PROGETTI IV CALL 2023/2024

Titolo progetto:	L'offerta di servizi digitali da parte dei Comuni attraverso l'utilizzo di metodologie Site-centric
Descrizione	In linea con quanto previsto dal Regolamento (UE) 2021/241 istitutivo del Recovery and Resilience Facility, che individua nella transizione digitale uno dei sei pilastri per le strategie di rilancio delle economie europee, la digitalizzazione della Pubblica Amministrazione (PA) rappresenta una delle principali sfide individuate dalle strategie di ripresa delineate dal Piano Nazionale di Ripresa e Resilienza (PNRR). L'analisi del livello d'implementazione dei sistemi d'identità digitale a livello locale diventa cruciale per la misurazione d'impatto degli investimenti pubblici legati al processo di digitalizzazione della PA, attraverso l'erogazione al cittadino dei servizi pubblici online. La statistica ufficiale, pur fornendo un Quadro dettagliato sulle caratteristiche strutturali delle istituzioni pubbliche e l'utilizzo delle tecnologie digitali da parte della PA locale, presenta dei limiti nella diffusione tempestiva dei dati legati alla digitalizzazione. L'utilizzo di metodologie web scraping nella raccolta dei dati permette una misurazione più accurata a fenomeni legati alla digitalizzazione rispetto ai metodi tradizionali, oltre a garantire un minore disturbo statistico. Il Progetto si propone di testare e validare le tecniche di machine learning piu' idonee per sistematizzare la raccolta e l'elaborazione dei dati, riducendo i costi e aumendo la tempestivita' di diffusione delle informazioni, in vista di un loro possibile riuso per in altri ambiti tematici.
Obiettivi	Attraverso l'applicazione di metodologie Site centric, combinate con fonti di statistica ufficiale, il progetto ha l'obiettivo di fornire una misurazione del livello di digitalizzazione delle amministrazioni comunali nell'erogazione dei servizi ai cittadini e alle imprese. L'analisi dei siti web dei comuni rilevati dal Censimento permanente delle Istituzioni pubbliche (Istat, 2022) permette, attraverso metodologie di web scraping per la raccolta dei dati di testare le opportune tecniche di machine learning, al fine di misurare se e in che misura i servizi digitali della Pubblica Amministrazione vengono erogati attraverso l'adozione dell'identità digitale.
Metodologia	L'approccio metodologico ha l'obiettivo di classificare e di analizzare la capacità dei comuni italiani di offrire servizi online, con particolare attenzione all'implementazione di sistemi di identità digitale (SPID, CIE) e dei servizi telematici per l'edilizia (SUE). Gli step metodologici condotti consistono nella: **Raccolta dei dati tramite web scraping:** Lo studio utilizza tecniche di web scraping per raccogliere informazioni direttamente dai siti web dei comuni italiani. Questa procedura ha permesso di estrarre dati testuali dalle pagine web comunali e di organizzare tali dati in una Term-Document Matrix (TDM), che rappresenta la base per le successive analisi. **Costruzione di un dataset etichettato:** La creazione di una procedura automatica ha permesso di assegnare la corretta etichetta alla variabile SPID. Inoltre, per garantire una maggiore precisione nel dataset finale e' stato realizzato un approccio semi-automatico per la variabile SUE . **Applicazione di algoritmi di machine learning:** Lo studio utilizza tre modelli di apprendimento automatico supervisionato per la classificazione: *Multinomial Naive Bayes* per la classificazione di caratteristiche discrete (es. conteggio di parole). *Bernoulli Naive Bayes* per la classificazione basata sulla presenza o assenza di termini. *Random Forest,* un modello a foresta casuale che riduce la varianza e migliora la capacità di generalizzazione grazie all'aggregazione di decisioni da molti alberi decisionali. **Selezione delle caratteristiche tramite Chi-quadro:** La selezione delle variabili è stata realizzata usando il test del Chi-quadro (χ^2), che ha permesso di identificare i termini più rilevanti per la classificazione dei comuni. **Valutazione dei modelli**: I modelli sono stati sottoposti a 10 run per ciascuna configurazione, al fine di tenere conto della variabilità nei risultati e sono stati valutati attraverso metriche di accuratezza e F1-macro score, applicando una suddivisione dei dati in un set di addestramento (70%) e un set di test (30%) per garantire la validità dei risultati.

	Generalizzazione dei modelli: I modelli selezionati, una volta addestrati e validati, sono stati applicati all'intero insieme di comuni italiani per classificare la presenza dei servizi digitali (SPID e SUE).
Risultati ottenuti	Per la classificazione della presenza del servizio SUE il modello che ha dato i migliori risultati è il Bernoulli Naive Bayes, con un F1-macro score del 74,1% e un'accuratezza del 79,1%. Riguardo la classificazione della presenza del servizio SPID, il modello migliore è stato il Random Forest, con un F1-macro score del 64,8% e un'accuratezza del 76,3%. Attraverso l'integrazione con informazioni rilevate dalla statistica ufficiale, lo studio ha evidenziato importanti differenze territoriali e dimensionali nella disponibilità dei servizi online tra i comuni italiani. In particolare, l'analisi consente in maniera sistematica di indentificare la presenza dei sistemi di identità digitale (SPID, CIE, CNS) e del servizio telematico per l'edilizia (SUE) sui siti web dei Comuni, garantendo una misurazione della disponibilità di accesso dei servizi pubblici locali per cittadini e imprese.
Membri del Team	Chiara Orsini (Istat), Andrea De Panizza (Istat), Fabrizio De Fausti (Istat) Marco Conti (Presidenza del Consiglio dei Ministri), Sergio Leonardi (Presidenza del Consiglio dei Ministri)