PROGETTI V CALL 2024/2025

Titolo progetto:	Metodologia di integrazione nel sistema integrato dei registri
Descrizione	Sviluppo di un modello ML supervisionato per la determinazione della matrice di correlazione tra la componente catastale del registro dei luoghi e il registro delle imprese ASIA. Lo scopo del progetto è individuare una strategia di integrazione utilizzando le categorie catastali degli immobili e i codici ATECO delle imprese al fine di ottimizzare il matching delle unità statistiche. In particolare una parte dell'integrazione è effettuata considerando la proprietà e la locazione (giuridica o fisica) dell'unita immobiliare in cui l'impresa ha sede legale o locale ed esercita la professione (Integrazione effettuata considerando l'indirizzo dell'unita locale). Questa parte dell'integrazione copre all'incirca il 30 % della totalità delle imprese Italiane e fornisce un benchmark validato per misurare la correlazione tra ATECO e categoria catastale dell'unità immobiliare. L'idea è di utilizzare questo benchmark tramite algoritmo ML per imputare l'immobile al restante 70% di imprese che non risultano proprietarie o locatarie sfruttando appunto la correlazione, 70% che è collocato unicamente tramite l'indirizzo a tutti gli immobili presenti su quell'indirizzo.
Obiettivi	Sviluppo di un modello ML supervisionato per la determinazione della matrice di correlazione tra la componente catastale del registro dei luoghi e il registro delle imprese ASIA. Lo scopo del progetto è individuare una strategia di integrazione utilizzando le categorie catastali degli immobili e i codici ATECO delle imprese al fine di ottimizzare il matching delle unità statistiche. In particolare una parte dell'integrazione è effettuata considerando la proprietà e la locazione (giuridica o fisica) dell'unita immobiliare in cui l'impresa ha sede legale o locale ed esercita la professione (Integrazione effettuata considerando l'indirizzo dell'unita locale). Questa parte dell'integrazione copre all'incirca il 30 % della totalità delle imprese Italiane e fornisce un benchmark validato per misurare la correlazione tra ATECO e categoria catastale dell'unità immobiliare. L'idea è di utilizzare questo benchmark tramite algoritmo ML per imputare l'immobile al restante 70% di imprese che non risultano proprietarie o locatarie sfruttando appunto la correlazione, 70% che è collocato unicamente tramite l'indirizzo a tutti gli immobili presenti su quell'indirizzo.
Metodologia	Il progetto ha sperimentato l'utilizzo di strumenti low-code di Data Science e Intelligenza Artificiale per supportare i processi di integrazione tra registri statistici, con particolare riferimento all'associazione tra le unità locali delle imprese e le unità immobiliari del Registro dei Luoghi RSBL. L'approccio metodologico è stato finalizzato a valutare l'efficacia di modelli di machine learning supervisionato riducendo tempi e complessità rispetto a soluzioni tradizionali. La soluzione proposta si basa su un'architettura integrata composta da database Oracle come sistema sorgente, Denodo come livello di virtualizzazione dei dati e RapidMiner come piattaforma low-code per la gestione dell'intero flusso di data engineering e modellazione. Il processo è stato articolato in quattro fasi: acquisizione dei dati, preparazione, sviluppo del modello e analisi delle performance. La preparazione dei dati ha incluso la gestione dei valori mancanti e anomali, la standardizzazione delle variabili e l'aggregazione dei codici ATECO per ridurre la complessità del problema. La fase di modellazione ha previsto la sperimentazione di diversi algoritmi di machine learning, con selezione finale di modelli di Gradient Boosted Trees, risultati i più performanti. I dati sono stati suddivisi in training set e test set e il tuning degli iperparametri è stato effettuato tramite procedure automatizzate ad interfaccia visuale.
Risultati ottenuti	La sperimentazione ha evidenziato risultati predittivi complessivamente incoraggianti, considerando l'elevato numero di classi associate ai codici ATECO. Il modello di Gradient Boosted Trees ottimizzato ha raggiunto un'accuratezza pari al 49%, migliorando le prestazioni del modello base. Risultati più rilevanti emergono dall'analisi delle metriche Top-N: la Top-2 accuracy si attesta intorno al 71%, mentre la Top-3 accuracy raggiunge l'84%, indicando che il codice corretto è spesso incluso tra le previsioni più probabili. Sebbene i modelli non siano ancora maturi per l'utilizzo in produzione, l'output consente di ridurre significativamente la dimensionalità del problema di associazione

	tra unità locali e immobili, supportando le attività di validazione dei registri. Dal punto di vista operativo, l'impiego di strumenti low-code ha permesso di completare le attività di sviluppo, test e confronto dei modelli in poche settimane, con un notevole risparmio di tempo rispetto ad approcci tradizionali. Inoltre, l'interfaccia visuale ha favorito la collaborazione tra figure con competenze diverse, confermando la validità dell'approccio come supporto ai processi di innovazione statistica.
Membri del Team	1. Francesco Altarocca (Referente) DIRM/DCIT 2. Enrico Orsini (Referente) DIRM/DCME 3. Armando d'Aniello DIRM/DCME 4. Annunziata Fiore DIRM/DCIT 5. Domenico Aprile DIRM/DCIT 6. Andrea Pagano DIRM/DCME 7. Simonetta Cozzi DIPS/DCSE