

**Sistema Intelligente di Raccomandazione e Analisi del Dominio Manga**

Caso di Studio di “Ingegneria della Conoscenza”

**Studente**: *Antonello Isabella*

**Matricola**: *737827*

**E-Mail**: *a.isabella1@studenti.uniba.it*

**URL Repository**: [*LINK*](https://github.com/AIsabella1/ICON)

AA 2024-2025

**Indice**

[**1 Introduzione** 3](#_Toc201336629)

[**2** **Architettura del sistema** 3](#_Toc201336630)

[**3 Raccolta e preparazione dei dati** 4](#_Toc201336631)

[**3.1** **Accesso API MyAnimeList** 4](#_Toc201336632)

[**3.2** **Dataset generati** 5](#_Toc201336633)

[**3.3** **Dataset Apprendimento Supervisionato** 5](#_Toc201336634)

[**4** **Knowledge Base** 6](#_Toc201336635)

[**4.1** **Fatti** 6](#_Toc201336636)

[**4.2** **Generazione da CSV** 7](#_Toc201336637)

[**5** **Motore di raccomandazione** 7](#_Toc201336638)

[**5.1** **Regole di raccomandazione** 7](#_Toc201336639)

[**5.2** **Menu interattivo** 12](#_Toc201336640)

[**6** **Apprendimento supervisionato** 13](#_Toc201336641)

[**6.1** **Preprocessing** 14](#_Toc201336642)

[**7.1** **Classificazione** 15](#_Toc201336643)

[**7.1.1** **Decision Tree** 15](#_Toc201336644)

[**7.1.2** **Random Forest** 18](#_Toc201336645)

[**7.1.3** **AdaBoost** 21](#_Toc201336646)

[**7.1.4** **KNN** 24](#_Toc201336647)

[**7.1.5** **Naive Bayes** 26](#_Toc201336648)

[**7.1.6** **XGBoost** 29](#_Toc201336649)

[**7.1.7** **Analisi Complessità** 32](#_Toc201336650)

[**7.2** **Target Piace** 32](#_Toc201336651)

[**7.3** **Confusion Matrix e Radar Plot** 33](#_Toc201336652)

[**9** **Risultati e confronto finale** 34](#_Toc201336653)

[**9.1** **Analisi simbolico vs statistico** 34](#_Toc201336654)

[**9.2** **Riflessioni sui modelli supervisionati** 35](#_Toc201336655)

[**10** **Conclusioni** 35](#_Toc201336656)

[**10.1** **Riepilogo del lavoro** 35](#_Toc201336657)

[**10.2** **Estensioni future** 36](#_Toc201336658)

[**11** **Appendice** 36](#_Toc201336659)

[**11.1** **Riferimenti** 36](#_Toc201336660)

# **1 Introduzione**

Questa documentazione presenta la progettazione e lo sviluppo di un sistema per l’analisi e la raccomandazione di contenuti nel dominio dei manga.  
L’obiettivo principale è combinare **AI Simbolica** e **AI SubSimbolica**, integrando:

* Dati reali provenienti dalle **API** pubbliche di [MyAnimeList](https://myanimelist.net/apiconfig/references/api/v2#section/);
* Tecniche di **apprendimento supervisionato**;
* Un motore di raccomandazione basato su **regole esplicite**.

Questo permette di costruire un sistemain grado di fornire raccomandazioni basandosi sulle preferenze dell’utente.

Gli **obiettivi** includono:

* Automatizzare il recupero di dataset personalizzati;
* Generare una knowledge base con fatti strutturati da dati reali;
* Implementare un motore di raccomandazione basato su regole;
* Addestrare modelli di classificazione supervisionata per stimare la probabilità di gradimento;
* Produrre visualizzazioni chiare e metriche comparative tra approcci (le immagini generate sono state inserite nella documentazione).

# **Architettura del sistema**

L’architettura del progetto è **modulare, estendibile** ed è suddivisa in componenti indipendenti, ciascuno responsabile di una fase specifica del progetto.  
Questa separazione consente di isolare chiaramente le attività di raccolta dati, ragionamento simbolico e analisi statistica.

Ogni componente ha una propria directory dedicata:

**Immagine che contiene testo, schermata, Carattere

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

Il contenuto generato dall'IA potrebbe non essere corretto.**

# **3** **Raccolta e preparazione dei dati**

La fase di raccolta e preparazione dei dati è fondamentale per garantire l’efficacia delle analisi successive.  
L’obiettivo è ottenere i **dataset**, partendo da fonti reali, sia per i moduli di apprendimento automatico che per quelli simbolici.

## **3.1 Accesso API MyAnimeList**

L’accesso ai dati dell’utente e alla top mille avviene tramite il protocollo **OAuth 2.0 con PKCE**, supportato dalle API di MyAnimeList (MAL). Questo consente di ottenere dati **aggiornati** senza esporre le credenziali dell’utente.

Utilizziamo le seguenti credenziali ottenute da [MAL](https://myanimelist.net/apiconfig) per effettuare le successive richieste:

* **CLIENT\_ID**, ID applicazione fornito da MAL;
* **CLIENT\_SECRET**, secret key privata;
* **REDIRECT\_URI**, dove ricevere la risposta.

Nel codice del progetto sono presenti le credenziali generate da me medesimo, così da poter eseguire i vari codici senza registrarsi al sito. Inoltre, è stata seguita la seguente [guida](https://myanimelist.net/blog.php?eid=835707) per impostare l’accesso ai dati.

Il processo si compone di tre fasi:

1. **Autenticazione,** lo script genera un **code\_verifier** (per lo scambio del token), costruisce **l’URL** (l'URL completo è visibile nello script) di autorizzazione e **apre automaticamente** il browser per permettere all’utente di concedere i permessi;

Immagine che contiene testo, Carattere, schermata, linea

Il contenuto generato dall'IA potrebbe non essere corretto.

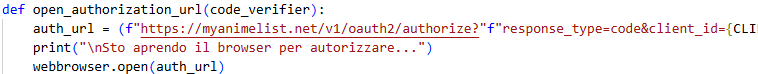


Immagine che contiene testo, schermata, Carattere, numero

Il contenuto generato dall'IA potrebbe non essere corretto.

1. **Cattura del codice,** un server HTTP locale (su localhost:8080) riceve il parametro code da MAL dopo l’autorizzazione;
2. Immagine che contiene testo, schermata, Carattere, numero

   Il contenuto generato dall'IA potrebbe non essere corretto.**Scambio del codice,** lo script invia una richiesta POST con il codice ricevuto per ottenere un **access\_token**, necessario per accedere agli endpoint.

## **3.2 Dataset generati**

Durante l’autenticazione e le chiamate API, vengono generati due dataset principali in formato CSV, utili per il motore di raccomandazione.

Il primo **top\_manga**, ègenerato dallo script top\_manga.py, contiene una classifica dei manga più popolari su MyAnimeList, ottenuta interrogando l’endpoint /v2/manga/ranking.

Ogni riga include:

* ID, Titolo, Generi;
* Punteggio medio, Rank, Popolarità;
* Stato (es. finished);
* Autori (nome e cognome).

L’API consente di estrarre 500 manga alla volta dalla top, aspettiamo 1 secondo dopo ogni richiesta per evitare il rate limit (e che il dispositivo sia segnalato come possibile Bot).

Questo dataset fornisce una visione oggettiva dei manga.

Primi cinque manga estratti dalla top mille: 

Il secondo **mangalist,** è generato da user\_manga.py, contiene la lista personale dei manga dell’utente, ottenuta dall’endpoint /v2/users/{username}/mangalist.

Per ogni manga letto:

* ID, Titolo, Generi;
* Stato di lettura;
* Punteggio dato dall’utente.

L’API per questo tipo di richiesta consente di estrarre 100 manga alla volta dalla lista utente, come per il dataset precedente anche in questo caso attendiamo 1 secondo per evitare il rate limit.

Questo file è pensato come base per l’analisi delle preferenze utente. Tuttavia, **non viene usato per l’apprendimento supervisionato,** dato che questo tipo di richiesta non consente di estrarre il punteggio globale, rank e popolarità.

Primi cinque manga estratti dell’utente ([MAL Utente](https://myanimelist.net/profile/beam_mopanies/statistics/manga-scores)):

Immagine che contiene testo, schermata, ricevuta, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

Consiglio di leggere le due funzioni direttamente dagli script, in top\_manga si chiama **get\_top\_manga,** mentre in user\_manga si chiama **get\_user\_mangalist**.

## **3.3 Dataset Apprendimento Supervisionato**

Il dataset finale “**dataset\_ml**”, usato per l’apprendimento supervisionato, è costruito dallo script user\_estesa.py.  
La costruzione del dataset è gestita dallo script, che:

* Effettua l’autenticazione via OAuth2;
* Estrazione lista manga dell’utente;
* Aggiunge a ciascun manga mean, rank e popularity.

Vengono esclusi i manga nello stato “**plan\_to\_read**”, poiché non ancora valutati dall’utente.

Per questo script dato che l’estrazione è molto lenta, e talvolta la connessione viene interrotta, è stato aggiunto un sistema che consente di effettuare un **retry automatico** delle richieste GET con lo stesso offset, così da non ripartire da zero.

Per estrarre le informazioni l’API consente di ottenere un massimo di 100 manga alla volta, come per i dataset precedenti abbiamo un’attesa fra una richiesta ed un'altra di 1 secondo.

Primi cinque manga estratti dell’utente ([MAL Utente](https://myanimelist.net/profile/Stark700/statistics/manga-scores)):

Immagine che contiene testo, Carattere, bianco, ricevuta

Il contenuto generato dall'IA potrebbe non essere corretto.

Come per I dataset precedenti consiglio di leggere la funzione direttamente dallo script, **get\_user\_mangalist\_extended** è una versione estesa di **get\_user\_mangalist** dove ho aggiunto sia un controllo in caso di errore 5xx, generando file temporanei per salvare l’offset attuale e i manga estratti sino a quell’offset, sia una nuova richiesta per ottenere le informazioni aggiuntive necessarie per il machine learning.

# **4 Knowledge Base**

La **Knowledge Base** rappresenta la componente simbolica del sistema, adatto alla definizione di regole e fatti logici interrogabili.

Essa raccoglie e struttura le informazioni derivanti dai dataset, in particolare:

* Manga presenti nella top globale;
* Manga letti dall’utente, con punteggio e stato.

La KB ha tre finalità principali:

* Rappresentare la conoscenza sotto forma di **fatti, interrogabili**;
* Consentire **inferenze** sulle preferenze dell’utente;
* Costituire la base per un **motore di raccomandazione**, sviluppato nel capitolo successivo.

Questa rappresentazione simbolica consente di eseguire **query** come:

* “**Esistono manga premiati che l’utente non ha ancora letto?**”;
* “**Quali generi compaiono più frequentemente nella sua cronologia?**”.

## **4.1 Fatti**

La **base di conoscenza** è composta da due principali **tipi di fatti**, che rappresentano:

1. Le **informazioni oggettive** (estratte da **top\_manga.csv**);
2. Le **preferenze personali dell’utente** (estratte da **mangalist.csv**).

Il predicato ***manga/8*** descrive ciascun manga presente nella **classifica globale**. La struttura è la seguente:

* **ID**: identificativo numerico del manga;
* **Titolo**: nome del manga (come **atomo testuale**);
* **Generi**: lista di **atomi** che rappresentano i generi (es. [action, fantasy]);
* **Mean**: **punteggio medio globale** (numero);
* **Rank**: posizione nella **classifica globale;**
* **Popolarità**: indice di **popolarità su MAL;**
* **Stato**: **stato editoriale** (es. finished, publishing);
* **Autori**: lista di **nomi** (come atomi).

Il predicato ***lettura\_utente/5*** rappresenta i **manga letti** (o pianificati) dall’utente, con le sue **valutazioni** e il **proprio stato di lettura**. La struttura è:

* **ID**: identificativo del manga;
* **Titolo**: titolo **normalizzato;**
* **Stato**: **stato di lettura** (es. reading, completed, plan\_to\_read);
* **PunteggioUtente**: valore da **1 a 10** assegnato dall’utente, altrimenti 0 se l’utente non ha fornito un punteggio o non ha ancora letto il manga;
* **Generi**: lista dei **generi associati.**

Questi due fatti costituiscono l’intera base interrogabile dal motore, consentendo **inferenze** **personalizzate**, **raccomandazioni**, **raggruppamenti** **per** **generi** e tanto altro ancora.

***Struttura dei fatti***

Ogni riga di **top\_manga.csv** viene trasformata in un fatto **manga/8**, esempio:



Ogni riga di **mangalist.csv** produce invece un fatto **lettura\_utente/5**, esempio:



## **4.2 Generazione da CSV**

Questo modulo legge i file CSV generati nella fase di raccolta dati (**top\_manga.csv** e **mangalist.csv**) e li converte in una serie di **fatti**, scritti nel file **knowledge\_base.pl,** generati automaticamente dallo script **crea\_kb.py**.

***Input***

* **top\_manga.csv**: contiene i manga della **top mille globale;**
* **mangalist.csv**: contiene i **manga letti e valutati** dall’utente.

***Operazioni principali***

1. **Lettura dei file CSV** tramite csv.DictReader;
2. **Pulizia dei dati**: normalizzazione dei nomi con **safe\_string()**, rimozione spazi, gestione di **valori mancanti;**
3. **Parsing dei campi multipli**: trasformazione di **generi** e **autori** in **liste;**
4. Scrittura di ogni fatto in knowledge\_base.pl.

Lo script utilizza la funzione **safe\_string()** per garantire la compatibilità sintattica con **Prolog**, evitando problemi dovuti a **spazi**, **caratteri speciali** o **virgolette**.

# **5 Motore di raccomandazione**

Sviluppato in **Prolog**, consente di sfruttare la **conoscenza** nella base dei fatti per eseguire ragionamenti interpretabili e generare raccomandazioni personalizzate.

Attraverso un insieme di **regole logiche definite manualmente**, il motore è in grado di:

* Identificare i generi preferiti dall’utente;
* Suggerire manga non letti ma in base ai gusti personali dell’utente;
* Valutare la compatibilità di un manga in base alla frequenza dei generi letti;
* Eseguire diverse forme di raccomandazione, come la scoperta di generi nuovi, contenuti premiati o “perle nascoste”.

Il motore è accessibile tramite un **menù interattivo**, che guida l’utente tra le funzionalità del sistema, offrendo un'esperienza **esplicativa** e **guidata** nell'esplorazione dei contenuti suggeriti.

## **5.1 Regole di raccomandazione**

Il motore, implementato in **Prolog** nel file **system.pl**, definisce un insieme di **regole** che operano sui fatti presenti nella **knowledge base** per generare **raccomandazioni personalizzate**. L’interazione avviene tramite il predicato **menu/0**, che offre un’interfaccia testuale con otto opzioni numerate, ognuna corrispondente a una specifica **funzionalità del sistema**.

Di seguito si riportano le principali regole implementate.

***1 - Visualizza i generi preferiti (ordinati per frequenza)***

Questa funzionalità mostra i generi dei manga effettivamente letti dall’utente, ordinati per frequenza decrescente.

Immagine che contiene testo, Carattere, schermata, linea

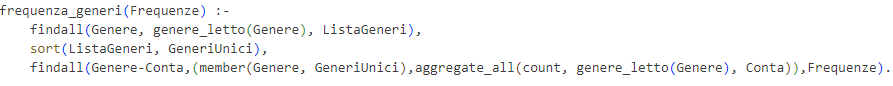
Il contenuto generato dall'IA potrebbe non essere corretto.

Immagine che contiene testo, Carattere, schermata, bianco

Il contenuto generato dall'IA potrebbe non essere corretto.Il processo si articola in più fasi:

* Immagine che contiene testo, schermata, Carattere, linea

  Il contenuto generato dall'IA potrebbe non essere corretto.Vengono estratti tutti i generi associati ai manga letti (escludendo quelli con stato plan\_to\_read) tramite il predicato genere\_letto(GenerePulito), che applica anche una normalizzazione per rimuovere l’underscore ad inizio genere;
* I generi estratti vengono deduplicati e ordinati alfabeticamente;
* Per ciascun genere, viene calcolata la frequenza con cui appare nei manga letti, producendo una lista nella forma genere-numero;



* Infine, la lista viene ordinata in ordine decrescente sulla base della frequenza, tramite il predicato generi\_ordinati/1.

Questo output consente di identificare i generi più apprezzati.

Esempio di output (la lista stampata è molto più lunga):  
Immagine che contiene testo, schermata, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

***2 - Consiglia 5 manga basati sui gusti più frequenti***

Questa funzionalità suggerisce manga non ancora letti, appartenenti ai generi preferiti dell’utente (letti almeno dieci volte, ho scelto questa misura sia per questa regola che per le successive per filtrare i manga con generi poco interessanti per un utente che ne ha letti pochi o abbastanza, confrontando la media dei lettori). I manga sono selezionati in modo casuale tra quelli compatibili.

Immagine che contiene testo, Carattere, linea, schermata

Il contenuto generato dall'IA potrebbe non essere corretto.

Il processo si sviluppa come segue:

* I generi vengono ordinati per frequenza tramite generi\_ordinati/1 e filtrati per selezionare solo quelli con almeno dieci letture (Count >= 10);
* Per ciascuno di questi generi, si costruisce una lista di manga che contengono quel genere e che non sono ancora stati letti (\+ lettura\_utente(...));
* La lista complessiva di candidati viene deduplicata e mescolata casualmente (random\_permutation/2);
* Infine, vengono selezionati casualmente 5 manga da proporre all’utente.

Questa raccomandazione si basa sui generi più letti, mantenendo però la casualità che permette la scoperta di titoli potenzialmente nuovi.

Esempio di output:

Immagine che contiene testo, Carattere, schermata, linea

Il contenuto generato dall'IA potrebbe non essere corretto.

***3 - Consiglia 5 manga di qualità ma poco popolari basati sui gusti più frequenti***

Suggerisce manga non ancora letti dall’utente che rispettano due condizioni: un punteggio medio elevato (≥ 8) e una popolarità bassa (> 1500, più alto è il valore, più i manga saranno poco popolari).

Il processo si articola così:

* Vengono considerati i generi letti almeno dieci volte;
* Per ciascun genere, si cercano manga che:
  + abbiano una media voto (Mean) maggiore o uguale a 8;
  + abbiano una popolarità (Pop) superiore a 1500, ovvero siano meno noti;
  + non siano già stati letti (\+ lettura\_utente(...)).
* I risultati vengono deduplicati, mescolati casualmente, e tra questi vengono selezionati i titoli.

Immagine che contiene testo, schermata, Carattere, algebra

Il contenuto generato dall'IA potrebbe non essere corretto.

Esempio di output:

Immagine che contiene testo, schermata, Carattere, linea

Il contenuto generato dall'IA potrebbe non essere corretto.

***4 - Consiglia 5 manga dalla tua lista "plan\_to\_read" basati sui gusti più frequenti***

Suggerisce titoli presenti nella lista “plan\_to\_read” dell’utente, che risultano compatibili.

Opera nel seguente modo:

* Immagine che contiene testo, schermata, Carattere, numero

  Il contenuto generato dall'IA potrebbe non essere corretto.Vengono considerati i generi letti almeno dieci volte;
* Per ogni manga nella lista “plan\_to\_read”, il sistema confronta i suoi generi con i generi preferiti dell’utente;
* Se almeno il 50% (così da consigliare manga con generi poco letti oltre quelli letti) dei generi del manga coincide con i generi preferiti, il titolo viene considerato compatibile e proposto come raccomandazione.

La funzione intersection/3 viene utilizzata qui e in regole successive per contare i generi in comune.

Esempio di output:

Immagine che contiene testo, Carattere, schermata, algebra

Il contenuto generato dall'IA potrebbe non essere corretto.

***5 - Consiglia 5 manga premiati basati sui gusti più frequenti***

Consiglia manga non ancora letti che hanno ricevuto un riconoscimento (award\_winning) e che condividono almeno due generi con quelli più letti dall’utente.

Il sistema procede come segue:

* Immagine che contiene testo, schermata, Carattere, numero

  Il contenuto generato dall'IA potrebbe non essere corretto. Vengono considerati i generi letti almeno dieci volte;
* Esamina ciascun manga con tag “award\_winning” e verifica che:
  + non sia già stato letto;
  + contenga almeno due generi tra quelli preferiti.
* Prima del confronto, i generi vengono normalizzati per evitare problemi di formato (underscore iniziali);
* I manga che soddisfano i criteri vengono proposti come raccomandazioni, includendo titolo, autori e stato.

Esempio di output:

Immagine che contiene testo, Carattere, schermata, linea

Il contenuto generato dall'IA potrebbe non essere corretto.

***6 - Consiglia 5 manga che combinano generi basati sui gusti più frequenti e generi mai letti***

Consiglia manga non ancora letti, ovvero una combinazione di generi letti frequentemente e generi completamente nuovi.

Il funzionamento è il seguente:

* Viene estratta l’intera lista dei generi presenti nella knowledge base;
* Si individuano:
  + I generi preferiti, letti almeno dieci volte;
  + I generi mai letti.
* Per ogni manga non ancora letto, si verifica se:
  + contiene almeno un genere preferito;
  + contiene almeno un genere mai letto.
* I generi dei manga vengono normalizzati prima del confronto;
* Solo i manga che soddisfano entrambi i criteri vengono inclusi nelle raccomandazioni.

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

Il contenuto generato dall'IA potrebbe non essere corretto.***

Esempio di output:

Immagine che contiene testo, Carattere, schermata, linea

Il contenuto generato dall'IA potrebbe non essere corretto.

***7 - Valuta la compatibilità di una lista di generi rispetto alle preferenze***

Dati uno o più generi forniti dall’utente, valuta quanto questi siano compatibili con i gusti di lettura già espressi, in base alla frequenza con cui ciascun genere è stato letto in passato.

La procedura funziona così:

* Viene calcolata la frequenza di ciascun genere letto dall’utente (escludendo i manga in plan\_to\_read);
* Immagine che contiene testo, Carattere, schermata, numero

  Il contenuto generato dall'IA potrebbe non essere corretto.Ogni genere fornito viene normalizzato (rimozione underscore iniziali) e confrontato con le frequenze note;
* A ciascun genere viene assegnato un punteggio:
  + ≥ 20 letture → 3 punti;
  + 10–19 letture → 2 punti;
  + 5–9 letture → 1 punto;
  + < 5 o mai letto → 0 punti.
* Viene calcolata la media aritmetica dei punteggi ottenut;
* In base alla media risultante, restituisce un giudizio sintetico:
  + Media ≥ 2 → Molto compatibile;
  + Media ≥ 1 → Abbastanza compatibile;
  + Media < 1 → Poco compatibile.

Immagine che contiene testo, schermata, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

Esempio di output:

Immagine che contiene testo, Carattere, schermata, linea

Il contenuto generato dall'IA potrebbe non essere corretto.

Immagine che contiene testo, Carattere, schermata, linea

Il contenuto generato dall'IA potrebbe non essere corretto.

Immagine che contiene testo, Carattere, schermata, linea

Il contenuto generato dall'IA potrebbe non essere corretto.  
***8 -* *Verifica lettura di un manga specifico***

Permette all’utente di inserire il titolo di un manga e verifica se è stato già letto/non letto/vorrebbe leggerlo.

Immagine che contiene testo, Carattere, schermata, algebra

Il contenuto generato dall'IA potrebbe non essere corretto.Funzionamento:

* L’utente digita il nome del manga;
* Il titolo viene **normalizzato** sostituendo gli spazi con underscore, in modo da renderlo compatibile con quanto scritto nella knowledge base;
* Verifica se esiste un fatto lettura\_utente/5 corrispondente al titolo:
  + Se sì, mostra lo **stato della lettura** (es. reading, completed) e il **punteggio assegnato**.
  + Se no, comunica che il manga non è stato letto;
  + Se è un manga plan\_to\_read, avvisa l’utente.

Esempio di output:







## **5.2 Menu interattivo**

Il **sistema di raccomandazione** offre un **menu interattivo**, sviluppato all'interno del file system.pl, che consente all’utente di ricevere **suggerimenti** in base alle proprie preferenze.

Immagine che contiene testo, schermata, Carattere, numero

Il contenuto generato dall'IA potrebbe non essere corretto.Per avviare il menu, è necessario consultare il file system.pl, quindi digitare i seguenti comandi:

Una volta eseguito, presenta una **lista di opzioni**, ognuna delle quali attiva una specifica **regola di raccomandazione o analisi**. Selezionando un’opzione (tramite il numero corrispondente, ovvero per la prima opzione scrivere ‘1.’ senza le virgolette), il sistema esegue la regola associata e **mostra il risultato a video**.

Immagine che contiene testo, schermata, Carattere, numero

Il contenuto generato dall'IA potrebbe non essere corretto.

Le opzioni del menu sono gestite tramite il predicato **esegui\_scelta(N)**, dove N rappresenta l'opzione selezionata. Ad esempio, l'invocazione di:

Immagine che contiene testo, schermata, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

esegue il seguente:

* attiva la regola raccomanda\_random/1;
* seleziona **cinque risultati casuali**;
* li visualizza a schermo;
* infine, **richiama il menu principale** per consentire nuove interazioni.

# **6 Apprendimento supervisionato**

In questo capitolo descriviamo l’applicazione di tecniche di **apprendimento supervisionato** con l’obiettivo di prevedere se un manga **possa piacere** **o meno**, sulla base di caratteristiche **numeriche** e **semantiche**.

L’obiettivo è costruire un **classificatore (binario)** capace di stimare se un manga possa piacere (1) o meno (0) a un determinato utente, utilizzando come feature:

* i **generi** associati all’opera (codificati binariamente);
* e alcune **metriche globali** tratte dal dataset (punteggio medio, rank, popolarità).

Questa fase rappresenta la componente **SubSimbolica** del sistema: l’apprendimento si basa sull’analisi dei dati dell’utente attraverso modelli statistici in grado di generalizzare le sue preferenze.  
I risultati costituiscono una base **predittiva** utile sia per la **raccomandazione diretta**, sia per l’**integrazione con il motore di raccomandazione**.

## **6.1 Preprocessing**

Dopo l’estrazione del **dataset\_ml**, i dati vengono **puliti** e **trasformati** per risultare compatibili con gli algoritmi di **apprendimento automatico**:

* **Filtraggio**: vengono considerati solo i manga valutati dall’utente (score > 0);
* **Tokenizzazione dei generi**: i generi sono convertiti in **liste uniformate** (es. "Action, Drama" => ["action", "drama"]);
* **Binarizzazione dei generi**: tramite **MultiLabelBinarizer**, ogni genere diventa una **colonna booleana** (1 se presente, 0 altrimenti), utile per assegnare più generi ad un singolo esempio;
* **Normalizzazione numerica**: **punteggio medio**, **rank** e **popolarità** sono trattati come **variabili numeriche continue;**
* **Oversampling della classe minoritaria** (randomico), dato che l’utente ha più punteggi minori di 7 rispetto a quelli maggiori uguali a 7.

Questa fase avviene nel file **apprendimento\_supervisionato.py**:

Immagine che contiene testo, schermata, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

Immagine che contiene testo, Carattere, schermata

Il contenuto generato dall'IA potrebbe non essere corretto.

Il file risultante è un **dataset** (matrice di feature (X) con target binario (y)), in cui le **colonne** combinano:

* **Feature binarie,** una per ogni genere;
* **Feature numeriche**, come Punteggio Medio, Rank e Popolarità;
* Un **target opzionale,** costituito dal Punteggio Utente e/o dalla **classe binaria** (Piace/Non Piace) per la **classificazione.**

## **6.2 Target Piace**

Per trasformare il problema del gradimento in una classificazione binaria supervisionata, è stata definita una variabile target denominata **Piace**, che assume valore:

* 1 se il manga ha ricevuto un **punteggio medio (score)** **maggiore o uguale a 7** da parte dell’utente;
* 0 in tutti gli altri casi.

Questa soglia è stata scelta sulla base delle seguenti considerazioni:

* Nei principali siti di valutazione di anime e manga (es. **MyAnimeList**, **AnimeClick**), il punteggio 7 rappresenta una soglia interpretata come **“buono”** o **“gradito”**, a differenza di voti inferiori che indicano disinteresse o insoddisfazione.
* L’analisi esplorativa del dataset ha mostrato che la **distribuzione dei punteggi** presenta una **curva centrata attorno a 6.5–7**, rendendo la soglia 7 un punto discriminante utile per separare classi bilanciate con sufficiente densità.
* Soglie più alte (es. ≥8) avrebbero ristretto eccessivamente la classe positiva, causando sbilanciamento e perdita di informazioni. Soglie più basse (es. ≥6) avrebbero incluso molti titoli neutri o mediocri, riducendo la qualità predittiva.

## **6.3 Classificazione**

### **7.1.1 Decision Tree**

Il **Decision Tree** è un algoritmo di classificazione supervisionata che rappresenta il processo decisionale tramite una struttura ad **albero gerarchico**.  
In questa struttura:

* ogni **nodo interno** rappresenta una condizione su una feature del dataset;
* ogni **ramo** corrisponde a un valore o a una soglia della feature;
* ogni **foglia** rappresenta una classe di output (in questo caso: Piace = 1 oppure Non piace = 0).

L’albero viene costruito in modo ricorsivo: ad ogni passo, l’algoritmo seleziona la feature e la soglia che **massimizzano la purezza** dei sottoinsiemi risultanti.

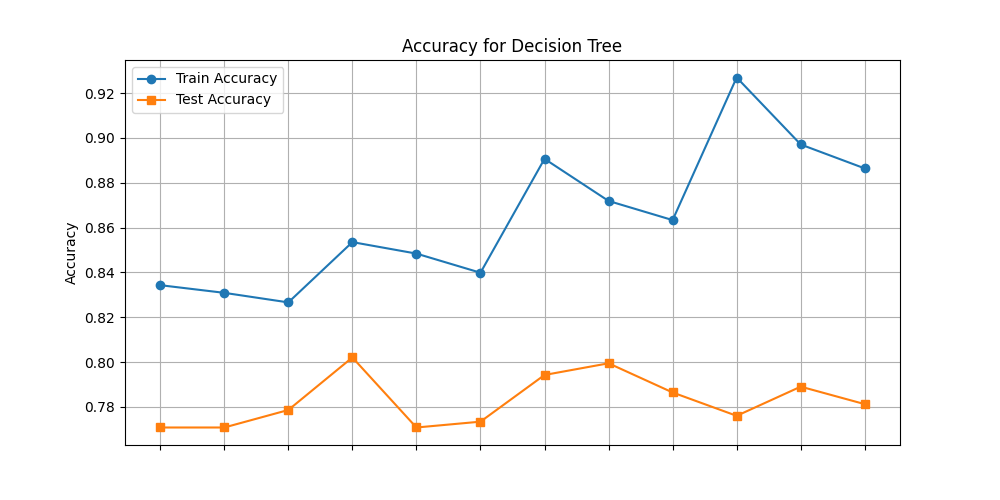
***Iperparametri principali***

* **max\_depth**: profondità massima dell’albero. Un valore troppo elevato può portare a overfitting; troppo basso può generare underfitting;
* **min\_samples\_leaf**: numero minimo di campioni richiesti in un nodo foglia;
* **splitter**: strategia usata per scegliere la feature su cui dividere ("best" o "random").

***Motivazioni della scelta del modello***  
Il Decision Tree è stato selezionato come **modello iniziale** per l’analisi supervisionata in quanto:

* è **semplice da implementare** e **altamente interpretabile**;
* consente una **visualizzazione intuitiva** del processo decisionale;
* rappresenta una **baseline efficace**, utile per confrontare le prestazioni con metodi più complessi (come Random Forest o XGBoost);
* offre la possibilità di osservare con chiarezza il fenomeno di **overfitting**, facilmente controllabile tramite i suoi iperparametri principali.

***Grafico delle Accuratezze***

Il grafico mostra l’andamento dell’**accuracy sul training set** (linea blu) e sul **test set** (linea arancione) per diverse combinazioni di iperparametri del Decision Tree:

* max\_depth: {4, 6, 8, 10};
* min\_samples\_leaf: {1, 5, 10};
* class\_weight: 'balanced'.

**Osservazioni principali:**

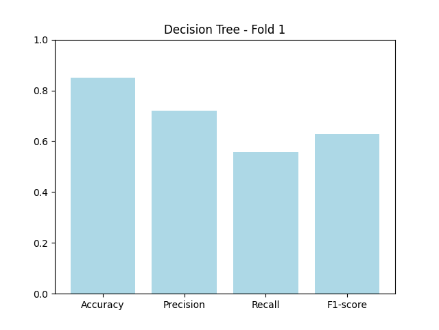
* **Overfitting contenuto**:  
  L’accuracy sul training set mostra valori elevati (fino al 93%), ma non raggiunge mai il 100%, segno che l’albero non ha memorizzato completamente i dati.  
  La distanza tra training e test accuracy è visibile, ma non estrema: questo indica che l’overfitting è presente ma **ben gestito** grazie ai vincoli introdotti;
* **Effetto della profondità (max\_depth)**:  
  Al crescere della profondità, l’accuracy sul training set tende ad aumentare (come atteso), ma quella sul test set **non segue lo stesso andamento**.  
  Le performance migliori sul test si osservano per profondità intermedie (es. max\_depth=6–8), suggerendo che un modello troppo profondo cattura rumore e perde capacità di generalizzazione;
* **Effetto della dimensione minima delle foglie (min\_samples\_leaf)**:  
  Valori maggiori di min\_samples\_leaf contribuiscono a ridurre la complessità del modello, limitando gli split su insiemi molto piccoli.  
  Questo ha un **effetto** migliora la stabilità dell’accuracy sul test set;
* **Test accuracy stabile ma non eccelsa (~77–80%)**:  
  Indica che il modello ha raggiunto un buon compromesso tra bias e varianza, ma la capacità predittiva è ancora **limitata**, giustificando l’esplorazione di modelli più sofisticati.

**Conclusione**

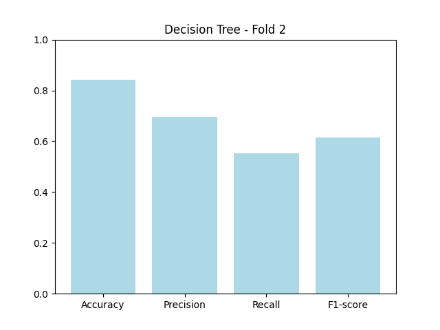
Il grafico conferma che, grazie all’uso degli iperparametri (max\_depth, min\_samples\_leaf, class\_weight='balanced'), è stato possibile **contenere l’overfitting**, mantenendo buone prestazioni sul test set.  
Questa configurazione rende il Decision Tree un valido punto di partenza, sebbene siano auspicabili modelli più robusti per migliorare ulteriormente la generalizzazione.

***Valutazione tramite Cross Validation (5 Fold)***

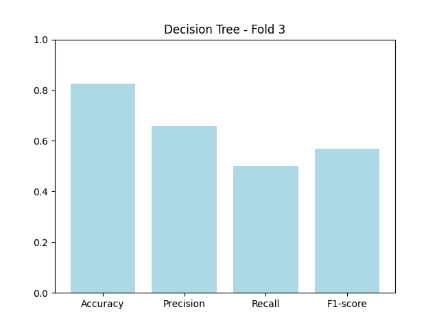
Per valutare la stabilità e le capacità predittive del modello Decision Tree, è stata eseguita una **validazione incrociata a 5 fold**. Di seguito vengono riportati i risultati per ciascun fold, accompagnati da un’analisi delle variazioni osservate tra le diverse partizioni del dataset.

**Fold 1**

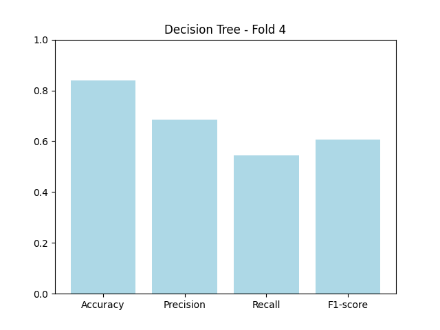
* **Accuracy**: 0.849;
* **Precision**: 0.721;
* **Recall**: 0.557;
* **F1-score**: 0.628.

**Fold 2**

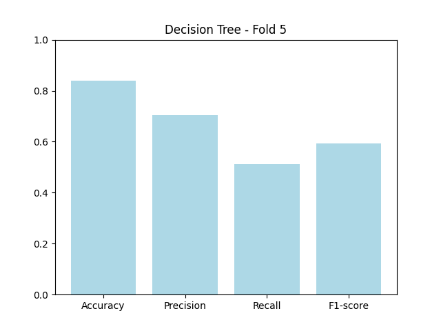
* **Accuracy**: 0.843;
* **Precision**: 0.696;
* **Recall**: 0.552;
* **F1-score**: 0.615.

**Fold 3**

* **Accuracy**: 0.825;
* **Precision**: 0.657;
* **Recall**: 0.500;
* **F1-score**: 0.568.

**Fold 4**

* **Accuracy**: 0.838;
* **Precision**: 0.686;
* **Recall**: 0.545;
* **F1-score**: 0.608.

**Fold 5**

* **Accuracy**: 0.838;
* **Precision**: 0.703;
* **Recall**: 0.511;
* **F1-score**: 0.592.

**Variazioni tra i fold**

* **Accuracy**:  
  L’accuratezza si mantiene relativamente stabile in tutti i fold, oscillando in un intervallo ristretto tra **0.825 e 0.849**. Il valore massimo si registra nel Fold 1, mentre il minimo nel Fold 3. Questa stabilità indica che il modello riesce a generalizzare bene anche su suddivisioni differenti del dataset;
* **Precision**:  
  Anche la precisione presenta una buona costanza, variando tra **0.657 (Fold 3)** e **0.721 (Fold 1)**. Il modello tende a essere più preciso nei fold in cui anche l’accuracy è più alta. Questo suggerisce che, quando il modello è più sicuro nella classificazione, tende anche a commettere meno falsi positivi;
* **Recall**:  
  Il **recall è la metrica che mostra maggiore variabilità** tra i fold, passando da **0.500 (Fold 3)** a **0.557 (Fold 1)**. Questo comportamento evidenzia come la capacità del modello di identificare correttamente i manga graditi possa risentire della particolare suddivisione del training set. Tuttavia, non si rilevano fluttuazioni eccessive;
* **F1-score**:  
  Di conseguenza, anche l’F1-score mostra variazioni limitate, con valori compresi tra **0.568 (Fold 3)** e **0.628 (Fold 1)**. La tendenza segue quella del recall, ma in forma più smussata.

**Considerazioni finali**

Nel complesso, le performance del Decision Tree risultano **coerenti tra i diversi fold**, con variazioni contenute.  
La metrica più instabile è il **recall**, seguita dall’**F1-score**, il che conferma una leggera tendenza del modello a **sottostimare i positivi**.

### **7.1.2 Random Forest**

La **Random Forest** è un algoritmo di classificazione supervisionata basato su un **ensemble di alberi decisionali**.  
Ogni albero viene addestrato su un **sottoinsieme casuale del dataset** (tecnica del *bootstrapping*) e, a ogni nodo, utilizza solo un **sottoinsieme casuale di feature** per determinare la suddivisione (*bagging* + *feature randomness*).

La predizione finale è ottenuta mediante **voto di maggioranza** tra i singoli alberi.

Questa strategia consente di:

* **Ridurre l’overfitting** tipico del singolo Decision Tree;
* **Aumentare la robustezza e la stabilità** del modello;
* Ottenere stime **più generalizzabili**, anche in presenza di dati rumorosi.

***Iperparametri principali***

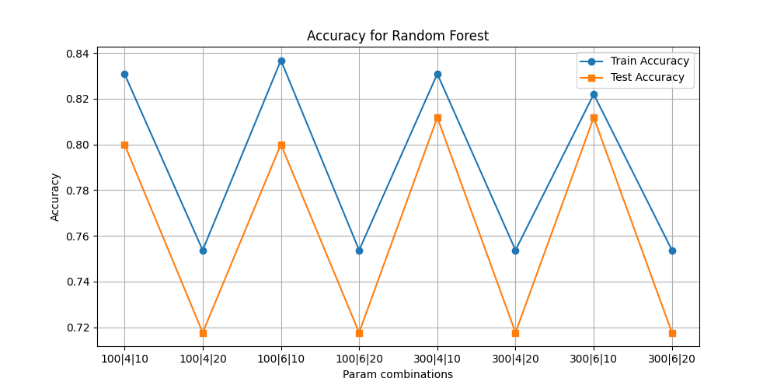
* **n\_estimators**: numero di alberi nella foresta. Un numero maggiore migliora la stabilità delle predizioni, a scapito del tempo di calcolo;
* **max\_depth**: profondità massima degli alberi;
* **min\_samples\_leaf**: numero minimo di campioni richiesti in un nodo foglia. Aiuta a prevenire overfitting sui singoli alberi;
* **max\_features**: numero di feature considerate per ogni split. Regola la diversificazione tra gli alberi;
* **class\_weight**: peso da assegnare alle classi per gestire dataset sbilanciati.

***Motivazioni della scelta del modello***

La **Random Forest** è stata selezionata come **secondo passo naturale** dopo il Decision Tree, per analizzare quanto un approccio ensemble possa migliorare la **capacità predittiva** e la **stabilità** del sistema.  
Il modello risulta:

* **Robusto** anche su dataset complessi o con feature ridondanti;
* **Meno incline all’overfitting**, grazie alla diversità tra gli alberi;
* Efficiente anche con un tuning limitato, rendendolo adatto a scenari reali.

***Grafico delle Accuratezze***

Il grafico mostra l’andamento dell’accuracy sul training set (linea blu) e sul test set (linea arancione) per diverse combinazioni di iperparametri della Random Forest:

* n\_estimators: {100, 300, 500};
* max\_depth: {6, 10, 12};
* min\_samples\_leaf: {1, 5, 10};
* max\_features: {'sqrt', 'log2', None};
* class\_weight: 'balanced'.

**Osservazioni principali:**

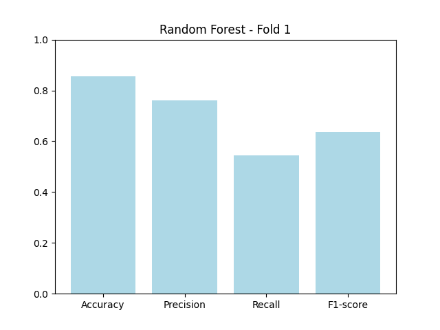
* **Overfitting contenuto**:  
  L’accuracy sul training set raggiunge valori molto alti (fino al 94%), ma resta sempre inferiore al 100%, segno che il modello non ha completamente memorizzato i dati.  
  La distanza tra le due curve è visibile ma contenuta, indicando che l’overfitting è presente, ma **ben controllato** dalla natura ensemble del modello;
* **Effetto del numero di alberi (n\_estimators)**:  
  Aumenti nel numero di alberi tendono a **stabilizzare l’accuracy**, ma non sempre migliorano la performance sul test.  
  La stabilità migliora con n\_estimators ≥ 300, anche se oltre un certo punto l’incremento è marginale;
* **Effetto della profondità (max\_depth)**:  
  Profondità maggiori portano a train accuracy più alta, ma anche a una maggiore distanza dal test set, segno di overfitting.  
  Le migliori performance sul test si osservano con max\_depth=10, valore che bilancia espressività e generalizzazione;
* **Effetto della dimensione minima delle foglie (min\_samples\_leaf)**:  
  Valori più alti per min\_samples\_leaf riducono la complessità dei singoli alberi, contribuendo a migliorare la generalizzazione e **ridurre l’instabilità tra le curve**;
* **Test accuracy stabile (~80–84%)**:  
  Pur non raggiungendo valori altissimi, la test accuracy si mantiene stabile in quasi tutte le combinazioni, confermando la **robustezza del modello** anche in presenza di parametri diversi.

**Conclusione**

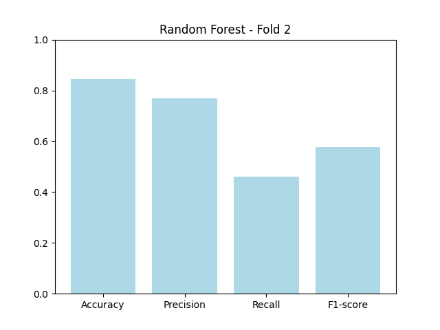
Il grafico conferma che la Random Forest, grazie all’uso combinato di bootstrapping, randomizzazione e aggregazione, **riesce a contenere l’overfitting** pur mantenendo prestazioni elevate.  
Le accuratezze sul test set indicano che si tratta di un classificatore **affidabile e ben bilanciato**, ideale per un sistema di raccomandazione che richiede stabilità e generalizzazione.

***Valutazione tramite Cross Validation (5 Fold)***

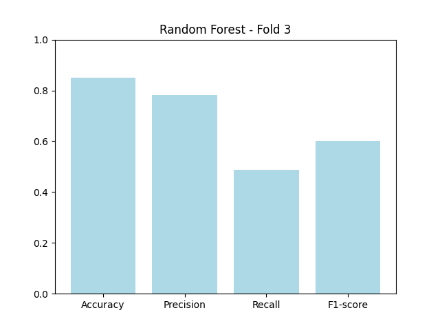
Per valutare la stabilità e le capacità predittive della Random Forest, è stata eseguita una validazione incrociata a 5 fold. Di seguito vengono riportati i risultati per ciascun fold, accompagnati da un’analisi delle variazioni osservate tra le diverse partizioni del dataset.

**Fold 1**

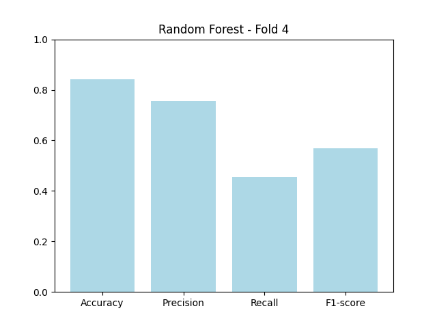
* Accuracy: 0.857;
* Precision: 0.762;
* Recall: 0.545;
* F1-score: 0.636.

**Fold 2**

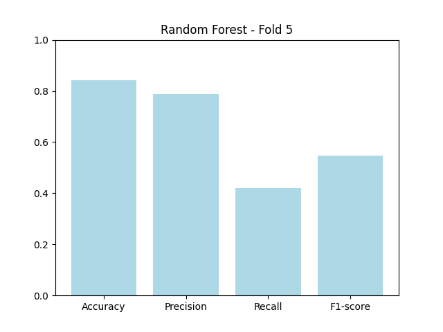
* Accuracy: 0.846;
* Precision: 0.769;
* Recall: 0.460;
* F1-score: 0.576.

**Fold 3**

* Accuracy: 0.851;
* Precision: 0.782;
* Recall: 0.489;
* F1-score: 0.601.

**Fold 4**

* Accuracy: 0.841;
* Precision: 0.755;
* Recall: 0.455;
* F1-score: 0.567.

**Fold 5**

* Accuracy: 0.841;
* Precision: 0.787;
* Recall: 0.420;
* F1-score: 0.548.

**Variazioni tra i fold**

* **Accuracy**:  
  I valori sono **molto stabili**, oscillando in un range ristretto tra 0.841 e 0.857. Il modello mantiene buone prestazioni su tutte le partizioni, indicando un’elevata capacità di generalizzazione.
* **Precision**:  
  Anch’essa stabile e sempre elevata (tra 0.755 e 0.787). La Random Forest tende a **limitare i falsi positivi**, classificando come “piace” solo i casi con alta probabilità.
* **Recall**:  
  È la metrica più variabile (da 0.420 a 0.545), suggerendo che il modello **fatica a catturare tutti i manga apprezzati**.
* **F1-score**:  
  Si mantiene tra 0.548 e 0.636, seguendo l’andamento del recall. Le variazioni non sono eccessive e confermano un buon bilanciamento tra precisione e richiamo.

**Considerazioni finali**

Nel complesso, la Random Forest **dimostra maggiore stabilità e performance globali superiori** rispetto al Decision Tree.  
Il **recall rimane il punto debole**, ma le alte precisioni e le accuratezze stabili rendono questo modello **molto affidabile** in un contesto di raccomandazione.

### **7.1.3 AdaBoost**

**AdaBoost** (*Adaptive Boosting*) è un algoritmo di classificazione supervisionata basato su un approccio **ensemble iterativo**, che combina più **classificatori deboli** (solitamente alberi decisionali a profondità uno, detti *decision stump*) per costruire un classificatore forte.

Il funzionamento si basa su un meccanismo adattivo:

* Dopo ogni iterazione, gli esempi classificati erroneamente ricevono **un peso maggiore**;
* Il classificatore successivo viene **indirizzato a correggere gli errori** del precedente;
* Ogni classificatore viene **pesato in base alla propria accuratezza**;
* Le predizioni finali si ottengono tramite **una votazione pesata** di tutti i classificatori.

Questo schema consente di concentrare l’attenzione del modello su istanze “difficili”, migliorando progressivamente le performance complessive.

***Iperparametri principali***

* **n\_estimators**: numero di classificatori deboli. Valori più alti permettono una maggiore espressività del modello, ma aumentano il rischio di overfitting;
* **learning\_rate**: coefficiente che controlla il peso assegnato a ciascun classificatore. Valori bassi rallentano l’apprendimento ma favoriscono la generalizzazione;
* **base\_estimator** (opzionale): permette di definire un classificatore personalizzato, ad esempio un Decision Tree con max\_depth > 1, che può essere configurato con ulteriori iperparametri (es. min\_samples\_split, min\_samples\_leaf, ecc.).

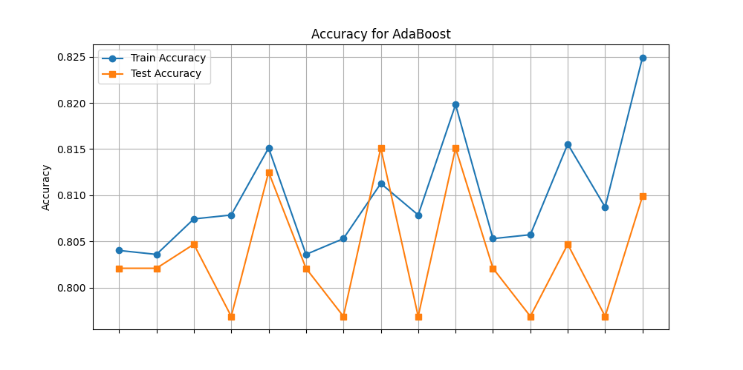
***Motivazione della scelta***

AdaBoost è stato selezionato per la sua **capacità di trasformare classificatori semplici in modelli predittivi efficaci**, sfruttando una strategia adattiva mirata a **correggere progressivamente gli errori**.

Le principali motivazioni includono:

* **Robustezza contro l’overfitting**, se adeguatamente regolarizzato (via learning\_rate o limitando la complessità del base learner);
* **Adattabilità** alle caratteristiche del dataset, soprattutto in presenza di **istanze soggettive o difficili**, come nel caso delle preferenze utente;
* **Buone prestazioni empiriche** in scenari reali, anche con dataset di dimensioni moderate.

***Grafico delle Accuratezze***

Il grafico mostra l’andamento dell’accuracy sul training set (linea blu) e sul test set (linea arancione) per diverse combinazioni di iperparametri di AdaBoost:

* n\_estimators: {50, 100, 200};
* learning\_rate: {0.05, 0.1, 0.3, 0.5, 1.0}.

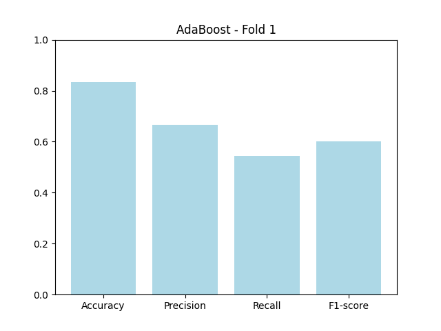
**Osservazioni principali:**

* **Overfitting contenuto**:  
  L’accuracy sul training set non supera mai il 83%, e si mantiene vicina a quella del test set. Questo comportamento è indicativo di un **modello ben regolarizzato**, che evita la memorizzazione dei dati;
* **Effetto del numero di stimatori (n\_estimators)**:  
  L’incremento di n\_estimators porta generalmente a un miglioramento della train accuracy, ma non sempre ha effetti positivi sulla test accuracy.  
  Si osserva che oltre i 100 stimatori, il beneficio tende a stabilizzarsi o addirittura a regredire in alcune combinazioni;
* **Effetto del tasso di apprendimento (learning\_rate)**:  
  Valori troppo alti (es. 1.0) possono causare oscillazioni nell’accuracy di test, sintomo di **eccessivo adattamento** ai dati.  
  I risultati più stabili si osservano con learning\_rate compresi tra **0.1 e 0.3**, che garantiscono un buon compromesso tra velocità di apprendimento e generalizzazione;
* **Accuratezza test stabile (~80–82%)**:  
  Le curve di test accuracy sono poco variabili tra le combinazioni, segno che il modello è **poco sensibile agli iperparametri**, ma anche che la sua **capacità predittiva è limitata**.

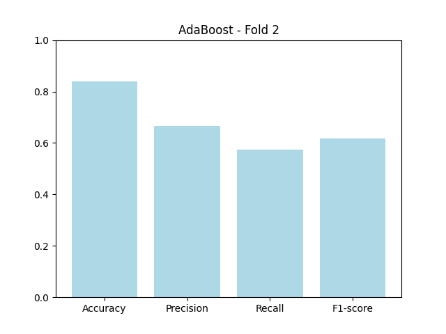
**Conclusione**

Il grafico mostra che AdaBoost è un **modello ben bilanciato**, capace di mantenere prestazioni costanti senza rischiare overfitting.  
La **regolarizzazione tramite learning\_rate** si rivela efficace per evitare eccessiva complessità. Sebbene le prestazioni siano inferiori rispetto a Random Forest o XGBoost, la **semplicità e la stabilità** del modello lo rendono un’ottima scelta in contesti moderati.

***Valutazione tramite Cross Validation (5 Fold)***

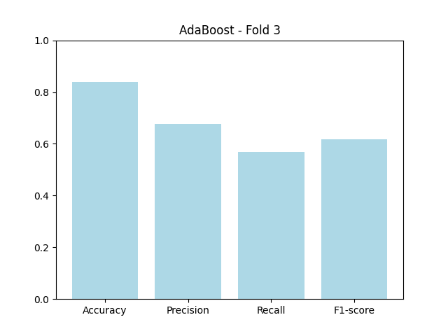
Per valutare la stabilità del modello AdaBoost, è stata eseguita una validazione incrociata a 5 fold. Di seguito vengono riportati i risultati ottenuti:

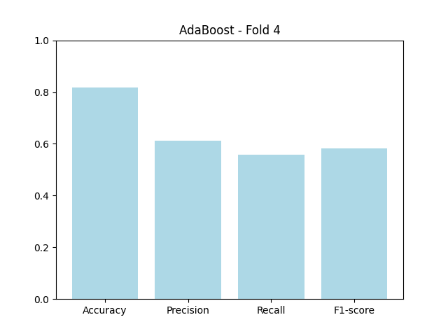
**Fold 1**

* Accuracy: 0.833;
* Precision: 0.667;
* Recall: 0.545;
* F1-score: 0.600.

**Fold 2**

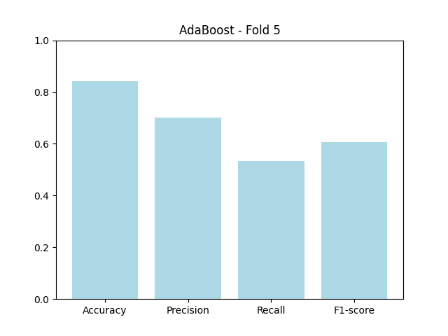
* Accuracy: 0.838;
* Precision: 0.667;
* Recall: 0.575;
* F1-score: 0.617.

**Fold 3**

* Accuracy: 0.838;
* Precision: 0.676;
* Recall: 0.568;
* F1-score: 0.617.

**Fold 4**

* Accuracy: 0.817;
* Precision: 0.613;
* Recall: 0.557;
* F1-score: 0.583.

**Fold 5**

* Accuracy: 0.841.
* Precision: 0.701;
* Recall: 0.534;
* F1-score: 0.606.

**Variazioni tra i fold**

* **Accuracy**:  
  L’accuracy mostra una buona stabilità (0.817–0.841), confermando che AdaBoost è in grado di mantenere prestazioni affidabili su differenti suddivisioni del dataset.
* **Precision**:  
  Varia da 0.613 a 0.701. Il modello tende a **contenere i falsi positivi**, mantenendo una precisione soddisfacente anche nei fold con performance inferiori.
* **Recall**:  
  Più sensibile alla variazione dei dati (0.534–0.575), ma comunque **più alto rispetto a modelli conservativi** come Random Forest.  
  AdaBoost riesce meglio a catturare gli esempi positivi.
* **F1-score**:  
  Si mantiene tra 0.583 e 0.617, dimostrando un buon equilibrio tra precisione e richiamo.

**Considerazioni finali**

AdaBoost si dimostra un modello **coerente, stabile e ben regolarizzato**.  
È meno soggetto a overfitting rispetto a modelli ensemble più complessi e fornisce una buona capacità di generalizzazione anche con un numero moderato di stimatori.

### **7.1.4 KNN**

Il **K-Nearest Neighbors (KNN)** è un algoritmo di classificazione supervisionata basato sul concetto di **similarità tra istanze**.  
Per classificare un nuovo esempio, il modello **non costruisce un modello esplicito**, ma memorizza i dati di training e, durante la fase di predizione, **identifica i *k* esempi più vicini** secondo una metrica (tipicamente la distanza euclidea). La classe viene assegnata in base alla **maggioranza dei vicini**.

***Iperparametri principali***

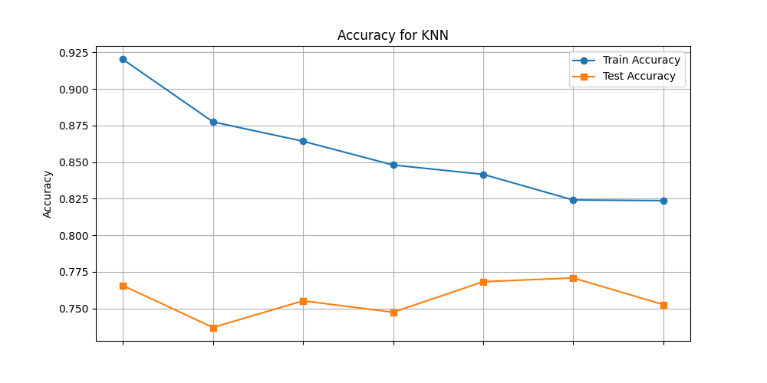
* **n\_neighbors**: numero di vicini considerati per la classificazione. Valori più alti **favoriscono la generalizzazione**, ma possono rendere **meno netti i confini decisionali**;
* **weights**: strategia di ponderazione dei vicini:
  + "uniform": tutti i vicini contribuiscono allo stesso modo;
  + "distance": i vicini più prossimi pesano di più nella decisione.
* **metric**: metrica di distanza utilizzata per misurare la similarità.

***Motivazioni della scelta del modello***  
Il **KNN è stato selezionato come benchmark non parametrico**, utile per valutare l’efficacia della classificazione **basata su prossimità tra istanze**.

In particolare:

* Permette di osservare l’impatto delle **feature numeriche e binarizzate** sulla definizione di similarità tra manga;
* Consente di analizzare la **sensibilità del sistema** rispetto alla scelta di *k*, evidenziando eventuali fluttuazioni di performance;
* Fornisce un punto di confronto con **modelli più strutturati o ensemble**, mostrando i limiti e i punti di forza di un approccio puramente geometrico.

***Grafico delle Accuratezze***

Il grafico mostra l’andamento dell’accuracy sul training set (linea blu) e sul test set (linea arancione) per diverse combinazioni del parametro n\_neighbors del modello KNN:

* n\_neighbors: {3, 5, 7, 9, 11, 15, 21}.

**Osservazioni principali:**

* **Overfitting evidente per valori bassi di k**:  
  Con k=3, l’accuracy sul training set raggiunge il 92%, mentre quella sul test si mantiene bassa (~76%). Questo indica un **forte overfitting**: il modello si adatta troppo ai dati locali del training set.
* **Effetto del numero di vicini (n\_neighbors)**:  
  All’aumentare di k, la curva di training accuracy si abbassa (segno di minore complessità del modello), mentre quella del test diventa **più stabile**, con un **massimo attorno a k=11**.
* **Test accuracy bassa e piatta (~74–77%)**:  
  Le prestazioni sul test set **non mostrano miglioramenti significativi** con l’aumento di k. Il modello fatica a generalizzare, probabilmente a causa di feature eterogenee e rumore nei dati.

**Conclusione**

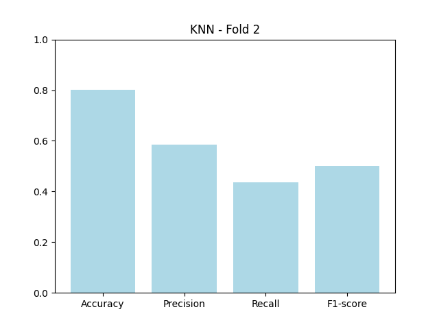
Il KNN si dimostra **molto sensibile alla scelta di k**, con prestazioni deboli sul test set e una forte tendenza all’overfitting per valori piccoli.  
L’assenza di un modello interno e la dipendenza diretta dai dati rendono KNN **inadatto a problemi in cui le feature non sono normalizzate** o i dati sono sparsi.  
Tuttavia, il modello resta un **benchmark utile per confrontare la classificazione basata su similarità pura**.

***Valutazione tramite Cross Validation (5 Fold)***

Immagine che contiene testo, schermata, diagramma, Rettangolo

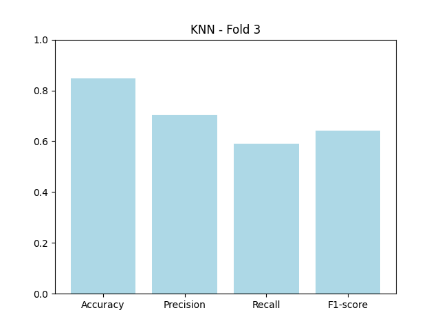
Il contenuto generato dall'IA potrebbe non essere corretto.Per valutare le prestazioni del modello KNN, è stata condotta una validazione incrociata a 5 fold. Di seguito i risultati:

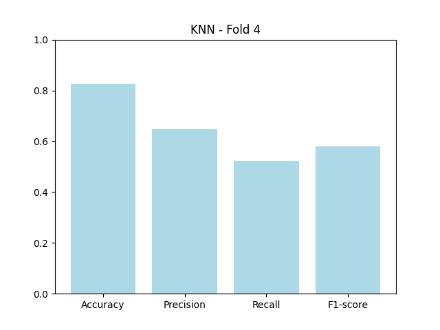
**Fold 1**

* Accuracy: 0.839;
* Precision: 0.676;
* Recall: 0.568;
* F1-score: 0.617.

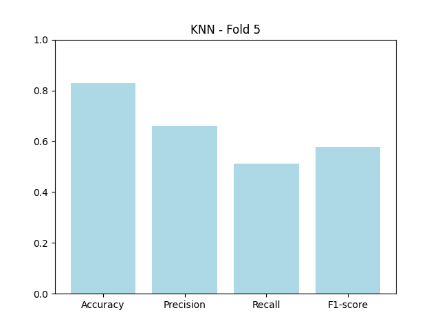
**Fold 2**

* Accuracy: 0.802;
* Precision: 0.585;
* Recall: 0.437;
* F1-score: 0.500.

**Fold 3**

* Accuracy: 0.849;
* Precision: 0.703;
* Recall: 0.591;
* F1-score: 0.642.

**Fold 4**

* Accuracy: 0.825;
* Precision: 0.648;
* Recall: 0.523;
* F1-score: 0.579.

**Fold 5**

* Accuracy: 0.828;
* Precision: 0.662;
* Recall: 0.511;
* F1-score: 0.577.

**Variazioni tra i fold**

* **Accuracy**:  
  Relativamente stabile (range: 0. 802–0.849), con un leggero calo nel Fold 2. Il modello mostra **una discreta capacità di generalizzazione**, ma non eccelle;
* **Precision**:  
  Variabile da 0.585 (Fold 2) a 0.703 (Fold 3), confermando che **l’efficacia nel classificare correttamente i positivi dipende fortemente dalla distribuzione locale dei dati**;
* **Recall**:  
  È la metrica più instabile (0.437–0.591), sintomo di **sensibilità agli outlier e ai dati poco densi**;
* **F1-score**:  
  Oscilla tra 0.500 e 0.642. Rispecchia la media tra precision e recall, rivelando **prestazioni moderate** con un equilibrio appena sufficiente.

**Considerazioni finali**

Il KNN si conferma un **modello semplice ma debole** nel contesto del problema affrontato.  
Soffre di **overfitting** con k piccoli e **scarsa generalizzazione** anche con k maggiori.  
Nonostante ciò, rappresenta un utile **termine di paragone per valutare l’efficacia di modelli più sofisticati**, soprattutto quelli basati su apprendimento strutturato o ensemble.

### **7.1.5 Naive Bayes**

Il **Naive Bayes** è una famiglia di classificatori probabilistici basati sul **Teorema di Bayes**, che assume (in maniera *naïve*) l’**indipendenza tra le feature** condizionatamente alla classe.  
Nonostante questa assunzione forte e raramente verificata nei dati reali, il modello si è dimostrato **efficace in numerosi contesti pratici**, in particolare su dataset con **feature binarie o categoriali**.

Nel nostro caso, il classificatore è stato applicato su dati misti:

* Feature binarie (generi binarizzati);
* Feature numeriche (punteggio, rank, popolarità).

***Iperparametri***

Una delle principali caratteristiche del Naive Bayes è la **semplicità di configurazione**:

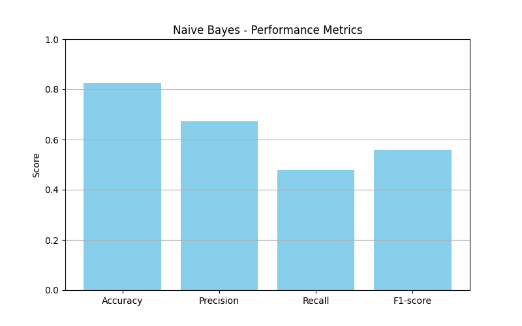
* Non è necessario effettuare tuning avanzato.
* Il modello **non ha iperparametri critici** nella configurazione standard (eccetto alpha per il Laplace smoothing, non modificato).
* La semplicità computazionale consente **addestramento e predizione rapidi**, anche su dataset estesi.

***Motivazione della scelta***

Il Naive Bayes è stato introdotto come **baseline probabilistica interpretabile**, per le seguenti ragioni:

* **Efficienza computazionale**: è tra i modelli più rapidi da addestrare ed eseguire;
* **Adattabilità a feature indipendenti**: si adatta bene alla rappresentazione binaria dei generi manga;
* **Elevato recall**: la tendenza a classificare positivi con maggiore libertà consente di **non escludere manga potenzialmente graditi**, risultando utile in un contesto di raccomandazione;
* **Trasparenza**: il modello fornisce una stima esplicita della probabilità a posteriori, facilitando l’interpretazione dei risultati.

***Grafico delle Metriche***

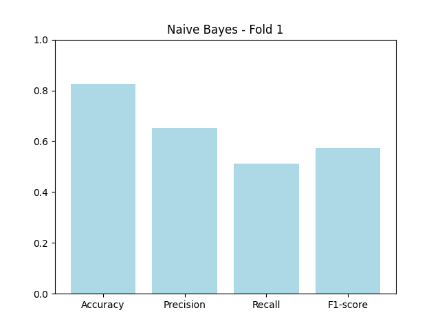
Il grafico a barre mostra le performance medie ottenute dal modello **Naive Bayes**, in termini di:

* **Accuracy**: ~82.7%;
* **Precision**: ~67.3%;
* **Recall**: ~47.9%;
* **F1-score**: ~55.8%.

**Osservazioni principali:**

* **Accuracy stabile e buona** (~82–84%):  
  Indica che il modello è in grado di classificare correttamente la maggior parte delle istanze, nonostante la semplicità dell’approccio.
* **Precisione elevata rispetto alla complessità del modello**:  
  Dimostra che, pur con assunzioni forti di indipendenza tra feature, il modello è in grado di ridurre i falsi positivi in modo consistente.
* **Recall relativamente basso (~48%)**:  
  Il modello **tende a essere conservativo**, classificando positivi solo se molto “convinto”, con il rischio di escludere alcuni manga che avrebbero potuto piacere.  
  Questo comportamento è coerente con la natura probabilistica e la forte indipendenza assunta tra le feature.
* **F1-score medio** (~56%):  
  Rappresenta un equilibrio accettabile tra precision e recall, pur penalizzato dal richiamo limitato.

***Valutazione tramite Cross Validation (5 Fold)***

Il modello è stato valutato tramite cross-validation a 5 fold. Di seguito i risultati per ogni partizione:

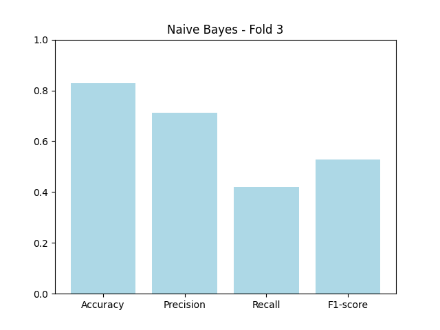
**Fold 1**

* Accuracy: 0.826;
* Precision: 0.652;
* Recall: 0.511;
* Immagine che contiene testo, schermata, diagramma, Rettangolo

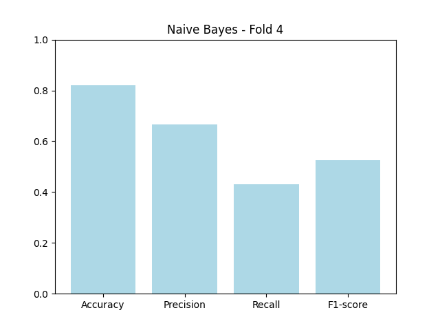
  Il contenuto generato dall'IA potrebbe non essere corretto.F1-score: 0.573.

**Fold 2**

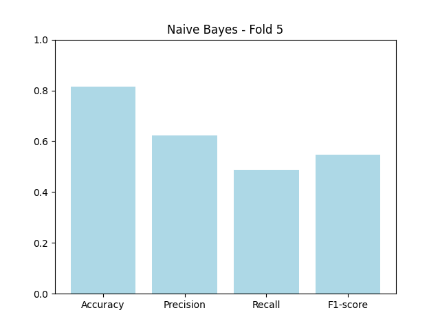
* Accuracy: 0.846;
* Precision: 0.712;
* Recall: 0.540;
* F1-score: 0.614.

**Fold 3**

* Accuracy: 0.828;
* Precision: 0.712;
* Recall: 0.420;
* F1-score: 0.529.

**Fold 4**

* Accuracy: 0.820;
* Precision: 0.667;
* Recall: 0.432;
* F1-score: 0.524.

**Fold 5**

* Accuracy: 0.815;
* Precision: 0.623;
* Recall: 0.489;
* F1-score: 0.548.

**Variazioni tra i fold**

* **Accuracy**:  
  Stabile tra 81.5% e 84.6%, confermando la **robustezza del modello** anche con dati parzialmente diversi;
* **Precision**:  
  Oscilla tra 0.623 e 0.712, dimostrando una **buona capacità di limitare i falsi positivi** nei vari fold;
* **Recall**:  
  Più instabile (range: 0.420 – 0.540), riflette la **difficoltà del modello nel catturare tutti i positivi**, specialmente in fold più sbilanciati;
* **F1-score**:  
  Varia tra 0.524 e 0.614. Segue l’andamento del recall, ma meno accentuato, confermando un bilanciamento solo moderato tra precision e recall.

**Considerazioni finali**

Il Naive Bayes si conferma un **modello semplice ma sorprendentemente efficace**.  
La sua capacità di raggiungere una buona accuratezza **senza tuning** e con una **computazione minima** lo rende un'ottima baseline.  
Tuttavia, la **limitata capacità di recall** suggerisce prudenza nel suo impiego in contesti dove è fondamentale **non perdere esempi positivi**.

### **7.1.6 XGBoost**

**XGBoost** è un algoritmo di boosting avanzato basato su **alberi decisionali**.  
Costruisce il modello in modo **sequenziale**, correggendo a ogni iterazione gli errori residui dei classificatori precedenti, tramite **ottimizzazione del gradiente**.

Rispetto ai metodi tradizionali, XGBoost introduce **numerose ottimizzazioni**:

* Gestione della **regolarizzazione** per ridurre l’overfitting;
* Strategia di **shrinkage** per evitare aggiornamenti troppo bruschi (tramite il parametro learning\_rate);
* Supporto a **subsampling** di dati e feature, per migliorare la generalizzazione;
* Elevata **efficienza computazionale**, grazie all’uso di strutture dati ottimizzate (es. DMatrix).

È un algoritmo particolarmente indicato per problemi complessi con:

* feature numeriche e categoriche;
* rumore nei dati;
* classi sbilanciate.

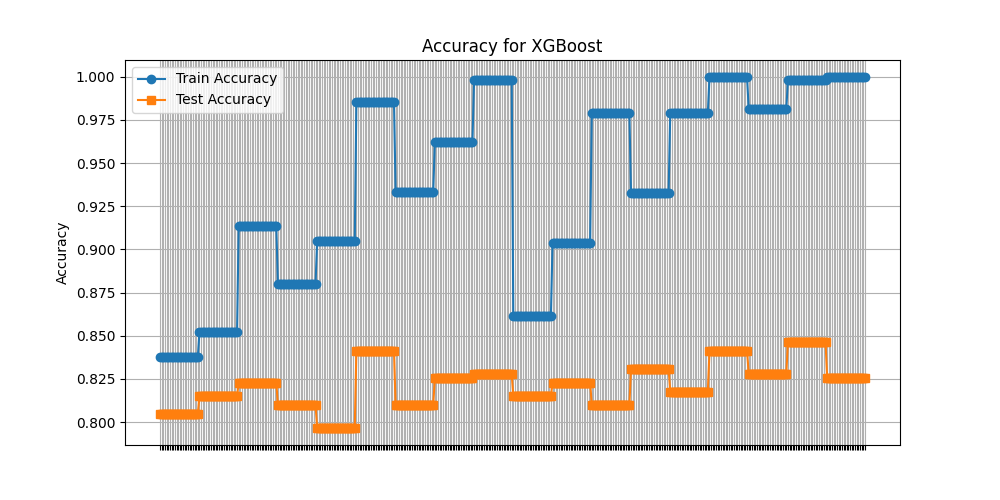
***Iperparametri principali***

* **n\_estimators**: numero totale di alberi generati. Più alberi => maggiore capacità, ma rischio di overfitting;
* **learning\_rate**: tasso di apprendimento, controlla il peso dato a ogni nuovo albero;
* **max\_depth**: profondità massima degli alberi, per limitare la complessità strutturale;
* **subsample**: frazione del dataset da usare per ciascun albero. Riduce varianza e migliora la generalizzazione;
* *(Altri iperparametri secondari: colsample\_bytree, gamma, reg\_alpha, reg\_lambda, non modificati in questa analisi).*

***Motivazione della scelta***  
XGBoost è stato selezionato per valutare un approccio di **boosting avanzato**, con le seguenti motivazioni:

* **Capacità di generalizzazione superiore** rispetto ai singoli alberi (grazie al boosting graduale);
* **Robustezza in presenza di rumore** e distribuzioni sbilanciate;
* **Regolarizzazione integrata**, utile per controllare overfitting in contesti reali;
* **Ottime prestazioni empiriche**, spesso superiori agli ensemble classici su problemi strutturati come la classificazione delle preferenze utente.

***Grafico delle Accuratezze***

Il grafico mostra l’andamento delle accuratezze su **training set** (linea blu) e **test set** (linea arancione) per diverse combinazioni di iperparametri di XGBoost, tra cui:

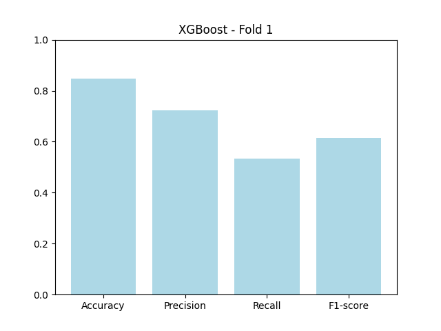
* n\_estimators: {100, 300};
* max\_depth: {3, 5, 7};
* learning\_rate: {0.05, 0.1, 0.3};
* subsample: {0.7, 0.9, 1.0};
* colsample\_bytree: {0.7, 1.0};
* scale\_pos\_weight: {1, 1.5}.

**Osservazioni principali:**

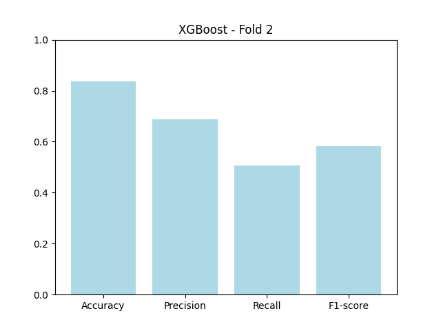
* **Overfitting visibile ma gestito:**  
  L’accuracy sul training set raggiunge spesso il 100%, ma quella sul test set resta più bassa (~81–84%).  
  Questo divario è previsto nei modelli boosting ad alta capacità, ma l’uso combinato di **regolarizzazione (reg\_alpha, reg\_lambda)**, **subsampling** e **learning rate basso** aiuta a contenerlo.
* **Effetto di learning\_rate e n\_estimators:**  
  L’aumento del numero di alberi (n\_estimators) migliora le prestazioni solo se accompagnato da un learning\_rate basso.  
  Un learning\_rate=0.05 con n\_estimators=300 tende a produrre risultati stabili e generalizzabili.
* **Contributo della regolarizzazione:**  
  L’introduzione di scale\_pos\_weight=1.5 ha migliorato leggermente il **recall**, compensando lo sbilanciamento tra le classi.
* **Stabilità delle prestazioni:**  
  Nonostante l’ampio spazio di ricerca, le combinazioni ben regolarizzate mostrano una **test accuracy stabile**, con valori attorno all’83–84%.

**Conclusione:**  
XGBoost dimostra un buon equilibrio tra accuratezza e capacità di generalizzazione, pur richiedendo attenzione alla scelta degli iperparametri. Le strategie adottate sono dimostrate efficaci nel contenere l’overfitting.

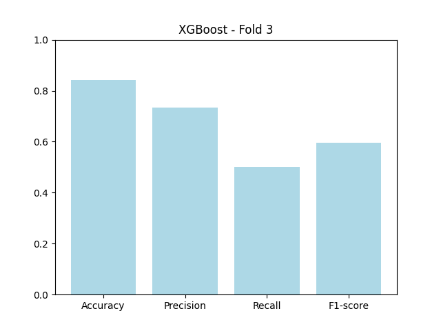
***Valutazione tramite Cross Validation (5 Fold)***

Per analizzare la robustezza del modello, è stata condotta una validazione incrociata 5-fold. I risultati per ciascun fold sono i seguenti:

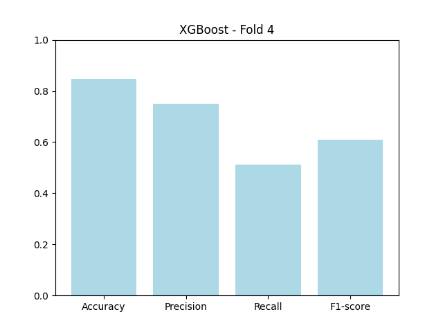
**Fold 1**

* Accuracy: 0.846;
* Precision: 0.723;
* Recall: 0.534;
* F1-score: 0.614.

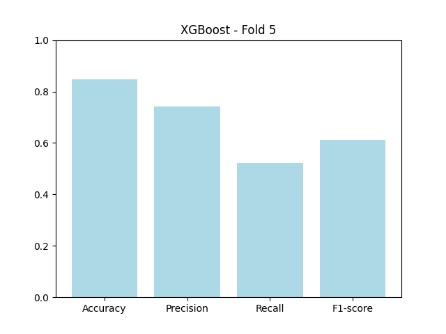
**Fold 2**

* Accuracy: 0.836;
* Precision: 0.688;
* Recall: 0.506;
* F1-score: 0.583.

**Fold 3**

* Accuracy: 0.843;
* Precision: 0.733;
* Recall: 0.500;
* F1-score: 0.595.

**Fold 4**

* Accuracy: 0.849;
* Precision: 0.750;
* Recall: 0.511;
* F1-score: 0.608.

**Fold 5**

* Accuracy: 0.849;
* Precision: 0.742;
* Recall: 0.523;
* F1-score: 0.613.

**Variazioni tra i fold**

* **Accuracy**: molto stabile tra i fold, con valori tra 0.836 e 0.849. Indica buona capacità generalizzativa.
* **Precision**: valori costantemente sopra 0.68, con un picco di 0.750 (Fold 4). Il modello è accurato nel classificare i positivi.
* **Recall**: variabile tra 0.50 e 0.53. Conferma una tendenza a privilegiare la precisione rispetto alla sensibilità.
* **F1-score**: coerente con gli altri modelli boosting (~0.58–0.61), suggerendo un bilanciamento discreto.

**Considerazioni finali**

XGBoost si conferma un modello **robusto e performante**, con accuratezza tra le più alte osservate.  
Mostra una leggera **preferenza per la precisione**, ma con configurazioni adeguate (es. bilanciamento classi), è possibile migliorare il recall.

È adatto come **modello di riferimento finale**, specialmente quando si cerca un compromesso ottimale tra **bias**, **varianza** e **scalabilità**.

### **7.1.7 Analisi Complessità**

Per ogni modello si è scelto di limitare la complessità strutturale e di bilanciare la sensibilità alle classi per mitigare overfitting e sbilanciamenti.

| **Modello** | **Tempo di training** | **Memoria** |
| --- | --- | --- |
| Decision Tree | O(n·d·log n) | O(n) |
| Random Forest | O(t·n·d·log n) (t = nestimators) | O(t·n) |
| AdaBoost | O(t·Tbase) | O(t·n) |
| KNN | O(1) training; O(n·d) inferenza | O(n·d) |
| Naive Bayes | O(n·d) | O(d·k) (k = # classi) |
| XGBoost | O(t·n·log n) | O(t·n) |

*n* = numero esempi, *d* = numero feature, *t* = numero stimatori.

**Decision Tree** presenta un costo di training di O(n·d·log n), dove n è il numero di esempi e d il numero di feature, e richiede memoria O(n) è quindi piuttosto scalabile su dataset di medie dimensioni, ma cresce più velocemente se aumentano gli attributi o i campioni.

**Random Forest**, con t stimatori, ha tempo di training O(t·n·d·log n) e memoria O(t·n); l’aumento degli alberi migliora la stabilità, ma allunga proporzionalmente i tempi e il footprint in memoria.

**AdaBoost** richiede O(t·Tbase) tempo, dove Tbase​ è il costo di training del classificatore debole (tipicamente un albero poco profondo), e O(t·n) spazio per mantenere pesi e predizioni intermedie. È più leggero di Random Forest finché gli stadi t restano contenuti.

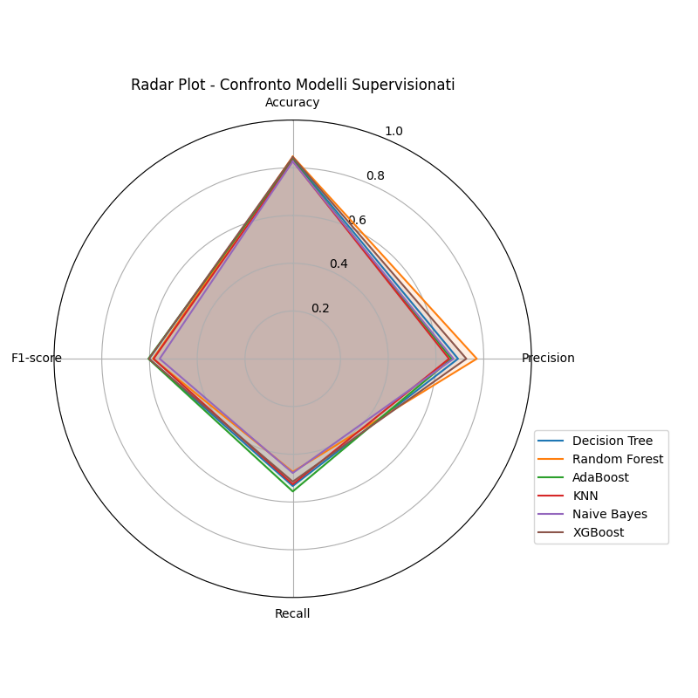
**K-Nearest Neighbors** non ha fase di training O(1), ma l’inferenza costa O(n·d) per ogni previsione e richiede memoria O(n·d) per conservare l’intero dataset, il che può diventare oneroso in fase di deployment su grandi insiemi di dati.

**Naive Bayes** scorre linearmente sui dati con O(n·d) sia in tempo che in spazio O(d·k) (con k numero di classi): ciò lo rende estremamente efficiente e adatto a scenari real-time o vincolati in memoria.

**XGBoost** fonde boosting e struttura ad albero, con tempo O(t·n·log n) e memoria O(t·n); pur generando modelli molto performanti, conviene mantenere basso t e profondità per contenere l’impatto computazionale.

Le notazioni Big-O riportate per ciascun algoritmo derivano dall’analisi delle operazioni fondamentali descritte nel libro (Poole & Mackworth, capitoli dedicati agli algoritmi corrispondenti ed internet). In sostanza, le formule standard dai testi di riferimento sono state adattate ai parametri (*n*, *d*, *t*).

## **7.3 Confusion Matrix e Radar Plot**

I grafici riepilogativi forniscono una visione sintetica ma efficace delle prestazioni dei modelli supervisionati analizzati.  
In particolare, il **radar plot** mostra le **metriche medie** (Accuracy, Precision, Recall e F1-score) dei sei classificatori:

* **Decision Tree;**
* **Random Forest;**
* **AdaBoost;**
* **KNN;**
* **Naive Bayes;**
* **XGBoost.**

**Osservazioni chiave:**

* **Random Forest** si conferma il modello con la **precision più alta**, segno della sua tendenza conservativa nel classificare positivi con alta confidenza.
* **Naive Bayes** evidenzia il **recall più elevato**, riuscendo a coprire un maggior numero di positivi a scapito di un numero maggiore di falsi positivi.
* **AdaBoost** e **XGBoost** mostrano un **profilo equilibrato**, con buone performance su tutte le metriche, risultando ideali per applicazioni generaliste.
* **KNN** e **Decision Tree** riportano **metriche più contenute**, ma comunque coerenti e stabili, utili come baseline interpretative.

Questa visualizzazione consente di **valutare visivamente i compromessi tra precisione e sensibilità**, aiutando a selezionare il modello più adatto agli obiettivi: massima precisione, elevata copertura o bilanciamento.

***Confusion Matrix***

Accanto al radar plot, la **confusion matrix** di AdaBoost mostra il comportamento del modello su un campione di test:

|  | **Predetto: 0** | **Predetto: 1** |
| --- | --- | --- |
| **Reale: 0** | 248 (TN) | 58 (FP) |
| **Reale: 1** | 13 (FN) | 65 (TP) |

**Analisi:**

* Il modello ha **classificato correttamente 248 manga non graditi** (True Negative), ma ha anche segnalato **58 falsi positivi**.
* Dei **78 manga apprezzati**, ne ha riconosciuti **65 come tali** (True Positive), mentre **13 sono stati ignorati** (False Negative).

Questo evidenzia un compromesso ben bilanciato:

* La **quantità di positivi recuperati (recall)** è buona;
* I **falsi positivi** sono contenuti;
* Il modello riesce a **identificare con successo molti manga potenzialmente apprezzati**, mantenendo basso il rumore.

In un contesto di raccomandazione, questo comportamento è auspicabile: **si riduce il rischio di consigliare titoli non graditi**, pur riuscendo a intercettare buona parte di quelli rilevanti per l’utente.

# **9 Risultati e confronto finale**

## **9.1 Analisi simbolico vs statistico**

Il progetto ha integrato due approcci distinti per la raccomandazione:

* **Apprendimento simbolico** (motore Prolog): genera raccomandazioni tramite regole deduttive esplicite (es. per autore, genere, stato, premi).
* **Apprendimento statistico** (modelli supervisionati e non supervisionati): apprende dai dati attraverso pattern e correlazioni tra feature e valutazioni.

L’approccio **simbolico** è altamente interpretabile e controllabile, ma è limitato nella capacità di generalizzare oltre quanto codificato. È ideale per esprimere preferenze logiche chiare.

L’approccio **statistico supervisionato** si è rivelato più efficace nella **predizione del gradimento**, con accuracies medie tra **0.828 (KNN)** e **0.847 (Random Forest)**.  
I modelli ensemble (Random Forest, AdaBoost, XGBoost) si sono distinti per l’equilibrio tra precisione e recall, fornendo **predizioni robuste e bilanciate** anche in presenza di dati eterogenei.

Al contrario, i metodi **non supervisionati (KMeans, GMM, Agglomerative)** non hanno prodotto risultati significativi:

* I cluster identificati da KMeans e GMM erano **fortemente sbilanciati**, con oltre 98% dei manga assegnati a un unico gruppo.
* L’Agglomerative Clustering ha mostrato **una segmentazione più bilanciata**, ma non coerente con gruppi rilevanti né utile per la raccomandazione.

Ciò conferma che, nel contesto del progetto, **l’apprendimento supervisionato è più adatto** a gestire il problema della previsione delle preferenze utente, mentre il clustering ha mostrato **limiti strutturali**.

## **9.2 Riflessioni sui modelli supervisionati**

Sono stati testati sei modelli principali di classificazione supervisionata:

| **Modello** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| --- | --- | --- | --- | --- |
| Decision Tree | 0.839 | 0.692 | 0.533 | 0.602 |
| Random Forest | 0.847 | 0.771 | 0.474 | 0.586 |
| AdaBoost | 0.834 | 0.665 | 0.556 | 0.605 |
| KNN | 0.828 | 0.655 | 0.526 | 0.583 |
| Naive Bayes | 0.827 | 0.673 | 0.479 | 0.558 |
| XGBoost | 0.844 | 0.727 | 0.515 | 0.603 |

**Considerazioni:**

* **Random Forest** ha ottenuto le **migliori performance globali**, con precisione e stabilità elevate;
* **AdaBoost** e **XGBoost** hanno mostrato **ottimi compromessi** tra accuratezza e capacità di generalizzazione;
* **KNN** ha performato bene ma mostra sensibilità alla scala e al numero di vicini;
* **Naive Bayes** è risultato molto rapido ed efficiente, ma meno potente nei casi più complessi;
* **Decision Tree**, pur interpretabile, ha sofferto di una leggera tendenza all’overfitting.

Nel complesso, i modelli ensemble (Random Forest e XGBoost) si confermano i più adatti al task.  
La classificazione è stata svolta **su feature semplici ma informative**: generi (one-hot), punteggio medio, rank e popolarità.

# **10 Conclusioni**

## **10.1 Riepilogo del lavoro**

Il progetto ha sviluppato un sistema di raccomandazione ibrido per manga, integrando approcci **simbolici** (logica Prolog) e **statistici** (modelli di apprendimento supervisionato e clustering esplorativo).  
Il sistema, progettato in modo modulare, include le seguenti componenti principali:

* Raccolta e preprocessing dei dati da **MyAnimeList**;
* Generazione automatica della **knowledge base Prolog**;
* **Motore logico-simbolico** per raccomandazioni interpretabili e filtrabili;
* Sei **modelli supervisionati** per predire il gradimento dell’utente a partire da generi e metriche numeriche;
* Un modulo sperimentale di **clustering (KMeans, GMM, Agglomerative)**, con riduzione dimensionale (PCA) e selezione automatica del numero di cluster (Silhouette, Elbow);
* Visualizzazioni finali tramite **matrici di confusione**, radar plot e proiezioni PCA.

L’integrazione dei due paradigmi ha permesso di creare un sistema intelligente, estendibile e interpretabile.  
Tuttavia, ha evidenziato che **i modelli supervisionati risultano più efficaci e affidabili** rispetto agli approcci non supervisionati.

## **10.2 Estensioni future**

Il sistema è stato progettato in modo modulare ed estensibile. Alcune possibili evoluzioni includono:

* **Sperimentare altri modelli di classificazione**;
* Integrare una **web app interattiva** per navigare la knowledge base, generare raccomandazioni logiche e visualizzare i cluster in tempo reale;
* Introdurre un **modulo di feedback utente**, per raffinare le predizioni tramite apprendimento continuo.

# **11 Appendice**

## **11.1 Riferimenti**

**Fonti teoriche e accademiche**

* Poole, D., & Mackworth, A. – *Artificial Intelligence: Foundations of Computational Agents*  
  Online: <https://artint.info>  
  Concetti chiave:
  + Logica dei predicati, Prolog
  + Apprendimento supervisionato e clustering
  + Sistemi ibridi (simbolico + statistico)

**Librerie e tool di sviluppo**

* MyAnimeList API – <https://myanimelist.net/apiconfig>
* SWI-Prolog – <https://www.swi-prolog.org/>
* Owlready2 – <https://owlready2.readthedocs.io/>
* HermiT Reasoner – <https://www.hermit-reasoner.com/>
* Scikit-learn – <https://scikit-learn.org/>
* XGBoost – <https://xgboost.ai/>
* Pandas – https://pandas.pydata.org/
* NumPy – <https://numpy.org/>
* Matplotlib – <https://matplotlib.org/>
* Seaborn – https://seaborn.pydata.org/