***SocialBook***

***Progetto Fondamenti di Intelligenza Artificiale***

***(2020/2021)***

Studenti:

Barbato Alessia (0512105858)

Proietto Angelica (0512105762)

Russo Luca (0512105840)

******

# 1.Introduzione

SocialBook si propone come un social innovativo che permette la nascita di una vasta community unita dalla passione comune per i libri, dando agli utenti la possibilità di interagire tra loro.

Ogni utente, quando si registra, inserirà alcune informazioni personali di vario tipo, in particolare sui suoi interessi e preferenze letterarie, in modo da ricevere dei suggerimenti di altri utenti “simili”, ovvero utenti che appartengono allo stesso gruppo.

Il primo passo nell’analisi del problema consiste sicuramente nella definizione dell’ambiente, che si trova immediatamente al punto successivo.

# 2.Definizione dell’ambiente

L’ambiente viene definito tramite la rappresentazione schematica PEAS(*Performance, Environment, Actuator, Sensors*):

• **Prestazioni:** Le prestazioni dell’agente vengono valutate in base all’accuratezza con cui consiglia ad un utente altri utenti a cui potrebbe essere interessato.

• **Ambiente:** L’ambiente in cui l’agente opera è composto da utenti registrati a una piattaforma riguardante la tematica “libri”, ed è completamente osservabile, deterministico, episodico, statico e discreto. Inoltre, vi è presente un singolo agente.

• **Attuatori:**

• **Sensori:** I sensori dell’agentesono costituiti da un questionario di campionamento e da quello proposto all’utente che decide di iscriversi.

# 3.Raccolta dati e introduzione al pre-processing

Il dataset è la collezione dei dati (campioni), ognuno con le proprie caratteristiche, che rappresentano il dominio di interesse; viene utilizzato per addestrare e successivamente testare l’agente intelligente progettato.

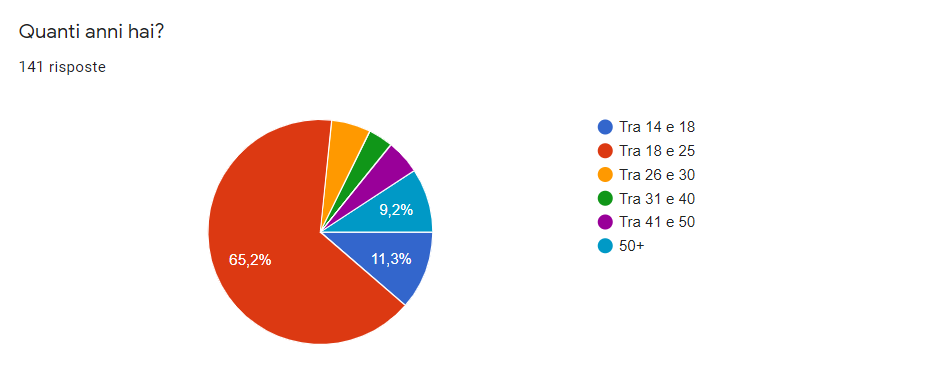
Per la natura dell’agente, il cui obiettivo è quello di “apprendere” e saper suddividere gli utenti in base alla similarità tra essi, si è reso necessario effettuare una sorta di indagine per poter raccogliere i dati di partenza, in modo da garantire il funzionamento efficace del modulo che abbiamo costruito.

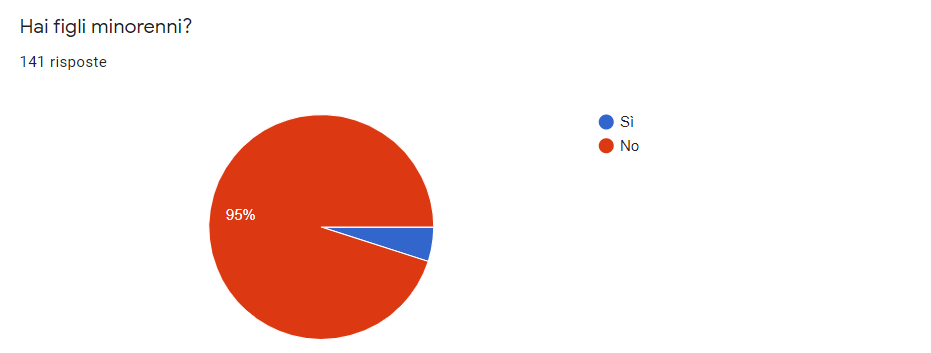
Per fare ciò, ci siamo serviti di un questionario sottoposto a quante più persone possibili, in modo da avere un dataset abbastanza ampio e non rischiare di mandare l’agente in overfitting.

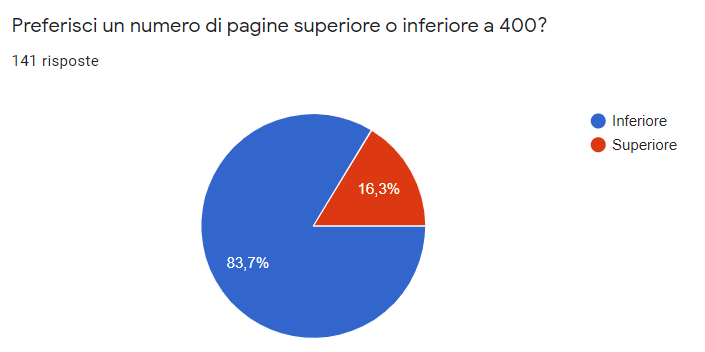
Il nostro questionario ([Socialbook (google.com)](https://docs.google.com/forms/d/1m8v-GKTujO15kvwWOejnmVg1X4RbXxKW61YECixXLTg/viewform?edit_requested=true)) raccoglie informazioni sia personali che riguardanti l’ambiente letterario.

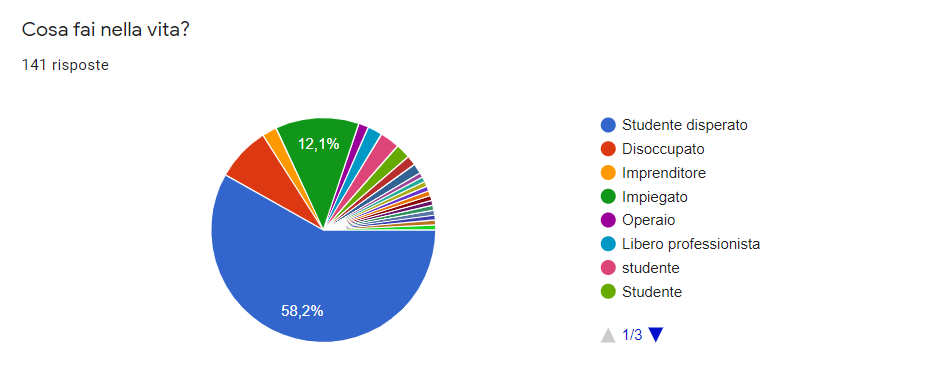
Il lettore, infatti, è caratterizzato dalle seguenti informazioni: l’età, l’essere genitori o meno di figli minorenni, l’occupazione, gli hobby, la quantità di libri letti in un anno, i generi letterari preferiti, la preferenza sul numero di pagine di un libro (inferiore/superiore a una certa soglia, pari a 400) e i criteri in base ai quali effettua la scelta del nuovo libro da leggere.

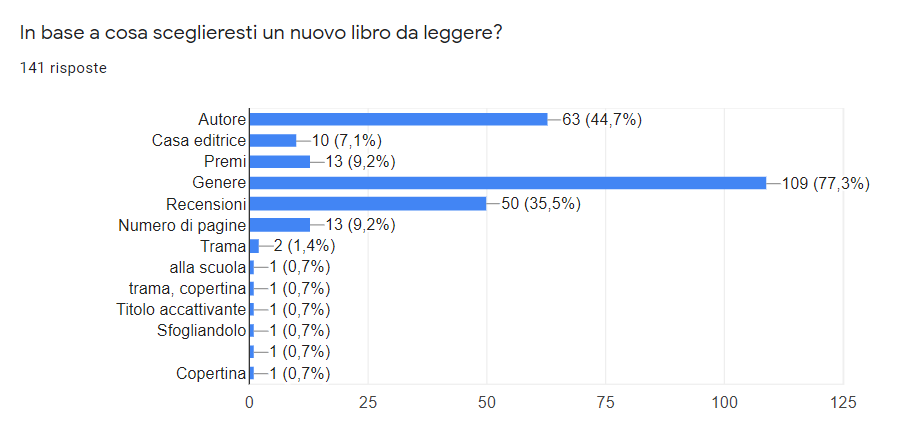
Purtroppo, non tutti i dati raccolti sono stati significativi e utili per il problema in questione, a causa di una scarsa distribuzione delle risposte. Di seguito, sono stati allegati degli esempi.











Nell’ultimo caso, abbiamo deciso di eliminare questa caratteristica per due ragioni:

• la preponderanza della risposta “genere” rispetto alle altre;

• la presenza della caratteristica “genere preferito” che può causare ridondanza.

Per questo motivo, si è deciso di filtrare il dataset di origine, con lo scopo di ridurre lo spazio di dimensione dei dati in input, andando a considerare solo le caratteristiche realmente utili per la risoluzione del problema, ovvero: hobby, numero di libri letti all’anno e genere preferito.

# 4. Definizione della tipologia di apprendimento

La profilazione utenti in questo caso si basa su campioni non associabili ad etichette. Dunque, l’unico approccio disponibile per questo tipo di apprendimento (non supervisionato) è quello del clustering, che restituisce un set di cluster contenenti degli oggetti raggruppati a seconda della similarità reciproca.

Gli algoritmi scelti e considerati per la realizzazione dell’agente intelligente in questione dovranno partizionare il dataset fino a copertura completa. Per soddisfare questa esigenza si è scelto di confrontare gli algoritmi *K-MEANS* e *DBSCAN*, entrambi esclusivi, agglomerativi e seriali.

Per questa ragione, è necessario stabilire due criteri:

• criterio di “bontà”: coefficiente di forma;

• criterio di similitudine.

## 4.1 Criterio di similitudine

Il passaggio successivo è stato quello di associare dei valori (numerici) alle varie opzioni di ogni caratteristica non numerica in base all’affinità tra loro. **//METTI DIZIONARI**

Ciò è stato fatto per permettere il corretto funzionamento degli algoritmi di clustering che si è deciso di utilizzare.

## 4.2 Algoritmi K-MEANS e DBSCAN (spiegazione e confronto)