veicoli a motore tradizionale e da elementi come le trattative sindacali e il costo delle materie prime per la produzione convenzionale. Le loro differenti strategie, valutazioni e fattori di mercato giustificano la correlazione relativamente bassa, pur appartenendo allo stesso settore

Infine, abbiamo ottenuto una correlazione tra JPMorgan e Wells Fargo del 0.811769. Significa che i rendimenti tendono a muoversi nella stessa direzione in modo molto consistente e con un’entità simile. Entrambe sono estremamente sensibili agli stessi fattori macroeconomici: Tassi di interesse, crescita economica, politiche della Federal Reserve, fiducia dei consumatori e regolamentazione bancaria. Anche se Wells Fargo ha affrontato delle difficoltà particolari e un periodo di ripresa, le dinamiche del settore finanziario la colpiscono in modo simile. Se uno registra un calo, è molto probabile che anche l’altro ne risenta.

GRAFICI DIAGNOSTICI A 3 SEZIONI

Immagine che contiene diagramma, testo, Diagramma, linea

Il contenuto generato dall'IA potrebbe non essere corretto.Immagine che contiene Diagramma, diagramma, linea, schermata

Il contenuto generato dall'IA potrebbe non essere corretto.

Immagine che contiene Diagramma, linea, schermata, diagramma

Il contenuto generato dall'IA potrebbe non essere corretto.

Immagine che contiene diagramma, testo, Diagramma, schermata

Il contenuto generato dall'IA potrebbe non essere corretto.

1. Istogramma con KDE:

La distribuzione dei rendimenti è simmetrica e centrata attorno allo zero, il che è coerente con una distribuzione normale.

Tuttavia, si nota che la curva KDE è piu’ appuntita rispetto alla normale, che ha forma campanile.

Inoltre, le code sono piu’ pesanti rispetto alla normale; quindi, c’è piu’ probabilità di eventi estremi. In altre parole, evidenziamo una deviazione della normalità.

1. Boxplot

La maggior parte dei dati è concentrata attorno allo zero.

Sono presenti molti outliers sia nella coda sinistra (rendimenti negativi) sia nella destra (rendimenti positivi), ben oltre i baffi del boxplot.

Questo conferma la presenza di rendimenti estremi che si verificano piu’ spesso di quanto previsto da una distribuzione normale.

1. QQ-Plot

Il QQ-Plot confronta i quantili teorici di una distribuzione normale con quelli osservati.

I punti centrali sono vicini alla linea rossa, di conseguenza, la parte centrale della distribuzione è abbastanza normale.

Tuttavia, le code si allontanano dalla linea, sia in alto che in basso. Ciò indica che i rendimenti presentano una distribuzione leptocurtica, ovvero con code piu’ pesanti della normale.

In sostanza, i rendimenti NON sono perfettamente distribuiti normalmente:

* Hanno code piu’ **pesanti** 🡪 Piu’ **eventi estremi**
* Sono simmetrici ma **leptocurtici**

Sono presenti **numerosi outliers**, soprattutto nei giorni di elevata volatilità.

Questi comportamenti indicano che per modellare correttamente questi rendimenti bisognerebbe usare distribuzioni che tollerano code pesanti.

STATISTICHE DESCRITTIVE UNIVARIATE ANNUALIZZATE

Per questa sezione sono stati utilizzati i rendimenti logaritmici mensili rispetto a quelli giornalieri in quanto così riduciamo il rumore, offrendo un’analisi piu’ stabile e coerente per confronti tra titoli e settori.

Immagine che contiene testo, schermata, Carattere, numero

Il contenuto generato dall'IA potrebbe non essere corretto.

* Media Annuale: Misura il rendimento medio annuo atteso del titolo.

Notiamo che tra tutti i titoli quello con la media annuale piu’ alta è NVDA con +55.7%, un rendimento eccezionale. Seguendo con AAPL, TSLA e JPM con 22.4%, 24.33% e 15.32%, con rendimenti elevati. I piu’ bassi, dall’altra parte, sono stati WFC e GM con +4.1% e +4.7%, rendendoli meno redditizi rispetto gli altri.

* Varianza Annuale e Deviazione STD Annuale:

Notiamo che TSLA ha la massima volatilità di tutti ottenendo 56.7%, ciò lo rende un titolo altamente speculativo e volatile. Le segue NVDA, GM e WFC con 45.7%, 33.1% e 28.5% rispettivamente: Un rischio alto-moderato. Infine, abbiamo AAPL e JPM con 27.5% e 24.1% con minor volatilità, il che li rendono i titoli piu’ “sicuri” rispetto agli altri.

* Asimmetria:

Notiamo che TSLA è stato l’unico titolo che ha ottenuto un valore di asimmetria POSITIVO, ottenendo un +0.43, il che implica una coda lunga destra della distribuzione: Alti rendimenti occasionali.

Tutti gli altri hanno ottenuto un valore di asimmetria della distribuzione negativo, sinonimo della tendenza di rendimenti negativi occasionali: WFC al primo posto con -0.70 (Piu’ frequenza di rendimenti negativi occasionali) fino a AAPL con -0.23 (Frequenza moderata di rendimenti negativi occasionali)

* Curtosi: Calcola la frequenza di eventi estremi rispetto a una normale.

AAPL ha ottenuto una curtosi negativa del -0.43, rendendolo l’unico titolo con meno eventi estremi rispetto a una normale.

Tutti gli altri titoli hanno conseguito una curtosi positiva, indice di una frequenza di eventi estremi maggiore rispetto a una normale: Al primo posto troviamo a WFC con 2.82 di curtosi (Molta presenza di eventi estremi).

In conclusione,

AAPL ha un buon compromesso tra rendimento e rischio.

NVDA ha un elevato rendimento e rischio, piu’ attraente per investitori aggressivi.

TSLA ha la massima volatilità ma con un gran potenziale di crescita.

GM ha un basso rendimento per il rischio assunto.

JPM ha un rendimento discreto e rischio contenuto, abbastanza bilanciato.

WFC ha un’alta probabilità di eventi estremi negativi, il che lo rende rischioso e poco attraente.

MATRICE VARIANZE/COVARIANZE E DI CORRELAZIONE

* Matrice di Varianze/Covarianze

Immagine che contiene testo, schermata, Carattere, numero

Il contenuto generato dall'IA potrebbe non essere corretto.

Dalla matrice di varianze e covarianze, la diagonale rappresenta le varianze dei singoli titoli e le celle fuori la diagonale rappresentano le covarianze tra i titoli.

Come abbiamo detto prima, la varianza è strettamente collegata al rischio e volatilità dei titoli e in questo caso, come abbiamo visto prima, **TSLA** ha la **varianza assoluta piu’ alta** con 0.0268, il che lo rende il titolo piu’ rischioso.

Dall’altra parte abbiamo **JPM** con la **varianza piu’ bassa in assoluto** del 0.0048, il che lo rende il titolo meno rischioso di tutti.

Per quanto riguardano le covarianze, sappiamo che misurano la tendenza di due titoli a muoversi insieme in termini assoluti.

In questo caso, abbiamo le seguenti coppie:

* TSLA – NVDA con una covarianza del 0.007570, il che si tratta di una covarianza forte: Si muovono fortemente insieme.
* TSLA – AAPL con una covarianza del 0.006770, anch’essa molto forte.
* NVDA – GM con una covarianza del 0.005196.

Questi titoli tendono a muoversi insieme con intensità e hanno un’alta varianza (alta volatilità e quindi rischio): Aumenta definitivamente il rischio complessivo del portafoglio.

* JPM – WFC con una covarianza del 0.004589, appartenente al top 3 titoli piu’ correlati tra di loro.
* AAPL – JPM hanno una covarianza del 0.001700: Un valore molto basso così come AAPL – WFC con una covarianza del 0.001657, anch’essa molto bassa.

Questi ultimi titoli sono buoni candidati per **diversificare** il portfolio: Covarianza positiva e movimenti deboli insieme.

* Matrice di Correlazione

Immagine che contiene testo, schermata, Carattere, numero

Il contenuto generato dall'IA potrebbe non essere corretto.

Possiamo osservare che i titoli piu’ correlati sono stati:

1. JPM – WFC con una correlazione del **0.7978:** Correlazione forte.
2. JPM – GM con una correlazione del 0.6661.
3. WFC – GM con una correlazione del 0.6176.

Visto che JPM e WFC appartengono al settore bancario, è coerente ottenere un risultato del genere, ovvero, che siano altamente correlate: Sono influenzate da variabili macro-simili.

Per quanto riguarda GM, appartenente al settore automobilistico e i titoli del settore bancario, hanno una correlazione decisamente alta, questo dovuto a che le vendite di auto dipendono fortemente dai finanziamenti al consumo. Se le banche vanno bene significa che i tassi di interesse sono favorevoli, il credito è disponibile e i consumatori sono propensi ad acquistare auto, quindi, ha senso che abbiano una correlazione alta.

GRAFICO ANDAMENTO DELLE CORRELAZIONI FRA TITOLI E SCATTERPLOTS

* AAPL – NVDA e S&P500

Immagine che contiene testo, schermata, diagramma

Il contenuto generato dall'IA potrebbe non essere corretto.

Immagine che contiene testo, linea, Diagramma, diagramma

Il contenuto generato dall'IA potrebbe non essere corretto.

Immagine che contiene testo, schermata

Il contenuto generato dall'IA potrebbe non essere corretto.

Immagine che contiene testo, Diagramma, diagramma, linea

Il contenuto generato dall'IA potrebbe non essere corretto.

Immagine che contiene testo, schermata, diagramma

Il contenuto generato dall'IA potrebbe non essere corretto.

La correlazione mobile tra i titoli tecnologici e l’indice di mercato S&P500 (basata su rendimenti logaritmici mensili con finestra di 6 mesi):

* Sia AAPL che NVDA mostrano una correlazione prevalentemente positiva e spesso elevata con l’indice S&P500. Questo e’ un comportamento atteso per titoli di grande capitalizzazione, soprattutto quelli che fanno parte dell’indice stesso e che ne influenzano pesantemente la performance. Significa che nella maggior parte dei periodi, quando il mercato complessivo sale, anche questi titoli tendono a salir, e viceversa.
* Ci sono differenze nella stabilita’ e volatilita’ delle correlazioni:
  + La correlazione tra AAPL – SP500 e’ generalmente piu’ stabile e costantemente elevata rispetto a quella di NVDA. Ci sono periodi in cui si avvicina a 1.0 e raramente scende in territorio negativo in modo significativo. Questo riflette il ruolo di AAPL come “blue chip” consolidata e la sua forte influenza sul mercato, rendendola un titolo che tende a muoversi molto in linea con l’andamento generale.
  + La correlazione tra NVDA – SP500 è nettamente piu’ volatile e imprevedibile. Anche se raggiunge picchi di correlazione elevati simili a quelli di AAPL, mostra anche cadute brusche e periodi in cui la correlazione diventa significativamente negativa (Fine 2017/inizio 2018, metà 2021, metà 2023). Questo suggerisce che, pur essendo un gigante tech, NVDA è piu’ suscettibile a cicli di settore specifici (semiconduttori, AI) e a sentiment di mercato che a volte possono divergere o addirittura contrapporsi a quelli dell’indice generale.
* In sintesi, il grafico rivela che entrambi i titoli sono molto sensibili all’andamento del mercato, ma con gradi di stabilità diversi nella loro relazione. AAPL segue la marea del mercato in modo piu’ prevedibile, mentre NVDA, pur essendo anch’essa un gigante tech, ha dei motori interni che a volte le permettono di navigare con maggiore indipendenza o persino controcorrente rispetto al mercato generale.

La correlazione mobile tra AAPL e NVDA (basata su rendimenti logaritmici mensili con finestra di 6 mesi):

* Mostra **instabilità nel tempo**: Oscilla frequentemente tra valori positivi e negativi. Durante l’intero periodo la loro correlazione è stata prevalentemente positiva.
* Si è evidenziato la correlazione piu’ bassa in assoluto nell’intero periodo a partire da un calo nella correlazione tra i titoli passando **da un 0.75 fino a un -0.75** nel giro di qualche mese.
* Nella maggior parte del tempo la correlazione oscillava tra valori positivi; quindi, **prevalentemente hanno avuto una correlazione positiva** a parte qualche eccezione.
* Queste variazioni indicano che la relazione tra le due azioni **non è costante** (**Correlazioni volatili**), anche se appartengono allo stesso settore. Ciò può essere dovuto a cambiamenti nelle condizioni macroeconomiche o nei tassi d’interesse, dinamiche aziendali interne, ecc.

Come cambiano le correlazioni in funzione dei rendimenti? La dispersione dei punti nello scatterplot conferma o no la relazione lineare fra i due rendimenti?

* I punti dello scatterplot formano una nuvola che si estende dal quadrante in basso a sinistra e in alto a destra, indicando una correlazione positiva, confermando visivamente il coefficiente di correlazione positivo di 0.551793 che si è ottenuto precedentemente. Ciò significa che, quando i rendimenti di un titolo sono positivi, anche quelli dell’altro tendono ad essere positivi, e viceversa.
* I punti non sono perfettamente allineati su una retta, ma sono piuttosto dispersi attorno a una tendenza centrale. Questo è coerente con un coefficiente di correlazione del 0.55, che è moderatamente forte ma non indica una dipendenza perfetta. Quindi, anche se tendano a muoversi spesso nella stessa direzione, l’entità dei loro movimenti non è sempre proporzionale e ci sono giorni in cui uno si muove significativamente di piu’ o meno dell’altro.
* Infine, si possono evidenziare alcuni punti **outlier** molto lontani dalla parte centrale (Nuvola principale) che indicano giorni in cui si hanno avuto rendimenti insoliti rispetto alla loro relazione tipica.
* TSLA – GM e S&P500

Immagine che contiene testo, schermata

Il contenuto generato dall'IA potrebbe non essere corretto.Immagine che contiene testo, linea, Diagramma, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

Immagine che contiene testo, schermata, diagramma

Il contenuto generato dall'IA potrebbe non essere corretto.Immagine che contiene testo, schermata, diagramma

Il contenuto generato dall'IA potrebbe non essere corretto.

Immagine che contiene testo, linea, Diagramma, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

La correlazione mobile tra i titoli automobilistici e l’indice di mercato S&P500 (basata su rendimenti logaritmici mensili con finestra di 6 mesi):

* La correlazione di Tesla con l’S&P500 è estremamente volatile e dinamica, oscillando frequentemente tra valori molto alti (vicini a +1) e profondamente negativi (fino a circa -0.8).
* Ci sono periodi in cui Tesla si muove in quasi perfetta sincronia con il mercato (Picchi ~+1.0 nel 2015, fine 2019, 2020-2021, fine 2023). Questi possono corrispondere a periodi di forte sentiment di mercato verso la tecnologia o a shock macroeconomici che colpiscono tutti.
* Tuttavia, sono molto evidenti i periodi di correlazione negativa marcata (Fine 2017-inizio 2018, metà 2019, fine 2021-inizio 2022, fine 2023). Questi indicano che per finestre di 6 mesi, Tesla ha spesso avuto rendimenti che andavano in direzione opposta a quelli del mercato generale. Questo riflette la natura dirompente di Tesla: La sua indipendenza da specifici annunci di prodotto, problemi di produzione, e il sentiment altamente speculativo attorno al titolo, che possono disaccoppiarla dal mercato piu’ ampio.
* Anche la correlazione di GM con l’indice S&P500 è volatile, ma con oscillazioni negative meno estreme di quelle di Tesla.
* Generalmente, tende a rimanere in territorio positivo, spesso con valori moderati o forti (tra 0.4 e 0.8), indicando che GM, in quanto azienda industriale matura e parte integrante dell’economia tradizionale, tende a seguire l’andamento del mercato generale.

La correlazione mobile tra TSLA e GM (basata su rendimenti logaritmici mensili con finestra di 6 mesi):

* La correlazione è estremamente dinamica e oscillante, passando frequentemente tra valori alti e negativi significativi.
* Questo pattern conferma la fondamentale differenza tra Tesla (Innovatore, alta crescita, volatile) e General Motors (Azienda in transizione, piu’ stabile ma con sfide significative, tradizionale). Le loro risposte a condizioni economiche possono essere molto divergenti, portando a periodi in cui i loro rendimenti si muovono in direzioni opposte o indipendentemente.

Come cambiano le correlazioni in funzione dei rendimenti?

* I grafici dimostrano che le correlazioni non sono fisse ma estremamente dinamiche (Non staticità e volatilità). Variazioni significative si verificano nel corso del tempo, a volte molto brusche.
* Le correlazioni cambiano a seconda dell’impatto dei fattori specifici e quelli di mercato: Quando dominano fattori specifici dell’azienda (Es. Un annuncio di prodotto di Tesla), la correlazione con il mercato può diminuire. Invece, quando dominano fattori macroeconomici o di sentiment generale di mercato (es. Variazioni tassi di interesse, recessioni, forte crescita economica), la correlazione con l’s&p500 tende ad aumentare.
* In piu’, abbiamo anche i “Periodi di Stress”, nel quali a partire da forti stress di mercato (es. Crolli o recessioni), le correlazioni tra la maggior parte degli asset, con il mercato, tendono ad aumentare per via del contagio. Tutti gli asset tendono a muoversi verso il basso insieme. I picchi di correlazione elevata potrebbero coincidere con tali periodi.

La dispersione dei punti nello scatterplot conferma o no la relazione lineare fra i due rendimenti?

* S&P500 e TSLA
  + La nuvola di punti presenta una tendenza generale positiva, ma è molto dispersa. I punti sono sparsi su un’ampia area, con molti outlier lontano dalla massa centrale.
  + Conferma l’esistenza di una relazione lineare positiva, ma questa relazione è debole/moderata e altamente volatile. La dispersione evidenzia che, anche se TSLA tenda a muoversi nella stessa direzione dell’S&P500, lo fa con grande variabilità nell’entità del movimento. Non è una relazione stretta.
* S&P500 e GM
  + La nuvola di punti mostra una tendenza positiva piu’ evidente rispetto a quella di TSLA, e i punti sono leggermente piu’ raggruppati attorno a una linea centrale.
  + Anche in questo caso si conferma l’esistenza di una relazione lineare positiva moderata. La dispersione è inferiore a quella di TSLA VS SP500.
* TSLA e GM
  + La nuvola di punti è molto dispersa e quasi circolare, con una leggera tendenza positiva. I punti si estendono ampiamente in tutti e quattro i quadranti.
  + Si conferma una relazione lineare positiva ma molto DEBOLE. La vasta dispersione indica che i rendimenti giornalieri di TSLA e GM si muovono spesso in modo indipendente, anche se a volte si muovono nella stessa direzione.
* Immagine che contiene testo, Diagramma, linea, diagramma

  Il contenuto generato dall'IA potrebbe non essere corretto.Immagine che contiene testo, linea, Diagramma, Carattere

  Il contenuto generato dall'IA potrebbe non essere corretto.JPM - WFC e S&P500

Immagine che contiene testo, schermata, diagramma

Il contenuto generato dall'IA potrebbe non essere corretto.Immagine che contiene testo, schermata

Il contenuto generato dall'IA potrebbe non essere corretto.Immagine che contiene testo, schermata, diagramma

Il contenuto generato dall'IA potrebbe non essere corretto.

La correlazione mobile tra i titoli finanziario e l’indice di mercato S&P500 (basata su rendimenti logaritmici mensili con finestra di 6 mesi):

* Sia JPM che WFC mostrano una correlazione costantemente positiva e generalmente alta con l’indice SP500 per la maggior parte del periodo. Questo e’ atteso per grandi istituzioni finanziarie che sono fortemente sensibili alle condizioni macroeconomiche e alle politiche monetarie che guidano il mercato generale.
* Si notano pero’ delle fluttuazioni significative. Entrambe le correlazioni tendono a mostrare dei cali significativi vicino allo zero (es. fine 2017, meta’ 2019, fine 2021, meta’ 2023 per JPM; fine 2015, inizio 2018 per WFC). Questi cali possono indicare periodi in cui il settore bancario e’ influenzato da fattori specifici (es. Preoccupazioni sui tassi di interesse, rischio di credito o eventi specifici delle banche) che lo disaccoppiano temporaneamente dall’andamento piu’ ampio del mercato.
* Ci sono picchi di correlazione evidenti, molto vicini a 0.9 e 1.0, soprattutto durante i periodi di forte stress di mercato (es. inizi 2020 con la pandemia) dove tutti gli asset tendono a muoversi all’uniscono con il mercato.

La correlazione mobile tra JPM e WFC (basata su rendimenti logaritmici mensili con finestra di 6 mesi):

* L’andamento della correlazione tra questi titoli dello stesso settore mostra una correlazione estremamente elevata e relativamente stabile, quasi costantemente superiore a 0.8 e spesso vicina a 0.9 o 1.0.
* Anche se ci sono dei cali temporanei (es. fine 2018, fine 2020, meta’ 2022), la correlazione rimane quasi sempre fortemente positiva e raramente scende sotto 0.6.
* Questa alta correlazione riflette il fatto che, nonostante il percorso di recupero di WFC, entrambe sono soggette in modo predominante agli stessi fattori che influenzano il settore bancario: Tassi d’interesse, regolamentazione, politiche della Federal Reserve, salute dell’economia e fiducia dei consumatori nel sistema finanziario.

Come cambiano le correlazioni in funzione dei rendimenti?

* I grafici illustrano chiaramente che la correlazione, sia tra azioni dello stesso settore che tra azioni e l’indice di mercato, non è statica ma è un fenomeno dinamico e variabile nel tempo.
* Possiamo vedere che durante periodi di incertezza o stress di mercato (es. all’inizio del 2020 per via della pandemia) tutte le correlazioni schizzano verso l’alto, indicando che in situazioni di panico, quasi tutti gli asset di muovono nella stessa direzione, riducendo i benefici della diversificazione.
* I cali nella correlazione possono riflettere periodi in cui il settore bancario sono influenzate da fattori specifici che non sono completamente riflessi nel mercato generale. Ad esempio, una particolare fase del ciclo dei tassi di interesse o cambiamenti di normativa potrebbero avere un impatto specifico e variare la loro correlazione.
* Infine, la correlazione con il mercato per il settore bancario tende a rimanere su livelli piu’ elevati rispetto ad altri settori piu’ dirompenti, a causa della loro stretta dipendenza dall’economia reale e dalla politica monetaria.

Come cambiano le correlazioni in funzione dei rendimenti?

* Le tendenze osservate sono coerenti con il concetto di correlazione Asimmetrica: Si tende a osservare che le correlazioni aumentano in periodi di rendimenti negativi elevati (mercato in calo). Questo e’ il fenomeno del contagio in cui, durante le crisi, gli asset tendono a perdere valore all’uniscono, indipendentemente dai loro fondamentali individuali. Le punte verso l’alto nelle correlazioni mobili spesso coincidono con tali periodi.
* In periodi di rendimenti positivi o di mercato piu’ calmo, le correlazioni possono essere piu’ basse o piu’ variabile, poiché i fattori specifici dell’azienda/settore possono avere un peso maggiore sulla performance.

La dispersione dei punti nello scatterplot conferma o no la relazione lineare fra i due rendimenti?

* S&P500 e JPM
  + La nuvola di punti è stretta e chiaramente allineata lungo una diagonale ascendente. C’è una dispersione minima attorno a questa linea, indicando una forte relazione lineare positiva tra i rendimenti di JPM e quelli dell’SP500.
* S&P500 e WFC
  + La nuvola di punti anch’essa molto stretta e allineata lungo una diagonale ascendente. Questo conferma una forte relazione lineare positiva tra i rendimenti di WFC e quelli dell’SP500.
  + Questo conferma una forte relazione lineare positiva tra i rendimenti di WFC e quelli dell’SP500.
* JPM e WFC
  + La nuvola è estremamente compatta e densamente raggruppata attorno a una diagonale ascendente.
  + Si conferma una relazione lineare estremamente forte e positiva. La quasi totale assenza di dispersione significativa lontano dalla linea centrale evidenzia come questi due titoli si muovano quasi in perfetta sincronia su base giornaliera.

# MODELLO ARIMA

* METODOLOGIA DI MODELLAZIONE

1. Sono stati utilizzati i rendimenti logaritmici mensili per addestrare e testare il modello ARIMA, prevedendo, naturalmente, i rendimenti logaritmici mensili in quanto questi dati sono meno volatili nel tempo e sono piu’ stabili (piu’ stazionari).
2. È stata utilizzata la funzione auto\_arima (dalla libreria “pmdarima”) per determinare automaticamente l’ordine (p, d, q) ottimale del modello ARIMA.
3. Una volta addestrato il modello, sono stati previsti i rendimenti logaritmici futuri. Questi sono stati poi riconvertiti in prezzi di chiusura mensili utilizzando il logaritmo inverso (np.exp()) e l’ultimo prezzo di chiusura reale disponibile.
4. Le prestazioni del modello sono state valutate utilizzando le metriche dell’errore RMSE e MAE per quantificare la differenza tra i prezzi reali e quelli previsti nel periodo di test/previsioni.

I risultati sono stati:

* Immagine che contiene testo, linea, Diagramma, diagramma

  Il contenuto generato dall'IA potrebbe non essere corretto.AAPL
* Il p-value del test di Dickey-Fuller è estremamente piccolo e molto inferiore ai livelli di significatività comuni come 0.05. Possiamo rifiutare l’ipotesi nulla di non stazionarietà. Questo conferma che la serie dei rendimenti logaritmici mensili di AAPL è stazionaria, rendendola idonea per la modellazione ARIMA.
* La funzione auto\_arima ha identificato il miglior modello ARIMA per i rendimenti logaritmici mensili di AAPL come (0, 0, 0). Questa configurazione del modello significa che non sono state trovate correlazioni significative con i valori passati della serie

# STRATEGIA DI TRADING: CROSSOVER DELLE MEDIE MOBILI

Questo capitolo esplora l'applicazione dell'analisi tecnica per la costruzione di una strategia di trading automatizzata e la sua valutazione tramite backtesting, confrontando le performance con una strategia "Buy & Hold".

Per questa sezione, è stata sviluppata una strategia di trading algoritmica basata sulla **tecnica di crossover delle medie mobili**. Questa strategia è ampiamente utilizzata nell'analisi tecnica per generare segnali di acquisto e vendita, basandosi sul principio che il superamento di una media mobile di breve periodo al di sopra di una di lungo periodo (e viceversa) può indicare un cambiamento nella tendenza del prezzo.

La strategia si basa sul confronto tra due medie mobili (Moving Averages - MA) calcolate sui prezzi di chiusura del titolo AAPL:

* **Media Mobile di Breve Periodo (MA\_short):** Cattura le tendenze di prezzo più recenti.
* **Media Mobile di Lungo Periodo (MA\_long):** Rappresenta la tendenza di prezzo di fondo.

Vengono generati segnali di trading:

* **Segnale di Acquisto (Long):** Quando la MA\_short attraversa al di sopra della MA\_long, viene generato un segnale di acquisto (posizione long, Signal = 1). Ciò suggerisce l'inizio di una tendenza al rialzo.
* **Segnale di Vendita/Chiusura (Short):** Quando la MA\_short attraversa al di sotto della MA\_long, viene generato un segnale di vendita (posizione short, Signal = -1 per posizionarsi al ribasso, o chiudere la posizione long). Ciò indica una potenziale inversione di tendenza al ribasso.
* **Assenza di Posizione:** Se non c'è un segnale chiaro o se si sceglie di non posizionarsi, il segnale può essere 0 (posizione neutra). Nel nostro caso, la strategia prevede di essere sempre posizionati, o long (+1) o short (-1), basandosi sul crossover.

La strategia è stata sottoposta a backtesting sui dati storici di AAPL. Per identificare la combinazione ottimale delle medie mobili, è stato eseguito un processo di ottimizzazione testando diverse combinazioni dei periodi short e long delle medie mobili.

* Il periodo short è stato variato da 10 a 100 giorni (passo di 10).
* Il periodo long è stato variato da 50 a 300 giorni (passo di 10), assicurando sempre che short < long.

L'efficienza di ciascuna combinazione è stata misurata utilizzando l'**Indice di Sharpe** annualizzato, una metrica che valuta il rendimento aggiustato per il rischio della strategia (rendimento in eccesso per unità di deviazione standard del rendimento).

L’ottimizzazione ha rilevato che la combinazione di medie mobili con il miglior indice di Sharpe per AAPL e’:

* Short: 100 periodi
* Long: 300 periodi
* **Sharpe: 0.6696783865510522**

Immagine che contiene testo, Diagramma, linea, diagramma

Il contenuto generato dall'IA potrebbe non essere corretto.

Per valutare l'efficacia della strategia di trading basata sul crossover delle medie mobili, le sue performance sono state confrontate con una strategia passiva di "Buy & Hold" (acquisto e mantenimento del titolo per l'intero periodo). La metrica chiave per questo confronto è l'Indice di Sharpe.

L'Indice di Sharpe della strategia "Buy & Hold" per il titolo AAPL sull'intero periodo di analisi è risultato pari a:

* **Sharpe Buy&Hold: 0.943026362667236**

Il Buy&Hold è superiore in termini di rischio-rendimento: Lo Sharpe Ratio del Buy&Hold, superiore rispetto alla strategia di crossover, indica che il ritorno medio annuo è migliore.

La strategia crossover con Sharpe Ratio di 0.67 non è affatto scarso. Significa che la strategia ha un profilo rischio/rendimento accettabile e gestisce bene l’esposizione al rischio.

Inoltre, questa strategia ha dei vantaggi significativi:

* La strategia Crossover ha il vantaggio di ridurre le perdite nei drawdown, cosa che il Buy&Hold non fa.
* Il crossover può ridurre la volatilità e il rischio di forti ribassi.

In sintesi, la strategia crossover a 100/300 giorni è molto piu’ protettiva rispetto al Buy&Hold durante i mercati ribassisti, il che fornisce protezione nei periodi negativi.

# CAPM

Per stimare la sensibilità di ciascun titolo rispetto al mercato, è stato calcolato il **beta** tramite una regressione lineare dei rendimenti mensili in eccesso di ciascun titolo rispetto ai rendimenti in eccesso dell’indice S&P 500 (^GSPC), considerato come proxy del mercato.

* Tasso risk-free: Ho utilizzato il tasso del T-Bill USA a 1 anno parti al 4.11% annuo, equivalente a 0.3425% mensile.
* Rendimento atteso di mercato: Ho assunto un rendimento medio storico dell’8% annuo dell’S&P500, equivalente al 0.06667% mensile.
* Rendimento in eccesso del mercato: Calcolato come la differenza tra il rendimento semplice mensile dell’S&P500 e il tasso risk-free.
* Rendimento in eccesso dei titoli: Calcolato sottraendo il risk-free ai rendimenti mensili di ciascun titolo.

È stato quindi applicato un modello di regressione OLS per ogni titolo, con formula:

* (Ri – Rf) = alpha + Beta(Rm - Rf) + eps

I valori di alpha e beta ottenuti sono riportati nella seguente tabella:

Immagine che contiene testo, schermata, Carattere, numero

Il contenuto generato dall'IA potrebbe non essere corretto.

Questi valori evidenziano una forte esposizione al rischio di mercato per NVDA e TSLA (Sono piu’ volatili del mercato), mentre JPM e WFC mostrano beta piu’ bassi anch’essi positivi, rendendoli piu’ volatili rispetto il mercato.

Utilizzando i beta stimati, è stato calcolato il rendimento atteso annuo secondo la formula del CAPM.

I risultati sono i seguenti:

* **AAPL**: 8.93%
* **NVDA**: 11.14%
* **TSLA**: 11.22%
* **GM**: 9.65%
* **JPM**: 8.54%
* **WFC**: 8.60%

Questi rendimenti riflettono l’intensità dell’esposizione al mercato: Titoli piu’ volatili e piu’ sensibili al rischio sistematico, come TSLA e NVDA, mostrano un rendimento atteso piu’ elevato.

ESPOSIZIONE AI FATTORI DI RISCHIO FAMA-FRENCH: MODELLO A 3 FATTORI

È stato stimato un modello multifattoriale per ogni titolo:

(Ri – Rf) = Alpha + β\_mkt(Rm - Rf) + β\_smb \* SMB + β\_hml \* HML + Eps

I fattori utilizzati (SMB, HML, RF) sono stati scaricati dal database Fama-French a 3 fattori con frequenza mensile:

Immagine che contiene testo, schermata, Carattere, numero

Il contenuto generato dall'IA potrebbe non essere corretto.

Immagine che contiene testo, schermata, Carattere, numero

Il contenuto generato dall'IA potrebbe non essere corretto.

🡪

È stata rinominata la colonna “date\_ff\_facators” a semplicemente “Data” e poi è stata impostata come indice del DataFrame.

Immagine che contiene testo, schermata, Carattere, numero

Il contenuto generato dall'IA potrebbe non essere corretto.

* AAPL
  + β\_mkt = 1.24 🡪 Mostra come AAPL è molto sensibile al mercato: Quando il mercato sale o scende, AAPL tende a muoversi ancora di piu’ nella stessa direzione.
  + β\_smb = -0.214 🡪 Ha un’esposizione leggermente contraria rispetto al fattore SMB (Small Minus Big), il che conferma che AAPL non segue un comportamento simile alle piccole imprese bensì similmente alle grandi imprese (Big Caps): Basso rischio legato alla taglia. Di conseguenza, AAPL ha basso rischio, piu’ stabilità, meno volatilità e piu’ liquidità.
  + β\_hml = -0.567 🡪 Si comporta come un growth stock: AAPL reagisce positivamente quando le growth vanno bene. basso valore contabile rispetto al prezzo, ma con alte aspettative di crescita future. Si conferma che AAPL è una società in rapida espansione rispetto l’intero mercato finanziario.
  + α = 0.0077 🡪 Buona capacità di sovraperformare il modello. Supera il rendimento previsto da Fama-French dal suo rischio.
  + R² = 0.535 🡪 Il modello spiega solo il 53.5% delle variazioni dei rendimenti: AAPL presenta anche comportamenti molto imprevedibili o influenzati da altri fattori.
* NVDA
  + β\_mkt = 1.79 🡪 Fortemente influenzata dai movimenti di mercato: Alta volatilità.
  + β\_smb = -0.087 🡪 Ha un’esposizione leggermente orientata verso grandi aziende, ma quasi neutra.
  + β\_hml = -0.929 🡪 Molto simile a una growth stock pura, tipica delle aziende tech innovative.
  + α = 0.0363 🡪 Ottima sovraperformance rispetto al modello.
  + R² = 0.418 🡪 Il modello spiega solo il 41.8% delle variazioni dei rendimenti: NVDA presenta piu’ comportamenti imprevedibili o influenzati da altri fattori che il modello non riesce a spiegare/conoscere.
* TSLA
  + β\_mkt = 1.75 🡪 Anche questa azienda ha un’altissima esposizione al mercato: Molto volatile.
  + β\_smb = 0.752 🡪 Si comporta come una small Cap: Profilo di rischio da impresa giovane o in crescita.
  + β\_hml = -1.089 🡪 Azione molto growth, lontana dal profilo “value”.
  + α = 0.0158 🡪 Sovraperformance moderata.
  + R² = 0.293 🡪 Il modello spiega solo il 29.3% delle variazioni dei rendimenti: Ci sono altri fattori che influenzano il prezzo che sono imprevedibili dal modello.
* GM
  + β\_mkt = 1.27 🡪 Buona sensibilità al mercato.
  + β\_smb = 0.788 🡪 Si comporta come una small Cap anche se è una grande impresa: Riflette un rischio operativo alto.
  + β\_hml = 0.608 🡪 Comportamento da value stock: Sottovalutata dal mercato.
  + α = -0.0025 🡪 Leggera **sottoperformance** rispetto al mercato.
  + R² = 0.577 🡪 Il modello spiega solo il 57.7% delle variazioni dei rendimenti.
* JPM
  + β\_mkt = 1.09 🡪 Segue moderatamente il mercato.
  + β\_smb = 0.0607 🡪 Quasi neutra: Né Small ne’ big cap.
  + β\_hml = 0.846 🡪 Forte esposizione a value stock: Tipico dele banche solide.
  + α = 0.0055 🡪 Leggera sovraperformance.
  + R² = 0.748 🡪 Il modello spiega molto bene il rendimento (75%).
* WFC
  + β\_mkt = 1.08 🡪 Simile a JPM: Segue il mercato in modo equilibrato.
  + β\_smb = 0.139 🡪 Leggera tendenza verso Small Cap, ma quasi neutra.
  + β\_hml = 0.980 🡪 Altamente value stock: Classica banca sottovalutata.
  + α = -0.0026 🡪 Leggera sottoperformance.
  + R² = 0.617 🡪 Il modello spiega circa il 62% delle variazioni.

ESPOSIZIONE AI FATTORI DI RISCHIO FAMA-FRENCH: MODELLO A 5 FATTORI

Questa sezione presenta i risultati dell’analisi dell’esposizione dei titoli selezionati ai fattori di rischio sistematici, utilizzando il modello a 5 fattori di Fama-French, con rendimenti e fattori su base mensile. Questo modello estende il modello a 3 fattori aggiungendo altri 2 fattori: RMW (redditività) e CMA (investimenti).

Sono stati scaricati dal database Fama-French a 5 fattori con frequenza mensile.

Immagine che contiene testo, schermata, Carattere, numero

Il contenuto generato dall'IA potrebbe non essere corretto.Di seguito, si presenta la tabella dei risultati ottenuti:

* 1. Alpha (Rendimento extra mensile)
* La maggior parte dei titoli mostra un alpha leggermente positivo (da 0.000276 a 0.036675), ad eccezione di GM che ha un alpha leggermente negativo e, quindi, una sottoperformance sottile rispetto al mercato.
* NVDA spicca con l’alpha piu’ elevato, con una sovraperformance di circa 3.67% di rendimento mensile.
  1. β\_mkt (Rischio di Mercato Mensile)
     + Tutti i titoli presentano un β\_mkt positivi e superiori a 1. Questo è coerente con la loro natura large cap dell’SP500 e conferma la loro sensibilità ai movimenti del mercato generale.
     + TSLA (1.777) e NVDA (1.671) hanno i beta di mercato piu’ alti, riflettendo la loro maggiore **volatilità** e **sensibilità** ai trend di mercato rispetto agli altri titoli. Dall’altra parte, i titoli bancari (JPM e WFC) hanno beta piu’ vicini a 1, indicando una sensibilità’ al mercato simile alla media.
  2. β\_smb
     + Principalmente negativo per i Large-Cap (AAPL, NVDA, JPM e WFC), come atteso. TSLA e GM mostrano, invece, una leggera esposizione positiva, indicando un comportamento simile alle piccole imprese (Small-Cap’s).
  3. β\_hml
     + I risultati sono coerenti con la natura delle azioni:
       - AAPL, NVDA e TSLA mostrano un comportamento di Growth Stock (valore negativo), performano meglio quando il mercato favorisce la crescita.
       - GM, JPM, WFC mostrano un comportamento di Value Stock, indicando la loro sensibilità ai fattori che favoriscono i titoli value.
  4. β\_rmw (Redditività)
     + Positivo per AAPL, NVDA, indicando un’alta redditività.
     + Negativo per TSLA, GM, JPM, WFC, indicando una bassa redditività: Spendono molto, margini bassi; Oppure, la loro redditività potrebbero non essere sempre allineati con questo fattore.
  5. β\_cma (Investimenti)
     + Negativo per NVDA, TSLA, JPM e WFC: Coerente con aziende che investono **aggressivamente** per la crescita o che hanno subito fasi di investimento intenso.
     + Positivo per AAPL, GM, indicando politiche di investimento piu’ **conservative**.
  6. R-squared (Potere esplicativo del modello)
     + Il modello a 5 fattori riesce a spiegare una porzione sostanziale della varianza nei rendimenti mensili.
     + JPM (78.8%) e WFC (66.6%): Il modello è molto efficace nello spiegare i rendimenti delle banche, indicando che la loro performance è **fortemente legata a questi fattori sistematici.**
     + AAPL (55.3%) e GM (57.3%): Il modello spiega piu’ della metà della varianza, fornendo una buona comprensione delle loro fonti di rischio.
  + NVDA (44.2%) e TSLA (29.4%): Questi titoli presentano comportamenti molto imprevedibili o influenzati da altri fattori, sono altamente dinamici.

# SIMULAZIONE STRATEGIA DINAMICHE

Questo capitolo introduce la simulazione di una strategia di gestione dinamica del portafoglio basata sul modello CPPI (Constant Proportion Portfolio Insurance), confrontando le performance con una strategia passiva di Buy&Hold su un portafoglio di azioni selezionate.

L’obiettivo della CPPI è garantire un valore minimo del portafoglio (floor) pur partecipando ai potenziali rialzi del mercato.

IMPLEMENTAZIONE DELLA STRATEGIA CPPI

La strategia CPPI è stata implementata sul portafoglio dei sei titoli precedentemente analizzati (AAPL, NVDA, TSLA, GM, JPM, WFC), assegnando un peso uguale a ciascun titolo (Equal Weights) per la componente rischiosa del portafoglio. La logica della CPPI prevede l’allocazione dinamica del capitale tra un asset rischioso e un asset sicuro.

PARAMETRI INIZIALI DELLA SIMULAZIONE

* Equal Weights = np.repeat(1/6, 6)

Pesi fissi per il portafoglio rischioso

* Investimento Iniziale: 100.000 euro
* Floor iniziale: 80% dell’investimento iniziale (0.8 x 100.000 = 80.000 euro)

Il floor diventa **dinamico** successivamente, proteggendo l’80% del valore massimo raggiunto.

* Rendimento Asset Sicuro (Safe\_return): 0.02/252

Rendimento annuale del 2%, distribuito giornalmente.

* Moltiplicatore: 4
* Peso minimo Asset Rischioso: 0.2 (20%)

PRINCIPIO DI FUNZIONAMENTO

La CPPI calcola una “cushion” (cuscinetto) come la differenza tra il valore corrente del portafoglio e un valore “floor” (soglia minima). L’esposizione dell’asset rischioso e’ determinata moltiplicando questo cuscinetto per un “moltiplicatore”

* Cushion 🡪 Valore del portafoglio – Floor
* Esposizione all’Asset Rischioso 🡪 Moltiplicatore x Cushion
* Peso all’Asset Rischioso 🡪 min(Esposizione all’Asset Rischioso/Valore portafoglio, 1)

Assicurandosi che non superi il 100%

* Peso all’Asset Sicuro 🡪 1 – Peso all’Asset rischioso

Immagine che contiene testo, Diagramma, linea, diagramma

Il contenuto generato dall'IA potrebbe non essere corretto.La performance della strategia CPPI ottimizzata è stata confrontata con quella di un portafoglio Buy&Hold equivalente.

* Dal grafico, è evidente che la strategia Buy&Hold (Linea arancione) ha generato un valore finale del portafoglio cumulato significativamente piu’ elevato rispetto alla strategia CPPI (Linea blu). Il Buy&Hold raggiunge quasi 10 volte il valore iniziale (indicando una performance cumulativa di quasi 900%), mentre la CPPI si ferma a circa 4 volte il valore iniziale (Performance cumulativa del 397%).

Questo dimostra che la CPPI, pur proteggendo, sacrifica una parte dell’upside in un mercato caratterizzato da una forte tendenza al rialzo.

* La linea verde rappresenta l’allocazione dinamica agli asset rischiosi. Si osserva che durante i periodi di forte calo del mercato (es. inizio 2020, inizio 2022-fine 2023), l’allocazione agli asset rischiosi **diminuisce drasticamente**, talvolta toccando il peso minimo del 20%. Questo è il meccanismo di protezione della CPPI che **riduce l’esposizione al rischio**.

METRICHE DI PERFORMANCE

Immagine che contiene testo, Carattere, schermata, bianco

Il contenuto generato dall'IA potrebbe non essere corretto.Immagine che contiene testo, Carattere, schermata, bianco

Il contenuto generato dall'IA potrebbe non essere corretto.

Nel confronto tra le due strategie, emergono differenze significative in termini di rendimento, rischio e stabilità.

* Rendimento

La strategia Buy&Hold ha ottenuto una performance cumulativa molto superiore (+1.049%) rispetto alla strategia CPPI (+297%). Questo e’ coerente con la natura della Buy&Hold, che rimane completamente esposta agli asset rischiosi, beneficiando pienamente delle fasi di rialzo dei mercati.

* Volatilità Annualizzata (Rischio)

La strategia CPPI mostra una volatilità annualizzata dimezzata rispetto Buy&Hold (13.32% contro 26.58%). Questo conferma che la gestione **dinamica dell’allocazione CPPI** riduce il rischio complessivo del portafoglio.

* Drawdown Massimo (Protezione dal Rischio)

La strategia CPPI ha ottenuto un drawdown massimo di -19.32%, il che lo rende notevolmente inferiore rispetto a quello del Buy&Hold (-42.44%). Questo è il principale beneficio della CPPI, che attraverso il suo meccanismo di allocazione agli asset rischiosi dinamico, riesce a mitigare le perdite durante i cali significativi del mercato. Il grafico lo mostra con la linea blu della CPPI che è piu’ **liscia** e meno soggetta a picchi negativi profondi rispetto alla linea arancione.

Inoltre, dal grafico si può vedere il comportamento della linea verde dell’allocazione agli asset rischiosi evidenzia come la CPPI riduca l’esposizione durante le turbolenze di mercato per proteggere il capitale dal raggiungimento del floor.

* Efficienza del portafoglio (Sharpe Ratio)

Lo Sharpe Ratio della strategia CPPI è leggermente superiore rispetto al Buy&Hold (1.12 contro 1.01), indicando che ogni unità di rischio assunto ha generato un rendimento leggermente migliore: CPPI è piu’ efficiente, anche se meno redditizio nel lungo periodo.

La strategia CPPI ha gestito il rischio in modo piu’ efficiente, fornendo un **ritorno migliore** per il rischio preso.

CONCLUSIONI

Il backtesting dimostra che la strategia CPPI ha raggiunto con successo il suo obiettivo primario: fornire protezione del capitale e migliorare i rendimenti aggiustati per il rischio, come evidenziato dal suo Sharpe Ratio superiore e dal drawdown massimo significativamente inferiore. Questo è stato ottenuto a costo di sacrificare una porzione del potenziale di crescita assoluta in un mercato che ha visto forti trend rialzisti.

La strategia CPPI con allocazione Equals Weights si dimostra una soluzione prudente e bilanciata: Sacrifica parte del potenziale di rendimento per ottenere maggiore stabilità e protezione contro le perdite. Questa strategia sarebbe adatta per investitori con una bassa tolleranza al rischio.

D’altro canto, la strategia Buy&Hold, pur essendo piu’ esposta alla volatilità e al rischio di drawdown, massimizza la crescita di lungo periodo, rendendola ideale per chi può sopportare oscillazioni significative del portafoglio.

Immagine che contiene testo, schermata, schermo, diagramma

Descrizione generata automaticamente

Con il metodo “.plot()” possiamo plottare la frequenza dei ratings.

Dal grafico possiamo notare:

* Il rating più frequente è il 5
* Il rating meno frequente è il 2

PRINCIPALI STATISTICHE DESCRITTIVE: RATING

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

Utilizzando il metodo *describe()* abbiamo analizzato le principali statistiche descrittive dei rating e abbiamo ottenuto che:

* Il numero totale di valutazioni è pari a 105,583.
* La valutazione media è 4.23 indicando la tendenza generale verso recensioni positive.
* La deviazione standard è pari a 1.13, il che suggerisce che le valutazioni sono distribuite vicino alla media. La maggior parte delle recensioni sono comprese nell’intervallo [3.10, 5].

TOP 10 PRODOTTI PIU’ RECENSITI

Immagine che contiene testo, schermata, Carattere

Descrizione generata automaticamente

I prodotti più recensiti sono stati:

* **B00008OWZG** con 220 recensioni
* **B000002H97** con 205 recensioni
* **B000002H33** con 198 recensioni
* **B000002J09** con 194 recensioni
* **B000002UB3** con 190 recensioni
* **B00004ZAV3** con 184 recensioni
* **B000089RVX** con 184 recensioni
* **B000002U82** con 179 recensioni
* **B0002OERI0** con 178 recensioni
* **B000002UAU** con 175 recensioni

RATING MEDIO ASSEGNATO DAGLI UTENTI:BOXPLOT

Immagine che contiene testo, schermata, linea, diagramma

Descrizione generata automaticamente

Il boxplot rappresenta i rating medi assegnati dagli utenti, calcolati raggruppando i dati per “user\_id” e prendendo la media dei rating di ciascun utente.

Possiamo vedere che l’intervallo interquartile si trova tra circa il voto 4 e 5 il che significa che il 50% dei dati si trovano proprio qui. Questo indica una distribuzione concentrata verso l’alto.

Gli outlier, ovvero, i punti fuori dai baffi rappresentano gli utenti con un comportamento anomalo rispetto alla maggior parte e ci indicano che alcuni utenti hanno assegnato rating molto bassi compressi tra l’1 e il 3.

# Risultati del progetto

In questa sezione vengono presentati i risultati ottenuti attraverso le diverse fasi di sviluppo del sistema di raccomandazioni basato su collaborative filtering. Ogni passaggio è stato progettato per ottimizzare l'accuratezza delle previsioni, garantire una segmentazione efficace degli utenti e generare raccomandazioni personalizzate.

Il lavoro è stato suddiviso in diverse fasi principali, tra cui l’identificazione della configurazione ottimale per l’algoritmo K-NN, il riempimento della matrice di rating tramite tecniche di predizione, la segmentazione degli utenti in base alle preferenze e l’applicazione della Matrix Factorization come confronto.

I risultati sono analizzati con metriche come MSE e RMSE per valutare l’efficacia degli algoritmi utilizzati, mentre l’approccio di clustering degli utenti ha permesso di evidenziare pattern di preferenze utili per migliorare ulteriormente le raccomandazioni.

## IDENTIFICAZIONE DELLA CONFIGURAZIONE OTTIMALE DELL’ALGORITMO K-NN

### IMPORTAZIONE DEI MODULI NECESSARI

Prima di tutto dobbiamo importare dalla libreria “Surprise” i seguenti moduli:

* Immagine che contiene testo, Carattere, schermata, linea

  Descrizione generata automaticamenteDataset 🡪 Questa classe ci permette di trasformare i nostri dati in un formato compatibile con Surprise in modo di poter gestirli e manipolarli per l’addestramento dell’algoritmo K-NN.
* Reader 🡪 Ci permette di definire il range di valori dei rating.
* Accuracy 🡪 Contiene i metodi MSE e RMSE che ci permetteranno di valutare le prestazioni degli algoritmi di raccomandazione.
* KNNBasic 🡪 Algoritmo che utilizzeremo per calcolare le somiglianze tra utenti/prodotti e di effettuare raccomandazioni in base ai vicini più simili.
* Model\_selection 🡪 Ci permette di dividere i dati in trainset e testset e poi eseguire il cross-validation.

### PREPARAZIONE DEI DATI E SUDDIVISIONE DEL DATASET

Immagine che contiene testo, Carattere, linea, schermata

Descrizione generata automaticamente

Per preparare il dataset per l'addestramento del modello, abbiamo utilizzato la libreria Surprise. In primo luogo, abbiamo definito un oggetto Reader per specificare il range dei valori delle valutazioni, che nel nostro caso è compreso tra 1 e 5. Successivamente, il dataset è stato caricato utilizzando il metodo “load\_from\_df” per trasformarlo in un formato compatibile con la libreria.

Inoltre, abbiamo inizializzato l’algoritmo K-NN sulla variabile “algo”

Infine, il dataset trasformato è stato suddiviso in due sezioni: Il **trainset** e il **testset** utilizzando il metodo “train\_test\_split()” il cui riceve un parametro “test\_size” per specificare la porzione dell’intero dataset che vogliamo destinare al set di test, che in questo caso abbiamo scelto l’80% dei dati per l’addestramento e il 20% per il test.

### FASE DI ADDESTRAMENTO, PREDIZIONE E VALUTAZIONE DEL MODELLO

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

L’algoritmo KNN viene addestrato sul trainset con il metodo “fit()”. Questo passaggio permette all’algoritmo di apprendere le relazioni tra utenti e prodotti, successivamente calcola le similarità tra utenti/prodotti.

Dopo essere addestrato l’algoritmo viene utilizzato per effettuare predizioni sui dati destinati al test tramite il metodo “test()”. Tutte le predizioni le salviamo sulla variabile “predictions”.

Ora possiamo passare al calcolo degli errori MSE e RMSE sulle predizioni tramite il modulo “accuracy” che invoca i metodi “mse()” e “rmse()”.

In questo caso abbiamo che:

* **MSE: 0.9591**
* **RMSE: 0.9794**

Possiamo vedere che entrambi hanno un valore minore a 1, il che ci fa capire che l’errore non è così grande e quindi, **inizialmente, il modello è abbastanza accurato** anche se non abbiamo ancora trovato la configurazione ottimale.

### OTTIMIZZAZIONE DEL PARAMETRO K CON CROSS-VALIDATION

Attraverso il cross-validation possiamo provare diversi valori di K per l’algoritmo in modo tale di trovare il valore ottimo nel quale gli errori MSE e RMSE siano i minori in assoluto e, dunque, il modello sia il più accurato possibile.

Il processo del cross-validation suddivide i dati in 5 folds in cui l’algoritmo viene addestrato su 4 di essi e testato sull’ultimo, ripetendo questo processo per ogni fold.

Immagine che contiene testo, schermata, Carattere, software

Descrizione generata automaticamente

Per prima cosa creiamo degli array in cui appenderemo gli eventuali valori di mse e rmse al variare di K, questi si chiameranno “mse\_over\_k” e “rmse\_over\_k”. È molto importante tener traccia dei risultati ad ogni iterazione per, poi, poter rappresentarli graficamente tramite l’elbow method.

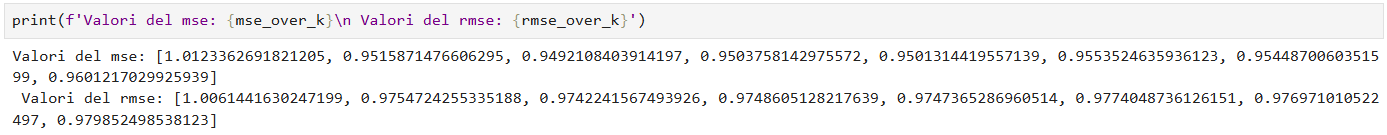
Abbiamo scelto i valori di K dal 5 al 41 con un incremento di 5 unità con il metodo “.arange()” in questa maniera l’algoritmo non diventa né troppo specifico, impostando K piccoli, né troppo generalista, impostando K molto grandi.

Un valore piccolo di K avrebbe reso il modello più focalizzato sui vicini stretti e quindi potrebbe verificarsi un overfitting, ovvero, che il modello diventerebbe troppo specifico rispetto ai dati di addestramento il che non permetterebbe al modello di apprendere pattern generalizzabili.

Dall’altro canto con una scelta grande di K renderebbe il modello più generalista il che potrebbe non catturare i pattern nei dati di addestramento, cioè, non riuscirebbe a rappresentare adeguatamente la complessità del problema.

Visto che il cross-validation deve addestrare e testare l’algoritmo su ogni fold ci saranno cinque valori di mse e rmse alla fine dell’esecuzione, procediamo a prendere il valore medio di questi valori e appenderli sugli array appositi per salvarli tramite il metodo “.append()”.

### RISULTATI DEI VALORI DI MSE E RMSE: CROSS-VALIDATION

****

Come possiamo osservare: I valori del MSE e del RMSE arrivano al minimo globale nella **seconda posizione**, ciò vuol dire che con **K = 15** si otterrebbe la migliore configurazione di K.

Procediamo a plottare i risultati tramite l’elbow method.

### VALORE DI K OTTIMALE: ELBOW METHOD

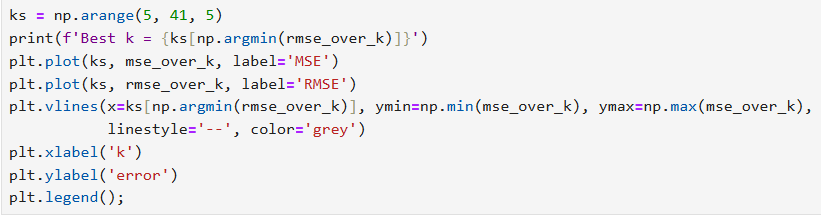


Immagine che contiene testo, diagramma, Diagramma, linea

Descrizione generata automaticamente

L’obiettivo di questo grafico è poter visualizzare l’andamento degli errori MSE e RMSE al variare del valore di K nel modello KNN.

Si è utilizzato l’Elbow method per individuare il valore di K basandosi sui valori degli errori calcolati durante la fase del cross-validation.

* L’asse x rappresenta i valori di K testati.
* L’asse y rappresenta i valori di errore MSE e RMSE.

Tra i valori di MSE e RMSE prendiamo in considerazione i valori dell’RMSE in quanto questi vengono rappresentati nella scala all’unità originale che in questo caso è [1, 5].

Il metodo *plot()*, per plottare le curve dei valori di MSE e RMSE, prende come parametri i valori dell’asse x (ks), i valori dell’asse y (mse\_over\_k/rmse\_over\_k) e un’etichetta (label = ‘MSE/RMSE’).

Dopodiché, tracciamo la linea verticale nel punto in cui si verifica il valore minimo assoluto dell’RMSE tramite il metodo *vlines()* che prende come parametri:

* Il punto dove l’RMSE è il minimo assoluto, ovvero, *np.argmin(rmse\_over\_k)*.
* I limiti **ymin** e **ymax** in cui sarà contenuta la linea verticale, lo stile e il colore della linea.

Infine, assegniamo i nomi degli assi x e y tramite i metodi *xlabel()* e *ylabel()*.

Osserviamo dal grafico che all’aumentare di K, l’errore inizialmente decresce fino a raggiungere un minimo per **K = 15**, dopodiché, inizia a crescere.

Nonostante ciò, questo è un **metodo euristico e non accurato**, serve solo per avere una prima idea di quanto potrebbe essere il valore di K, per questo motivo, esiste un altro metodo per trovare la configurazione ottimale dell’algoritmo KNN: **GridSearch**.

### CONFIGURAZIONE OTTIMALE: METODO GRID SEARCH

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamenteQuesto metodo ci permette di ricavare la configurazione ottimale per l’algoritmo KNN tramite una ricerca esaustiva della configurazione ottimale utilizzando diverse misure di similarità: MSD, Cosine, Pearson e diversi approcci: User-based e Item-based.

Eseguiamo i seguenti passaggi:

1. **Definiamo un dizionario** “*param\_grid*” in cui specifichiamo i valori da testare per ogni parametro.
2. K 🡪 Scegliamo i valori di K da **[15, 41].**
3. Sim\_options 🡪 Prendiamo in considerazione tutte le misure di similarità: **MSD, Cosine e Pearson** ed entrambi gli approcci: **User-Based e Item-Based**.
4. Invochiamo e inizializziamo il metodo GridSearch che prenderà come parametri:
5. KNNBasic 🡪 L’algoritmo da ottimizzare.
6. Param\_grid 🡪 La griglia di parametri da testare definita precedentemente.
7. Measures = [“rmse”, “mse”] 🡪 Le metriche di valutazione usate per il confronto dei risultati.
8. Cv = 5 🡪 Numero di folds per la cross-validation.
9. N\_jobs = -1 🡪 Numeri di processi paralleli per eseguire i calcoli. Quando impostato a -1 sfrutta tutti i core disponibili del processore.

Osserviamo che le metriche più basse ottenute durante la ricerca sono:

* **MSE = 0.8623**
* **RMSE = 0.9286**

E questi valori sono stati trovati con seguente **configurazione ottimale**:

* Best configuration = {'k': 24, 'sim\_options': {'name': 'msd', 'user\_based': False}}

In altre parole, con **K = 24, similarità MSD e con l’approccio Item-based** riusciamo a ottenere un modello KNNBasic il più accurato possibile.

## FILLING DELLA MATRICE DI RATING CON LA CONFIGURAZIONE OTTIMALE

### ALLENAMENTO DELL’ALGORITMO SULL’INTERO DATASET CON LA CONFIGURAZIONE OTTIMALE

**Immagine che contiene testo, Carattere, schermata, linea

Descrizione generata automaticamente**

Dopo aver trovato la configurazione ottimale l’algoritmo viene allenato utilizzando tutti i dati disponibili, ovvero, l’intero dataset in quanto non c’è più bisogno di riservare una parte dei dati al testset. Questo serve per sfruttare al massimo le informazioni presenti nel dataset e garantire che il modello sia il più accurato possibile prima delle predizioni.

Visto che l’obiettivo è utilizzare l’intero dataset per allenare il modello utilizzeremmo il metodo *build\_full\_trainset()* che ci permette di non splittarlo ma di unificarlo per il training.

### CREAZIONE DEL RATING\_MATRIX\_ORIGINALE

Immagine che contiene testo, Carattere, schermata, linea

Descrizione generata automaticamente

La creazione della matrice dei rating originale è un passaggio fondamentale in quanto ci permette di individuare i prodotti non ancora recensiti e sui quali fare predizioni, quindi, bisogna tenere traccia delle interazioni tra utente e prodotto.

L’obiettivo, quindi, è di creare una matrice in cui ogni riga rappresenti un utente e ogni colonna rappresenti un prodotto, per questo motivo abbiamo prima reperito tutti gli utenti e prodotti tramite il metodo *unique()*, dopodiché, tramite il metodo *DataFrame()* di pandas abbiamo creato la matrice:

* Impostando inizialmente tutte le celle a 0.
* Assegnando le righe tutti gli users\_id.
* Assegnando alle colonne gli items\_id.

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamentePer il popolamento della matrice con i rispettivi rating utilizziamo un ciclo in cui ogni riga rappresenta un’interazione user\_id 🡪 parent\_asin con un determinato rating, quindi, per ogni interazione viene aggiornato con il rating effettivo.

La matrice dei rating originale è stata creata e presenta: **2685 righe** (Numero di utenti univoci) e **3444 colonne** (Numero di prodotti univoci).

Tramite questa matrice, quindi, possiamo fare distinzione tra i prodotti recensiti o meno.

### PREDIZIONE DEL RATING DI TUTTI I PRODOTTI PER UN USER SPECIFICO

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

Prima di applicare il modello sull’intero dataset è importante testare le funzionalità del processo su un sottoinsieme dei dati su piccola scala.

Questo approccio porta molti vantaggi come:

* Riduzione dei tempi di esecuzione
* Costo computazionale minore
* Facilità di debug

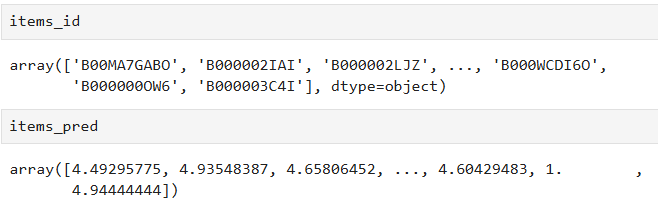
Abbiamo considerato l’utente “**AGTMZCWIWBH43TCW7UKG2YV2EKKA**” per il quale vogliamo calcolare le predizioni e la lista di tutti i prodotti univoci.

L’array “items\_pred” sarà una lista che conterrà i valori predetti per ogni prodotto.

Attraverso un ciclo controlliamo se ogni prodotto è stato recensito dall’utente in questione, si possono verificare due possibilità:

1. Se l’utente ha già assegnato un rating al prodotto allora aggiungiamo il valore di 0 alla lista delle predizioni “items\_pred” per indicare che non è necessaria una predizione.
2. Se l’utente non ha ancora assegnato un rating al prodotto allora viene utilizzato il metodo *algo.predict()* sull’utente e sul prodotto in questione e viene salvata in una variabile “res”. Possiamo reperire la predizione tramite il metodo *res.est* e l’aggiungiamo alla lista delle predizioni “items\_pred”.

Infine, la lista delle predizioni “items\_pred” viene convertita in un array Numpy per semplificare successive analisi.



A primo impatto possiamo osservare alcune delle predizioni che sono state salvate nella lista “items\_pred” per ogni prodotto.

Come ad esempio:

* Per il prodotto B00MA7GABO l’algoritmo ha calcolato un rating di 4.49295775.
* Per il prodotto B000002IAI l’algoritmo ha calcolato un rating di 4.93548387.
* Per il prodotto B000002LJZ l’algoritmo ha calcolato un rating di 4.65806452.

E così via per tutti i prodotti.

### TOP 10 PRODOTTI CONSIGLIATI PER L’UTENTE “AGTMZCWIWBH43TCW7UKG2YV2EKKA”

Dopo aver calcolato i punteggi predetti per i prodotti non recensiti da un utente specifico, abbiamo selezionato i 10 prodotti con rating predetti più alti.

Per prima cosa ordiniamo gli indici dell’array dei rating predetti in modo **crescente** tramite il metodo *np.argsort()*. Dopodiché invertiamo l’ordine in modo **decrescente** tramite l’operazione [::-1].

Usando gli indici ordinati in modo decrescente riorganizziamo sia l’array dei rating predetti che gli ID dei prodotti in modo che siano ordinati dal rating più alto al più basso.

Estraiamo una lista degli ID dei primi 10 prodotti, così come il loro punteggio.

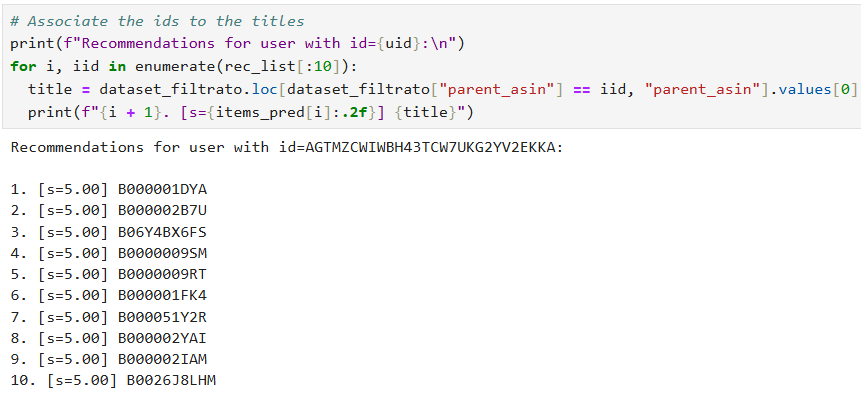
Ci rendiamo conto che i primi dieci prodotti con rating predetto più alti sono stati:

1. B000001DYA

Tutti i prodotti hanno avuto un rating predetto uguale a 5.0.

Questo accade dovuto al fatto che i dati hanno una tendenza generale verso punteggi positivi, infatti, osserviamo che quasi il 60% degli utenti hanno recensito con rating = 5. Quindi naturalmente i vicini degli utenti tenderanno a dare voti positivi.

1. B000002B7U
2. B06Y4BX6FS
3. B0000009SM
4. B0000009RT
5. B000001FK4
6. B000051Y2R
7. B000002YAI
8. B000002IAM
9. B0026J8LHM

Questo è un altro modo da rappresentare i top 10 prodotti con i loro rispettivi rating.

### FILLING DELLA MATRICE DEI RATING CON LA CONFIGURAZIONE OTTIMALE

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamenteDopo aver creato un sistema di raccomandazione per un solo utente procediamo a portarlo a grande scala creando la matrice dei rating predetti per ogni utente per ogni prodotto tenendo conto dei prodotti che sono stati e non sono stati recensiti per creare delle predizioni dove c’è bisogno.

Come primo passaggio estraiamo tutti gli utenti e i prodotti e li salviamo nelle variabili “user\_id” e “items\_id” rispettivamente.

Inizialmente, creiamo una lista vuota “filled\_rating\_matrix” che durante il ciclo, ad ogni iterazione, viene aggiunta una lista per ogni *uid*.

Dentro il ciclo annidato viene controllato se ogni prodotto è stato recensito o meno controllando la matrice dei rating originale in cui si verificano due casi:

1. Il rating == 0 🡪 Il prodotto NON è stato ancora recensito.
2. Il rating > 0 🡪 Il prodotto è già stato recensito.

Quindi:

* **Appendiamo uno zero** nella matrice dei ratings SE il prodotto è già stato recensito.
* **Appendiamo la predizione del rating** SE il prodotto non ha ancora un rating nella matrice dei rating originale. Come?

Invochiamo il metodo *predict(uid, iid)* e salviamo la risposta nella variabile “res”, dopodiché, appendiamo il *res.est* che rappresenta il rating predetto.

Appendiamo i valori nella posizione [-1] della matrice dei rating predetti in quanto con l’indice -1 accediamo sempre all’ultima lista aggiunta, cioè, l’ultima riga della matrice in costruzione che vogliamo popolare.

*(Se matrice\_rating\_filtrato.at[uid, iid] > 0 ALLORA filled\_rating\_matrix[-1].append(****0****))*

*(Se matrice\_rating\_filtrato.at[uid, iid] == 0 ALLORA filled\_rating\_matrix[-1].append(****res.est****))*

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

Alla fine del processo la matrice viene convertita in un array Numpy per una facile manipolazione successiva.

Come passaggio finale creiamo il DataFrame corrispondente alla matrice dei rating per ogni utente e prodotto in cui abbiamo fillato i rating mancanti con i rating predetti.

## SEGMENTAZIONE DEGLI UTENTI IN BASE ALLE PREFERENZE: ALGORITMO DI CLUSTERING K-MEANS CON COSINE SIMILARITY

Come prima considerazione, dobbiamo tener presente che l’algoritmo di clustering K-MEANS utilizza la distanza Euclidea (che è diversa alla similarità del coseno) per determinare la somiglianza degli utenti attraverso la distanza tra i punti e i centroidi dei clusters.

Immagine che contiene Carattere, testo, bianco, diagramma

Descrizione generata automaticamente

* Distanza Euclidea

Immagine che contiene testo, Carattere, bianco, schermata

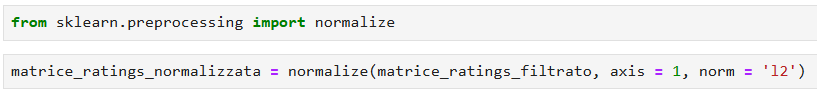
Descrizione generata automaticamente

* Cosine Similarity

Per utilizzare la similarità del coseno, al posto della distanza Euclidea, bisogna prima normalizzare vettori così abbiamo che se |A| = 1 e |B| = 1 ALLORA:

* Cosine\_similarity (A, B) = A . B
* Distanza Euclidea (A, B) = =

Quindi abbiamo che la distanza Euclidea dipende dall’angolo tra i vettori e non più dalle loro lunghezze. Possiamo dire, dunque, che la distanza Euclidea diventa equivalente alla similarità del coseno.



Importiamo il modulo ‘normalize’ di Sklearn.preprocessing che avrà lo scopo di normalizzare ogni riga della matrice originale dei rating utilizzando la norma l2.

*Abbiamo scelto la matrice dei ratings originale e non la matrice dei ratings predetti in quanto potrebbe portare a distorsioni nel clustering.*

La normalizzazione con norma l2 trasforma ogni riga della matrice (corrispondente ad ogni utente) in un vettore di lunghezza unitaria per poter finalmente utilizzare la similarità del coseno.

Immagine che contiene testo, Carattere, linea, numero

Descrizione generata automaticamente

Dopo aver normalizzato tutta la matrice dei ratings l’obiettivo è trovare il numero ottimale k di clusters per l’algoritmo K-Means per segmentare gli utenti in gruppi con comportamenti simili.

Creiamo l’array “wcss” (Within-Cluster Sum of Squares) in cui appenderemo ad ogni iterazione il valore della somma delle distanze al quadrato tra ogni punto all’intero del cluster e il centroide del cluster stesso.

Attraverso un ciclo fitiamo un valore diverso di K clusters che va da 1 al 121 con step di 5 e appendiamo il valore del wcss tramite l’attributo “km.inertia\_”, cosi via per tutti i valori di K clusters applicati nel modello.

Immagine che contiene testo, linea, Diagramma, diagramma

Descrizione generata automaticamente

Infine, plottiamo tuti i valori della lista “wcss” al variare del numero K di clusters e applichiamo l’Elbow method per trovare il miglior K.

Come possiamo osservare, il gomito si evince con **K = 20** clusters in cui si verifica un cambiamento significativo nella pendenza della curva.

* Prima del gomito la WCSS diminuisce rapidamente.
* Dopo il gomito la WCSS diminuisce più graduale.

Ora, applichiamo l’algoritmo K-Means alla matrice dei ratings normalizzata per assegnare ciascun utente a uno dei K = 20 clusters.

Immagine che contiene testo, Carattere, schermata, linea

Descrizione generata automaticamente

Per primo passaggio dobbiamo impostare i parametri del modello K-Means i quali sono i seguenti:

* N\_clusters = 20
* Init = ‘k-means++’ 🡪 Inizializza i centroidi in modo intelligente
* N\_init = 10 🡪 Ripete l’algoritmo 10 volte con diverse inizializzazioni e sceglie quella con WCSS più basso.
* Max\_iter = 300 🡪 Assegniamo un numero massimo di iterazioni per una singola esecuzione dell’algoritmo.
* Random\_state = 0 🡪 L’algoritmo prende lo stesso seed ad ogni iterazione cosi’ garantiamo clusters costanti.

Dopo aver impostato l’algoritmo K-Means con i parametri ottimali utilizziamo il metodo “fit\_predict()” per addestrare il modello sulla matrice dei ratings normalizzata con l’obiettivo di calcolare i centroidi dei clusters per poi assegnare, a ciascun utente, il cluster più vicino a lui.

La variabile *y\_means* fornirà le etichette dei cluster, indicando per ciascun utente a quale gruppo appartiene. Avrà la forma 🡪 [0, 17, 5, …, 9] con lunghezza = numero di utenti.

Infine, avendo segmentato gli utenti con successo creiamo un pandas DataFrame per visualizzare la relazione tra utente e cluster in cui avremmo due colonne ‘utente\_id’ e ‘cluster’.

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

Immagine che contiene testo, schermata, numero

Descrizione generata automaticamenteTramite il metodo ‘.value\_counts()’ possiamo vedere la frequenza degli utenti assegnati a ciascun cluster dopo l’applicazione dell’algoritmo K-Means con k = 20 clusters.

* Il **cluster 9** contiene il numero maggiore di utenti con 502 utenti. Ciò suggerisce che una porzione significativa di utenti ha preferenze molto simili rappresentate dal centroide di questo cluster.
* Il cluster 12 contiene il numero minore di utenti con solo 46 utenti.
* Ci sono diversi cluster (12, 19, 17, 7, 2, 1, 15, …) che hanno pochi utenti e ci sono diversi cluster (9, 18, 0, 14, 8, 6, 2, 5, …) che hanno molti utenti.

Possiamo osservare una **distribuzione non uniforme tra i cluster**: I Cluster hanno dimensioni variabili. Questa variazione significa che ci sono gruppi di utenti che hanno preferenze più specifiche/uniche rappresentati in cluster piccoli, così come, ci sono gruppi di utenti che hanno preferenze più comuni rappresentati in cluster grandi.

## CREAZIONE DELLA LISTA PERSONALIZZATA DI RACCOMANDAZIONI PER CIASCUN UTENTE

**Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente**

Dopo avere la matrice dei rating predetti per ogni utente e prodotto possiamo creare delle liste personalizzate di raccomandazioni in base alle predizioni dei rating.

In primo luogo, introduciamo la funzione “sort\_columns(row)” che ha il compito di ordinare le colonne (prodotti) di ogni riga (utente) in base al valore del rating predetto.

All’intero della funzione troviamo il metodo *sorted(a, b, c)* che ordina in ordine decrescente i prodotti in base al punteggio predetto e prende come parametri:

* Row.items() 🡪 restituisce una coppia (prodotto, rating)
* Key = lambda x : x[1] 🡪 Definisce la chiave di ordinamento, in questo caso vogliamo ordinare in base al **rating** quindi dovremmo prendere il secondo campo della coppia (x[indice = 1])
* Reverse = True 🡪 Indica che vogliamo un ordinamento decrescente

Infine, la funzione ritorna i nomi dei prodotti già ordinati in ordine decrescente in base al rating predetto.

Vogliamo applicare la suddetta funzione su tutte le righe (utenti) della nostra matrice dei rating predetti *res\_df.* Utilizziamo il metodo. apply() 🡪 res\_df.apply(sort\_columns, axis = 1) e creiamo un altro pandas DataFrame *rec\_lists.*

Come possiamo vedere dall’output di *rec\_lists* per ogni utente abbiamo una lista di prodotti consigliati, ad esempio, i top dei primi cinque prodotti consigliati per i primi tre utenti sono:

* Per l’utente “**AE7BV6IMNPZ3F266H7PXMH3BZQNQ**” i prodotti consigliati sono:

1. B000G1ALRU
2. B00006EXLQ
3. B00005OWEJ
4. B000002NHN
5. B00006J3UT

* Per l’utente “**AGTMZCWIWBH43TCW7UKG2YV2EKKA**” i prodotti consigliati sono:

1. B000002GKT
2. B000002UVZ
3. B000002O4I
4. B00005LNBH
5. B00000ICO0

* Per l’utente “**AGWDYYVVWM3DC3CASUZKXK67G6IA**” i prodotti consigliati sono:

1. B000002J0N
2. B01BGDXOFC
3. B0034CNSFU
4. B000NY17J6
5. B015EYVWPI

## FILLING MATRICE DEI RATING: MATRIX FACTORIZATION

Per applicare la tecnica del Matrix Factorization per riempire la matrice dei rating bisogna utilizzare l’algoritmo Singular Value Decomposition (SVD).

Quindi, come primo passaggio, bisogna importare SVD from surprise.

Immagine che contiene testo, Carattere, software, Pagina Web

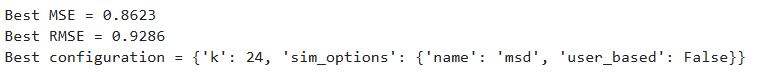
Descrizione generata automaticamente

Inizializziamo il modello SVD e dividiamo il dataset in un set di addestramento e un set di test utilizzando un rapporto di 80/20 come abbiamo fatto prima per il modello di K-NN.

Abbiamo ottenuto i seguenti valori:

* **MSE: 0.8168**
* **RMSE: 0.9038**

Bisogna tener conto che questi valori non sono ancora stati ottimizzati in quanto non abbiamo ancora trovato la configurazione ottimale dell’algoritmo SVD.

Possiamo, però, confrontare i valori di MSE e RMSE tra l’algoritmo KNN ottimizzato e l’algoritmo SVD non ancora ottimizzato:

* Il modello SVD ha un MSE leggermente più basso rispetto al KNN.
* Anche in termini di RMSE, il modello SVD mostra una performance superiore rispetto al KNN ottimizzato.

Questo suggerisce che una migliore capacità predittiva di SVD anche senza ottimizzazioni; quindi, proseguiamo e cerchiamo di trovare la configurazione ottimale dell’algoritmo SVD e fare il filling della matrice dei rating.

### CONFIGURAZIONE OTTIMALE DELL’ALGORITMO SVD: METODO GRID SEARCH

Per trovare la configurazione ottimale dell’algoritmo SVD non possiamo farlo tramite l’elbow method in quanto l’obiettivo non è trovare il numero K di vicini bensì altri parametri che non possono essere analizzati de rappresentati tramite una curva al variare degli errori di MSE e RMSE.

Il miglior metodo per l’algoritmo SVD è il Grid Search.

Immagine che contiene testo, schermata, Carattere

Descrizione generata automaticamente

I parametri ottimali dell’algoritmo SVD che vogliamo ottimizzare sono:

* N\_factors 🡪 Numero di fattori latenti per modellare le interazioni tra utente e prodotti: Varia tra 80 e 160 con step 10.
* N\_epochs 🡪 Numero di interazioni per l’apprendimento. Varia tra 10 e 50 con step di 10.
* Biased🡪 I bias rappresentano tendenze osservati nei dati.

Tramite il grid Search abbiamo trovato la configurazione ottimale esaustiva:

* **N\_factors = 90**
* **N\_epochs = 30**
* **Biased = True**

E otteniamo i risultati ottimizzati di MSE e RMSE:

* **MSE: 0.8064**
* **RMSE: 0.8980**

### CONFRONTO DEGLI ERRORI MSE E RMSE: MODELLI OTTIMIZZATI K-NN E SVD

Immagine che contiene testo, Carattere, schermata, algebra

Descrizione generata automaticamente

Ora che abbiamo ottimizzato entrambi i modelli possiamo confrontare i loro valori degli errori MSE e RMSE:

* Il modello ottimizzato SVD ha un MSE e RMSE **leggermente minore** rispetto al modello KNN.
* Il modello SVD ottimizzato mostra una **performance più accurata** rispetto al modello KNN.

Sebbene KNN si basa sulla similarità diretta tra utente e prodotto questo modello potrebbe non essere adatto a dataset con molte celle mancanti come in questo caso, per questo motivo l’algoritmo SVD funziona meglio: permette di catturare pattern latenti complessi.

### TOP 10 PRODOTTI CONSIGLIATI PER L’UTENTE “AGTMZCWIWBH43TCW7UKG2YV2EKKA”

Immagine che contiene testo, schermata, Carattere

Descrizione generata automaticamente

Utilizzando lo stesso procedimento come con il modello KNN troviamo il top 10 prodotti consigliati per lo stesso utente: AGTMZCWIWBH43TCW7UKG2YV2EKKA.

Ci rendiamo conto che ogni prodotto di questo top è completamente diverso al top precedente calcolato con il modello KNN.

### FILLING DELLA MATRICE DI RATING: MATRIX FACTORIZATION

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

Applichiamo la stessa logica per fare il filling della matrice dei rating così come abbiamo fatto con il modello KNN.

* Aggiungiamo uno zero se il prodotto è già stato valutato.
* Altrimenti, appendiamo la predizione del prodotto non ancora recensito tramite il metodo .*predict().*

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamenteOtteniamo la matrice fillata con i rating predetti per i prodotti non ancora valutati. Di conseguenza, procediamo a generare la lista degli n prodotti consigliati per ogni utente:

Possiamo osservare che tutte le liste sono completamente differenti alle liste generate tramite il modello KNN.

Il motivo per il quale possiamo evincere risultati diversi può essere al fatto della natura degli algoritmi in questione.

Da una parte, l’algoritmo KNN è un **modello locale** che si concentra sulle relazioni immediate tra utenti o prodotti, il che può essere più adatto in contesti in cui la similarità tra utenti/prodotti ha una grande importanza; Dall’altro lato, l’algoritmo SVD è un **modello globale** che cerca pattern generali nella matrice dei rating, il che può essere più efficace in scenari con dati sparsi (come nel nostro caso).

In sostanza, il modello SVD si è mostrato più adatto nel nostro caso in quanto la nostra matrice dei rating presentava dati molto sparsi, ecco perché abbiamo riscontrato valori di MSE e RMSE più bassi rispetto all’algoritmo KNN.

# Conclusioni

Abbiamo sviluppato un sistema di raccomandazione basato su **Collaborative Filtering** utilizzando due principali approcci: **K-Nearest Neighbors (KNN)** e **Matrix Factorization (SVD)**. Attraverso i sei step previsti, abbiamo analizzato, implementato e confrontato entrambi gli algoritmi, ottenendo i seguenti risultati principali.

1. Analisi Esplorativa dei Dati

L’analisi ha evidenziato una forte prevalenza di rating uguale a 5 (Quasi il 60%) e una minoranza di rating uguale a 2 (Il 4.2%). Inoltre, la matrice dei rating presentava una forte sparsita’: Più del 95% degli utenti e dei prodotti aveva meno di 10 recensioni, quindi, ha dovuto essere filtrato. Questo ha sottolineato l’importanza di scegliere algoritmi capaci di gestire dati sparsi.

1. Ottimizzazione dell’Algoritmo KNN

Tramite l’analisi delle combinazioni di similarità, del numero di vicini K e dalla modalità user-based e item-based, abbiamo individuato la configurazione ottimale: **Similarità: MSD, K = 24 e Item-based.** Questa configurazione ha ottenuto un **MSE: 0.6823** e un **RMSE: 0.9286.**

1. Filling della matrice dei rating e Raccomandazioni con KNN

Utilizzando il modello KNN, abbiamo riempito la matrice dei rating predicendo i valori mancanti e generato liste di raccomandazioni personalizzate per ogni utente. Questo approccio ha prodotto risultati adeguati anche se fortemente influenzati dalla similarità locale.

1. Segmentazione degli utenti tramite K-Means

Abbiamo segmentato gli utenti in base alle loro preferenze utilizzando l’algoritmo di clustering K-Means con cosine similarity utilizzando la matrice dei rating iniziale. Ci ha permesso di individuare **20 clusters** tramite l’elbow method in cui potevamo evincere la distribuzione non uniforme degli utenti tra i diversi clusters: Esistenza di gruppi di utenti con preferenze più specifiche e di gruppi di utenti con preferenze più comuni.

1. Matrix Factorization

Abbiamo applicato la Matrix Factorization tramite SVD, ottimizzando i parametri principali. La configurazione migliore ha ottenuto un **MSE di 0.8064** e un **RMSE di 0.8980**, superando il modello KNN in termini di accuratezza.

1. Confronto tra KNN e SVD

Confrontando i due approcci, il modello SVD si è dimostrato più preciso nella predizione dei rating mancanti (minori MSE e RMSE).

In sostanza, il sistema di raccomandazione sviluppato si è dimostrato più accurato utilizzando l’algoritmo SVD anziché l’algoritmo K-NN in quanto è più adatto a sistemi con dati scarsamente correlati (Come in questo caso in cui più del 95% dei prodotti/utenti aveva meno di 10 recensioni), grazie alla sua capacità di sfruttare informazioni latenti.