PROGETTI IV CALL 2023/2024

Titolo progetto:	Sperimentazione di tecniche di Machine Learning per l'individuazione delle amministrazioni pubbliche definite dal Regolamento UE n. 549/2013, Sistema europeo dei conti nazionali e regionali dell'Unione europea (SEC 2010)
Descrizione	Il settore S.13, secondo il SEC 2010, include unità istituzionali pubbliche e unità istituzionali a controllo pubblico, dal comportamento "non market". Il processo di classificazione statistica delle unità nel settore delle Amministrazioni pubbliche (settore S.13) richiede l'analisi e la valutazione del profilo istituzionale e del comportamento economico di ciascuna, sulla base dei criteri qualitativi (mercato di riferimento) e quantitativi (test market/non market o del 50%) dettati dal SEC 2010. L'analisi è basata sull'integrazione di fonti eterogenee, sia amministrative sia statistiche, strutturate e non strutturate. L'accesso a fonti documentali (ad esempio, leggi, statuti, regolamenti, note allegate al bilancio) è fondamentale. La classificazione di ciascuna unità si conclude con la elaborazione, la valutazione e la sintesi delle informazioni acquisite, riferite ad un congruo periodo di tempo (almeno tre anni). L'attività di ricerca prevede la definizione e l'implementazione di: – un modello di classificazione automatica supervisionato, basato su fonti documentali per l'attribuzione delle unità al settore pubblico o al settore privato; – un modello di classificazione automatica supervisionato, per l'attribuzione delle unità al settore delle Amministrazioni pubbliche (S.13).
Obiettivi	Recuperare tempestività e accrescere l'efficienza e la qualità del processo di classificazione; ridurre la soggettività tipica delle attività di *profiling* manuale
Metodologia	Il dataset utilizzato per la costruzione del primo modello è una matrice documenti/unità x parole: a ciascuna unità istituzionale (533 unità tra consorzi di diritto pubblico e di diritto privato e fondazioni) è associato il profilo lessicometrico del relativo statuto e quindi la variabile target (pubblico/privato). Gli algoritmi di apprendimento sperimentati sono i seguenti: 1. Naive-Bayes; 2. Modello logistico (e integrazione con analisi delle specificità); 3. Rete neurale; 4. Macchine a vettori di supporto (SVM). Il dataset utilizzato per l'individuazione del secondo modello contiene indicatori di controllo pubblico e test market/non market per tre anni; contiene inoltre variabili "strumentali" utili per isolare i gruppi di unità per la classificazione delle quali il SEC 2010 detta regole "speciali". Il dataset include le unità individuate come potenziali entranti in S.13; non include le unità appartenenti al settore delle Amministrazioni pubbliche, per definizione (*core*). La costruzione del dataset ha richiesto l'integrazione di registri, fonti amministrative, fonti statistiche. Gli algoritmi di apprendimento sperimentati sono i seguenti:

	1. Albero decisionale; 2. Modello logistico; 3. Macchine a vettori di supporto (SVM).
Risultati ottenuti	In considerazione dei risultati ottenuti nell'ambito del Laboratorio Innovazione, il gruppo di lavoro ha deciso di proseguire nell'attività di ricerca, tuttora in corso. I risultati illustrati sinteticamente a seguire sono, pertanto, da considerarsi provvisori. In riferimento al primo modello (controllo pubblico/privato), l'applicazione degli algoritmi di addestramento e classificazione logistico e con reti neurali basati sulla matrice *documents x types* (533x74.800) ha dato risultati in termini di accuratezza compresi tra il 64% e il 76%. I risultati risentono della eterogeneità della natura e della qualità delle fonti, costituite prevalentemente da file di tipo *.pdf*. L'estrazione dei contenuti testuali è risultata immediata nel caso dei documenti cosiddetti nativi digitali (file *.pdf* generati dalla conversione di un documento creato tramite un word processor), mentre per quelli generati attraverso una scansione di un documento cartaceo è stato necessario effettuare il riconoscimento ottico automatico (ocr). Tale operazione, che comporta spesso l'introduzione di errori ortografici, ha reso necessaria una fase di correzione. In taluni casi, i file sono risultati inutilizzabili e quindi sono stati scartati con conseguente perdita di osservazioni/informazione. L'eterogeneità delle strutture e dei contenuti dei documenti sono un ulteriore problema che merita di essere affrontato. Sviluppi futuri pertanto riguarderanno l'uso di documenti strutturati e l'individuazione di parole chiave rilevanti per la identificazione della struttura della governance e quindi per la classificazione dell'unità. In riferimento al secondo modello (classificazione in S.13), la sperimentazione degli algoritmi di addestramento (albero decisionale, logistico e SVM) e l'attività di progressivo affinamento del dataset originario hanno consentito di ottenere risultati soddisfacenti: la misura dell'accuratezza si è attestata su valori compresi tra l'85% per l'algoritmo dell'albero decisionale e il 94% per SVM. Tra gli interventi sul dataset si segnala l'introduzione di variabili utili a identificare gruppi di unità la cui classificazione statistica non deriva dai risultati del test del 50%, ma dall'applicazione di criteri qualitativi di valutazione della "tipologia di produttore". Il trattamento di tali gruppi di unità "speciali" nello sviluppo di modelli di classificazione è attualmente oggetto di approfondimenti.
Membri del Team	Guido Borà, Fiorella Boscaino, Alessio Canzonetti, Antares D'Achille, Annamaria D'Urzo, Adriano Pareto, Leonardo Poli, Giuseppe Sacco