# Abstract

Master di 1°livello in Analisi Dati e Business Intelligence per la Data Science

Titolo: Web Scraping e Data Quality per Indici di Benessere

Tutor accademico: Bonifacio Flavio
Tutor aziendale: Baldisserri Veronica
Studente: Peduto Francesco Mattia

Il progetto svolto presso l'azienda Metis Ricerche si basa sulla necessità di creare uno strumento in grado di raccogliere, elaborare e monitorare gli indici di benessere della popolazione. Questo progetto, svolto in Python, e battezzato "SWPyS, Scrape the Web. A Python Solution", sfrutta dati e parametri già raccolti altrove, per relazionarli e per creare dei modelli che mostrino il fenomeno d'interesse in modo più trasparente possibile. Il problema di base, che questo strumento cerca di arginare, è relativo al fatto che diverse fonti possano assegnare al medesimo fenomeno d'osservazione degli indici del tutto discordanti tra loro. Di SWPyS è stata realizzato il primo processo, lo Scraping Robot, adibito alla ricerca e pulizia dei dati sul web. L'Analytic Robot, che sarà dedicato all'analisi e correlazione di questi dati, necessita invece di ulteriori sviluppi.

Il processo generale che caratterizza lo Scraping Robot può essere riassunto nelle seguenti fasi:

- Ricerca dell'argomento d'interesse (es. Indici di Benessere).
- Estrazione parole ricorrenti, semanticamente più vicine all'indagine, da usare nei processi di filtraggio.
- Estrazione link principali dalla pagina.
- Iterazione sui link principali, estraendo sublink.
- Filtraggio link non affini alla ricerca.
- Iterazione su tutti i link, raccolta di tabelle da ogni pagina.
- Filtraggio tabelle non affini alla ricerca.
- Eliminazione tabelle duplicate e strutturalmente non significative.