

UNIVERSITÀ DEGLI STUDI DI SALERNO

DIPARTIMENTO DI SCIENZA AZIENDALI - MANAGEMENT &
INNOVATION SYSTEMS



Dottorato di ricerca in Big Data Management

XXXV Ciclo

Tesi di Dottorato

Verso sistemi AI Responsabili: un approccio Metodologico per l'integrazione della Explainable AI nei Sistemi di Supporto alle Decisioni

Tutor:

**Ch.mo Prof.
Carlo Blundo**

Candidato:

**Roberto Abbruzzese
Mat. 8801500029**

Coordinatore:

**Ch.mo Prof.
Valerio Antonelli**

ANNO ACCADEMICO 2021/2022

SINTESI DELLA TESI

I processi decisionali sono fondamentali per il successo di qualsiasi organizzazione, poiché influiscono direttamente sulla capacità dell'azienda di raggiungere i suoi obiettivi. Tali processi negli anni sono diventati estremamente "Knowledge Intensive" ovvero si sono trasformati in collezioni di attività che richiedono una significativa quantità di conoscenza o competenza specialistica. Inoltre, si basano fortemente sull'esperienza, l'intuizione, il giudizio degli esperti, costruiti su grossi volumi di informazioni che una volta trattati, lavorati e analizzati si trasformano in conoscenza preziosa, utile, appunto, a promuovere decisioni efficaci ed efficienti per l'azienda. L'adozione di un supporto tecnologico, al fine di migliorare tali tipologie di processi, è rappresentato sicuramente dai Decision Support Systems (DSS). I sistemi di supporto alle decisioni sono sistemi informativi che aiutano gli utenti a prendere decisioni basandosi su analisi, di varie complessità, operate sui dati aziendali e non, grazie all'utilizzo di tecnologie di diversa natura. Negli anni i DSS hanno avuto delle evoluzioni importanti, scandite in particolare dagli avanzamenti tecnologici avuti nel dominio dell'Information Technology, con particolare accezione, negli ultimi anni, dei progressi avuti nell'ambito dell'intelligenza artificiale (IA), da questo punto in poi Artificial Intelligence (AI). L'integrazione dell'AI, nei DSS nasce, infatti, dalla necessità di gestire la crescente complessità e il volume dei dati disponibili, legata, soprattutto, alle informazioni non strutturate. L'AI,

infatti, può elaborare rapidamente grandi quantità di dati, identificare modelli e tendenze, e fornire intuizioni preziose che possono migliorare l'efficienza e l'efficacia delle decisioni. Il crescente avanzamento tecnologico del Natural Language Processing, ha abilitato, inoltre, la costruzione di servizi AI based, sempre più avanzati e sempre più complessi da gestire e da integrare nei processi. Uno degli obiettivi che l'AI ha dimostrato di perseguire in questi anni è sicuramente legato al supporto nella riduzione del carico cognitivo sui decisori umani, permettendogli di concentrarsi su aspetti più strategici e creativi del processo decisionale. Ma l'integrazione dell'AI nei DSS presenta varie sfide. Problemi legati a errori di previsione, discriminazione, privacy, sicurezza, trasparenza e dipendenza eccessiva dalla tecnologia sono stati affrontati nel dettaglio e sintetizzati nel presente lavoro di tesi, per riuscire a costruire una visione complessiva chiara e precisa del fenomeno, e abilitare la costruzione di strumenti in grado di valutare e migliorare l'impatto sociale ed etico dell'uso inappropriato dell'AI. Tale studio approfondito della letteratura, ha contemplato anche un'analisi sulle tassonomie, principi, concetti dell'Explainable AI, utili anche alla costruzione di un linguaggio comune. Al fine di lenire ed affrontare le sfide promosse dall'introduzione dell'AI nei processi decisionali, l'obiettivo principale del presente lavoro è rappresentato dalla costruzione di un framework metodologico, XAI4DSS, in grado di supportare tutti gli stakeholder del processo per promuovere l'integrazione di tecnologie e sistemi di AI nei processi, nel rispetto dei principi, regolamenti, conservando però un estremo valore aggiunto in termini di performance. Il framework metodologico viene descritto elicitando tutte le sue componenti principali, rappresentate da fasi corredate da task e possibili deliverable prodotti. In particolare il framework si divide in 6 fasi, rispettivamente:

1. **Context Understanding:** analizzare il contesto ed individuare le priorità tra le differenti possibili spiegazioni considerando prospettive diverse e fattori diversi.
2. **Requirement for data collection:** definire i requisiti utili del processo

di data collection per migliorare la qualità della spiegazione da fornire ai vari stakeholder.

3. **XAI Implementation:** sulla base dei requirement definiti nelle fasi precedenti si provvede a sviluppare le opportune tecniche di XAI ai modelli AI opportunamente selezionati.
4. **XAI Identification:** comprendere i razionali statistici che sono alla base dei modelli utilizzati.
5. **Design and Development User Explanation:** costruire e presentare le informazioni in modo chiaro e accessibile considerando i requisiti espressi dalle fasi precedenti.
6. **User Learning:** quando i decisori umani sono coinvolti in modo significativo in un risultato assistito dall'AI, devono essere adeguatamente formati e preparati a utilizzare i risultati del modello in modo responsabile ed equo.

La metodologia di ricerca, adottata per lo sviluppo del presente lavoro di tesi, ha contemplato un approccio ibrido ed integrato sia top-down che bottom-up. Infatti, se da un lato lo studio dell'Explainable AI è partito con una fase il top-down, definendo teorie e principi generali della XAI, basati su letteratura esistente e conoscenze consolidate, successivamente e contemporaneamente è stato intrapreso un percorso strutturato di sperimentazione ed estensione degli strumenti XAI attualmente presenti in letteratura. Queste sperimentazioni, opportunamente analizzate, hanno contribuito all'identificazione di modelli, tecniche e insight specifici che hanno esteso ed integrato le teorie iniziali, creando un ciclo iterativo di apprendimento e affinamento nella comprensione della XAI. Nel presente lavoro sono stati riportati tre casi d'uso che hanno contribuito alla costruzione del framework metodologico. Il primo scenario riguarda l'uso della XAI per identificare e spiegare eventi di Pump & Dump nel settore delle criptovalute, con un'analisi basata su dati in formato tabulare. Lo studio ha dimostrato la fattibilità di identificare i Pump and Dump dalle fasi iniziali, evidenziando la superiorità degli scambi finanziari

nell'identificazione grazie a dati più ricchi e semanticamente più significativi. Per quanto concerne l'applicazione di tecniche di XAI sia grazie alla Features Contribution Analysis che alla Single Instance Explanation si è riusciti a dimostrare l'importanza della nuova caratteristica social introdotta nel nuovo approccio. Inoltre, con la Single Instance Explanation, l'obiettivo è stato quello di promuovere una spiegazione concreta per ogni singola predizione, fornendo uno strumento importante e necessario per l'adozione del nuovo modello all'interno di un processo decisionale. Nel secondo scenario, è stato affrontato il tema riguardante la segmentazione di testo nei documenti legali, attraverso l'utilizzo di reti neurali di tipo transformer. L'approccio proposto si è rivelato efficace, dimostrando ottime prestazioni grazie a tre componenti fondamentali: un'architettura basata su transformer, una fase complessa ed importante di pre-elaborazione del testo e l'introduzione di una caratteristica "spaziale" sull'allineamento delle righe. Lo studio, inoltre, grazie all'analisi della robustezza del modello ha evidenziato l'importanza della generalizzazione nel processare dati reali, spesso perturbati o rumorosi. Per quanto concerne l'integrazione di modelli e tecniche XAI nelle diverse fasi di sviluppo del modello di segmentazione testuale proposto, i risultati raggiunti si sono rilevati di particolare rilevanza. Infatti, la validazione del processo di etichettatura, Bias Labeling, che ha garantito l'accuratezza e l'affidabilità delle etichette, è risultata determinate nella valutazioni delle performance, dal momento che, gli errori commessi in fase di etichettatura, rilevati nelle prime fasi di sviluppo, determinavano risultati non performanti. La dimostrazione della Model Robustness, inoltre, ha assicurato che il modello fosse resistente a varie condizioni e input imprevisti. Infine, ma non meno importante, l'applicazione della XAI per la Model Explainability per l'utente finale nasce per garantire la comprensibilità e la trasparenza dei risultati, illustrando chiaramente il processo decisionale del modello. In conclusione, il terzo scenario proposto, nasce da un'idea sviluppata per affrontare i problemi e le performance osservati nel secondo caso d'uso, grazie all'impiego della XAI. Situato nel contesto della classificazione automatica di documenti legali e dell'effetto delle anomalie causate dai siste-

mi di riconoscimento ottico dei caratteri (OCR), l'intuizione proposta mira a utilizzare la XAI per perfezionare le prestazioni di un classificatore di testo legale. Questo metodo si avvale di un'interpretazione retroattiva dei risultati di un modulo XAI al fine di correggere le anomalie nei testi scansionati tramite OCR, che influiscono negativamente sulle prestazioni dei classificatori testuali. In particolare, l'approccio fornito dal framework, denominato REMOAC, prevede l'uso di approcci e tecniche di XAI non solo per rendere accessibili l'interpretazione e la spiegazione del sistema, ma anche per impiegare in maniera "retroattiva" i risultati degli approcci XAI (come SHAP) al fine di migliorare le prestazioni di un classificatore di testi. Questo permette di ottenere risultati accurati anche in contesti reali, promuovendo una correzione mirata e specifica dei token all'interno di una frase.

Il presente lavoro, quindi, ha affrontato la capacità delle tecniche di XAI di soddisfare diversi requisiti per un'AI affidabile nel contesto dei DSS in ambito Enterprise, razionalizzando il tutto e sintetizzando quanto analizzato in un framework metodologico, capace di scandire fasi e task di un processo di integrazione della XAI, non solo per indirizzare la scelta dello strumento ottimale in base a performance ed avanzamenti tecnologici, ma per gestire la complessità e promuovere l'importanza dei principi di un AI responsabile, trasferendo la necessità di un'analisi, profonda, del contesto e dei processi che lo caratterizzano. L'approccio principale adottato nella definizione del framework è stato quello di promuovere un processo di costruzione ed integrazione della XAI attraverso principi di co-design incentrato sull'uomo (human-centered design). Un co-design visto anche come un percorso partecipativo che coinvolge gli utenti finali rendendoli partecipanti attivi del processo di progettazione e facilitando le interazioni umano-macchina affinché siano efficaci ed utili. La sperimentazione e l'applicazione degli strumenti nei diversi casi d'uso individuati ha evidenziato, inoltre, l'importanza operativa e tangibile della XAI, in particolare durante le fasi di costruzione e sviluppo dei sistemi di machine learning. L'utilizzo retroattivo dei risultati della XAI ha, inoltre, delineato un nuovo approccio, promettente, sul come integrare la XAI nei sistemi AI

attraverso uno scenario pro-attivo, real time, che vede i risultati di particolari tecniche di XAI concorrere, direttamente, al miglioramento delle performance degli stessi sistemi AI analizzati. Le direzioni di ricerca tracciate dal presente lavoro, implicano la possibilità di proseguire gli studi di ricerca su due diverse direzioni. Per quanto concerne il framework metodologico, i prossimi studi potrebbero concentrarsi sulla costruzione di strumenti ad-hoc da correlare alle metodologie per aumentare l'operatività del framework stesso. Possiamo pensare ad esempio di progettare e costruire strumenti come matrici decisionali, checklist valutative, strumenti di compliance, moduli di generazione automatica per la produzione di documentazione a supporto, tutti pensati e sviluppati con l'ausilio dell'AI. Per quanto concerne invece REMOAC, si potranno investigare nuove applicazioni del metodo, e più in generale, esplorare, l'utilizzo pro-attivo delle tecniche di XAI per migliorare le performance dei modelli AI con un approccio automatico.