

Studentessa: Finiguerra Alessia

Matricola: 735326

Email: [**a.finiguerra1@studenti.uniba.it**](mailto:a.finiguerra1@studenti.uniba.it)

URL GitHub:

Immagine che contiene clipart, illustrazione, grafica, cartone animato

Il contenuto generato dall'IA potrebbe non essere corretto.

***SISTEMA DI PREDIZIONE DELLA PERSONALITA’: INTROVERSO VS EXTROVERSO***

A.A. 2024/2025

Sommario

[1. INTRODUZIONE 2](#_Toc200555626)

[1.1 STRUMENTI UTILIZZATI 2](#_Toc200555627)

[2. DATA EXPLORATION E DATA PREPROCESSING 2](#_Toc200555628)

[2.1 DATASET 2](#_Toc200555629)

[2.2 VALORI DUPLICATI O NULLI 3](#_Toc200555630)

[2.3 EDA: EXPLORATORY DATA ANALYSIS 3](#_Toc200555631)

[2.4 OUTLIER 5](#_Toc200555632)

# INTRODUZIONE

La personalità di una persona è l’insieme delle caratteristiche psicologiche che influenzano il modo in cui pensa, sente e si comporta nelle diverse situazioni. Essa ci rende unici e può comprendere tratti come l’emotività, la socievolezza, la stabilità e la creatività.

Tra le molte dimensioni della personalità, una classificazione degli individui può essere quella tra persone “introverse” e persone “estroverse”: le persone introverse tendono a essere più riservate, riflessive e a preferire ambienti tranquilli, mentre quelle estroverse sono generalmente più socievoli, energiche e stimolate dal contatto con gli altri.

L’obiettivo è quello di costruire un modello in grado di prevedere accuratamente la personalità di un individuo, utilizzando le principali tecniche di apprendimento automatico.

# ARGOMENTI DI INTERESSE

Gli argomenti affrontati in questo progetto sono:

* Apprendimento supervisionato: è una tecnica in cui il modello impara da un dataset fornito in input; attraverso l’applicazione di vari algoritmi, il modello viene addestrato per essere in grado di effettuare delle previsioni accurate;
  + Apprendimento supervisionato con iperparametri: attraverso la definizione di iperaparamentri si affinano le capacità del modello in modo da massimizzare l’accuratezza delle sue previsioni
* Sistema esperto: viene creata una base di conoscenza e si applica un motore inferenziale che sarà in grado di ragionare logicamente e arrivare ad una conclusione emulando le capacità decisionali di un esperto umano

# APPRENDIMENTO SUPERVISIONATO

## STRUMENTI UTILIZZATI

Il progetto è stato realizzato nel linguaggio Python utilizzando Visual Studio Code come ambiente di sviluppo. Tra le varie librerie necessarie abbiamo:

* **Pandas:** per la manipolazione e l’analisi dei dati;
* **Matplotlib e Seaborn:** per creare e gestire i grafici;

## DATA EXPLORATION

La fase di Data Exploration è la fase iniziale di ogni progetto di machine learning. Il suo obiettivo principale è quello di esplorare il dataset per comprenderne le sue caratteristiche.

### DATASET

Per valutare e fare previsioni, un modello di machine learning ha bisogno di un set di dati da cui “imparare”, che farà quindi da base da cui costruirà la sua conoscenza.

Per predire il tipo di personalità, il nostro modello ha utilizzato il dataset “personality\_dataset.csv” (disponibile su [Kaggle](https://www.kaggle.com/datasets/rakeshkapilavai/extrovert-vs-introvert-behavior-data/data)) contenente 2900 righe e 8 colonne.

Di seguito abbiamo la spiegazione delle varie features:

* **Time\_spent\_Alone:** numero di ore che l’individuo scorre da solo ogni giorno;
* **Stage\_fear:** presenza della paura da palcoscenico;
* **Social\_event\_attendance:** frequenza di partecipazione ad eventi sociali;
* **Going\_outside:** frequenza con cui l’individuo esce;
* **Drained\_after\_socializing:** presenza dellasensazione di svuotamento dopo aver socializzato;
* **Friends\_circle\_size:** numero degli amici più intimi;
* **Post\_frequency:** frequenza di pubblicazione sui social media;
* **Personality:** questa rappresenta la variabile **target** e indica appunto se l’individuo è “Introverso” o “Estroverso”.

### VALORI DUPLICATI O NULLI

Per garantire l’integrità dei dati e prevenire un addestramento distorto del modello, è stato ripulito il dataset eliminando eventuali valori duplicati o nulli presenti al suo interno.

### EDA: EXPLORATORY DATA ANALYSIS

L’EDA (Exploratory Data Analysis) è la fase iniziale dell’analisi dei dati in cui si esplorano, visualizzano e comprendono le caratteristiche principali di un dataset.

L’obiettivo è identificare pattern, anomalie, distribuzioni e relazioni tra variabili, senza ancora applicare modelli predittivi.

Immagine che contiene testo, diagramma, schermata, Parallelo

Il contenuto generato dall'IA potrebbe non essere corretto.Per visualizzare la frequenza con cui determinati valori compaiono nel dataset, sono stati creati dei grafici per confrontare visivamente come essi sono distribuiti rispetto alle relative caratteristiche.

Figura : Features numeriche

Immagine che contiene testo, schermata, diagramma, Rettangolo

Il contenuto generato dall'IA potrebbe non essere corretto.

Figura : Features categoriche

Immagine che contiene testo, diagramma, Piano, schermata

Il contenuto generato dall'IA potrebbe non essere corretto.Successivamente, è stata effettuata l’analisi della correlazione tra le varie features con la feature target “Personality”.

Figura : Features numeriche in relazione con il target

Come si può notare dai grafici, l'analisi esplorativa conferma che l'insieme delle feature numeriche è fortemente correlato con la variabile target “Personality”.

I grafici mostrano che per ogni metrica esaminata, come il tempo trascorso da soli o la partecipazione a eventi sociali, i comportamenti degli individui estroversi e introversi si raggruppano in modo nettamente distinto, confermando che ciascuna di queste feature contribuisce a definire il loro profilo di personalità.

Immagine che contiene testo, schermata, diagramma, Diagramma

Il contenuto generato dall'IA potrebbe non essere corretto.

Figura : Features categoriche in relazione con il target

Anche le features categoriche contribuiscono in modo determinante alla definizione della personalità di un individuo. Dai grafici, infatti, si può osservare una netta separazione nei comportamenti dei due gruppi.

### OUTLIER

All’interno di un dataset possono essere presenti dei valori che a primo impatto possono sembrare anomali, i cosiddetti “outlier”, ovvero dei valori che si discostano in modo significativo dal resto delle osservazioni.

Dopo un’attenta analisi del nostro dataset, si è deciso di non procedere con l’eliminazione degli outlier identificati, in quanto si è supposto che tali valori non rappresentino errori di misurazione o di inserimento, ma piuttosto variazioni autentiche e naturali del comportamento umano.

## DATA PREPROCESSING