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

ABSTRACT

Osservando la crescente evoluzione del fenomeno tecnologico dell'AI e la sua sempre più frequente e forte integrazione nei processi di descision making, questo lavoro di tesi vuole affrontare le problematiche e gli ostacoli che si presentano quando si vuole introdurre i sistemi AI nei Decision Support System. Per perseguire tale obiettivo, la prima parte della tesi si concentra sullo studio e analisi dei temi di Responsible AI ed Explainable AI, allo scopo di definire un quadro completo che contempli i requisti necessari da soddisfare e costruisca un linguaggio comune sui temi affrontati. Partendo da tale studio, la seconda parte della tesi affronta la costruzione di un framework metodologico per supportare l'introduzione di tecniche e modelli di XAI all'interno dei cicli di sviluppo dei sistemi di AI e dei relativi processi aziendali. Infine, il lavoro proposto, conclude il suo percorso di analisi sui temi affrontati, proponendo tre approcci diversificati delle tecniche di XAI su domini e tipologia di dati differenti. In particolare, con REMOAC, si vuole proporre un nuovo approccio nell'utilizzo dei risultati della XAI, illustrando una possibile modalità retroattiva e automatizzata delle informazioni estratte al fine di migliorare le performance dei sistemi AI analizzati.

INDICE

CAPITOLO 1

INTRODUZIONE

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. Tali processi, 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, sopratutto, 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 saranno affrontati nel dettaglio e sintetizzati nel Capitolo 2, 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, contempla 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 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. Nel paragrafo 3.1, viene descritto il framework metodologico elicitando tutte le sue componenti principali, rappresentate dai fasi, task e possibili deliverable prodotti. La metodologia di ricerca adottata 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 raffinamento nella comprensione della XAI. Nel Capitolo 4 sono stati riportati alcuni casi d'uso che hanno contribuito alla costruzione del framework metodologico. In particolare, lo studio condotto e descritto nel paragrafo 4.4 ha contribuito a costruire una nuova visione nell'utilizzo degli strumenti di XAI: parallelamente alla loro applicazione per capire il funzionamento dei modelli e verificare principi, qualità ed efficienza dei sistemi AI, con conseguente programmazione di task specifici previsti da chi utilizza i sistemi (come ad esempio la correzione dei modelli da parte dei data scientist o machine learning engineer), si è immaginato un utilizzo diretto e automatico, in una forma retroattiva, dei risultati delle tecniche di XAI per migliorare il sistema stesso di AI che era oggetto di spiegazione.

Infine, nel Capitolo 5, si discutono i risultati raggiunti e si delineiamo possibili lavori futuri.

CAPITOLO 2

VERSO L'AI RESPONSABILE

Lo studio e analisi della letteratura è stato costruito e condotto investigando due tematiche fondamentali per l'introduzione dell'AI nei processi aziendali: la Responsible AI e l'Explainable AI. I due temi sono fortemente correlati e una loro analisi approfondita ha rappresentato una solida base per poter sviluppare il framework metodologico e le soluzioni di XAI da integrare nei processi decisionali. Se la Responsible AI riguarda sopratutto, l'etica, la giustizia, la trasparenza e lo studio dell'interazione uomo-macchina nei sistemi di AI, lo studio sull'Explainable AI si concentra, in particolare, sulla comprensione delle tassonomie e concetti che la rappresentano, finalizzando il tutto con sperimentazioni riguardanti gli strumenti attualmente disponibili.

2.1 AI Responsabile

La Human Centered Responsible Artificial Intelligence (HCR-AI) si riferisce allo sviluppo e all'implementazione di sistemi di AI con un focus etico, tenendo conto delle implicazioni sociali, morali e di sicurezza. Questo approccio si concentra su aspetti come l'equità, la privacy, la sicurezza ed etica, assicurando che le tecnologie AI siano sviluppate e utilizzate in modo che favoriscano la prosperità e il benessere delle persone e della società nel suo insieme, piuttosto che semplicemente perseguire progressi tecnologici o guadagni economici. In sostanza, HCR-AI pone l'individuo e i valori etici al centro dello sviluppo dell'AI. La HCR-AI si può caratterizzare su quattro diverse verticali che possono aiutare la comprensione del tema:

- **Interpretabilità e spiegabilità**: rendere le decisioni dell'AI comprensibili per guadagnare la fiducia degli utenti. Per far ciò è necessario progettare sistemi costruiti e incentrati sugli utenti ([1], [2], [3]). Alcuni studi dimostrano che nei sistemi a supporto delle decisioni, se i sistemi AI sono supportati da spiegazioni chiare,comprensibili e utili, migliorano significativamente la qualità delle spiegazioni.[2].

- **Equità**: affrontare i pregiudizi nei sistemi AI per garantire la loro equità. Il concetto di equità non può essere descritto e rappresentato in una modalità unica ([4] [5],[6], [7], [8]). È necessario sempre analizzare il contesto e i desiderata degli stakeholder per determinare e declinare il concetto di equità per renderlo accertabile e comprensibile agli utenti coinvolti.

- **Sicurezza e Privacy**: esplorare il compromesso tra raccolta dati e protezione della sicurezza e della privacy. Entrambi i temi, nell'ambito dei sistemi informativi negli ultimi anni, sono diventati centrali ed estremamente complessa la loro trattazione. Con la diffusione dell'AI abbiamo assistito ad un esasperazione, giusta, del tema, che ha portata negli anni a diversi studi centrali sul tema privacy e sicurezza([9], [8];

[10],[3], [11], [12], [13],[6], [14]). Il tema della privacy, in particolare, in una sua accezione generale, implica che i dati personali devono essere raccolti in modo responsabile e consapevole. Questo significa ottenere un consenso informato, limitare la raccolta di dati al necessario, e utilizzare tecniche come l'anonimizzazione per proteggere l'identità degli individui. In aggiunta, bisogna garantire che i sistemi AI siano sicuri e resistenti contro minacce esterne. Tracciare tutti gli step, le informazioni, il processo in generale rappresenta uno requisito fondamentale nella costruzione di sistemi AI responsabile. Resta però da gestire un forte trade-off tra spiegabilità, privacy e sicurezza. In diversi casi per garantire trasparenza e spiegabilità può essere necessario avere accesso a informazioni coperte da privacy, a volte nemmeno utilizzate nemmeno dai sistemi ai analizzati, ma utili per comprenderne il significato.

- **Etica**: costruire sistemi AI che lavorino nel rispetto dell'etica ([15],[16],[17],[18],[19],[20],[21],[22],[23]), garantendo valori e diritti umani fondamentali. Rispetto alle direttrici appena descritte il tema dell'Ethical AI copre numerosi aspetti e diverse sfumature che indirizzano il tema stesso su verticali fortemente differenti. Anche in questo caso lo sforzo nella progettazione di sistemi HCR compliant risulta nella capacità di verticalizzare e personalizzare il tema nel conteso di utilizzo.

Garantire la HCR-AI significa quindi progettare e sviluppare i propri sistemi di AI con un approccio human-center ([24], [25], [26]). I metodi di design centrati sull'uomo (Human-Centered Design, HCD) per lo sviluppo dei sistemi di AI si concentrano sull'integrazione delle esigenze, delle esperienze e degli obiettivi degli utenti finali nel processo di progettazione. Per perseguire tale obiettivo è necessario comprendere le esigenze, i comportamenti e le preferenze degli utenti attraverso interviste, sondaggi, e osservazioni e prevedere una forte sinergia continuativa durante la progettazione e lo sviluppo dei sistemi. Infatti se da un lato nella fase di progettazione è fondamentale collaborare direttamente con gli utenti per garantire che il prodotto finale soddisfi le loro

esigenze, durante lo sviluppo è estremamente necessario prevedere prototipi ed iterazioni continue basate sul feedback degli utenti allo scopo di migliorare la soluzione. Tutti i sistemi poi devono essere oggetto di piani specifici di test, in particolare, stressando fortemente l' usabilità dei sistemi per valutare come gli utenti interagiscono con il sistema AI e identificare aree di miglioramento.

2.2 Explainable AI per la spiegabilità e interpretabilità dei modelli di AI

Lo studio condotto sull'analisi della letteratura per l'Explainable AI (XAI) evidenzia una crescita significativa del campo negli ultimi anni, dovuta soprattutto all'applicazione diffusa del machine learning, in particolare del deep learning. La XAI è un campo estremamente importante nell'ambito dell'AI che si concentra, di fatto, sulla creazione di sistemi, che, applicati a modelli di AI, sono in grado di fornire spiegazioni comprensibili e trasparenti delle loro decisioni e azioni. Mentre le tradizionali applicazioni di AI tendono ad operare come "scatole nere", in cui i processi decisionali interni sono opachi e difficili da interpretare, la XAI mira a rendere questi processi accessibili e comprensibili sia per gli sviluppatori sia per gli utenti finali. La necessità di integrare la XAI emerge, quindi, dal crescente impiego dell'AI in particolare in settori critici come la sanità, il diritto, la finanza e la sicurezza, dove comprendere il "perché" dietro una decisione AI è cruciale per promuovere l'adozione e l'integrazione nei processi decisionali. La XAI si occupa quindi di sviluppare tecniche e strumenti che permettono agli algoritmi di AI di "spiegare" le loro decisioni in un linguaggio comprensibile, contribuendo alla creazione di sistemi di AI più trasparenti, responsabili e fidati.

Perché XAI?

In [27] gli autori evidenziano la necessità dell'XAI, elencando quattro ragioni principali per cui le spiegazioni sono necessarie: giustificazione, controllo, miglioramento e scoperta. La giustificazione è una ragione chiave per l'XAI, permettendo agli utenti di comprendere il perché di una certa decisione, specialmente quando sono prese decisioni inaspettate. Il controllo è importante perché una maggiore comprensione del comportamento del sistema aiuta a identificare e correggere rapidamente errori. Il miglioramento continuo del sistema AI è un'altra ragione per la quale un modello se è spiegabile e comprensibile, è più facile correggerlo e migliorarlo. Infine, le spiegazioni

possono aiutare a scoprire nuove intuizioni utili a migliorare le decisioni finali. Inoltre se si analizza [28] troveremo delle forti intersezioni tra le linee guida per la progettazione di sistemi AI fornita dall'Europa con le capability proprie dei sistemi XAI.

- Human agency and oversight: i sistemi di AI dovrebbero potenziare gli esseri umani, consentendo loro di prendere decisioni informate e promuovendo i loro diritti fondamentali.

- Robustezza tecnica e sicurezza: i sistemi di AI devono essere resilienti e sicuri, con piani di emergenza in caso di problemi.

- Privacy e governance dei dati: rispetto della privacy e protezione dei dati, con meccanismi adeguati di governance dei dati.

- Trasparenza: trasparenza dei dati, del sistema e dei modelli di business dell'AI, con meccanismi di tracciabilità.

- Diversità, non discriminazione e equità: evitare pregiudizi ingiusti e promuovere la diversità.

- Benessere sociale e ambientale: l'accesso a sistemi di AI dovrebbe essere utile a tutti, comprese le generazioni future, garantendo la sostenibilità nel tempo e tenendo conto dell'impatto sociale e ambientale.

- Responsabilità: istituire meccanismi che garantiscano la responsabilità e l'affidabilità dei sistemi di AI e dei loro risultati. La verificabilità, che consente di valutare gli algoritmi, i dati e i processi di progettazione, svolge un ruolo fondamentale, soprattutto nelle applicazioni critiche. Inoltre, deve essere garantito un ricorso adeguato e accessibile.

Proseguendo lo studio sulle motivazioni poste alla base della necessità della comprensione dei modelli di AI, possiamo razionalizzare quanto analizzato, schematizzando e sintetizzando alcune direzioni comuni emerse nei diversi studi:

- Fiducia e Affidabilità: per fidarsi delle decisioni prese da un modello di machine learning, gli utenti e gli stakeholder devono capire come il modello arriva a tali decisioni. Questo è particolarmente cruciale in settori come la medicina, la finanza e il diritto, dove le decisioni possono avere conseguenze significative.

- Accountability e Conformità Legale: per garantire che i modelli siano conformi alle normative, come il GDPR, è necessario spiegare come i modelli trattano e interpretano i dati. Questo è importante per affrontare problemi di responsabilità legale e per rispondere ai diritti degli individui di comprendere come vengono prese decisioni che li riguardano.

- Rilevamento e Correzione di Bias: i modelli di machine learning possono involontariamente incorporare e perpetuare pregiudizi esistenti nei dati. Capire come funzionano aiuta a identificare e correggere questi bias, garantendo decisioni più eque e meno discriminate.

- Miglioramento e Ottimizzazione del Modello: comprendere il funzionamento dei modelli permette agli sviluppatori di migliorarne la precisione, l'efficienza e la generalizzazione. Questo è fondamentale per adattare il modello a nuovi dati o contesti.

Spiegabilità o interpretabilità

Nella letteratura, non c'è un accordo unanime sul significato di "spiegabilità" e "interpretabilità", spesso usati in modo intercambiabile. Diverse definizioni mettono in luce questa ambiguità [29], [30],[31], [32], [33]. Ad esempio, in [29] si definisce l'interpretabilità come la capacità di spiegare o fornire un significato in termini comprensibili per un essere umano, mentre la spiegabilità è associata all'idea di spiegazione come interfaccia tra umani e sistemi decisionali. D'altra parte [33] differenzia i termini suggerendo che l'interpretazione traduce concetti astratti in intuizioni utili per la conoscenza del dominio, mentre la spiegazione fornisce informazioni su come un modello è giunto a una decisione o interpretazione. Analizzando le diverse fonti

disponibili, è possibile procedere con una sintesi che prova a fornire una distinzione fra i termini fondamentali nel dominio della XAI:

- Understandability: la capacità di un modello di rendere comprensibile il suo funzionamento a un essere umano senza bisogno di spiegare la sua struttura interna. [34]
- Comprehensibility: si riferisce alla capacità di un algoritmo di apprendimento di rappresentare la conoscenza acquisita in modo comprensibile per gli esseri umani [35],[36],[37].
- Interpretability: la capacità di spiegare o fornire significati in termini comprensibili a un essere umano.
- Explainability: associata alla nozione di spiegazione come interfaccia tra esseri umani e decision maker, che è sia un proxy accurato del decision maker sia comprensibile agli esseri umani [38].
- Transparency: un modello è considerato trasparente se di per sé è comprensibile. I modelli trasparenti sono suddivisi in categorie basate sul grado di comprensibilità [39]

2.2.1 Tassonomia della XAI

La tassonomia e la terminologia nella spiegabilità dell'intelligenza artificiale (XAI) svolgono un ruolo cruciale nell'offrire chiarezza e coerenza nel campo in continua evoluzione della XAI. In questa sezione, esploreremo la tassonomia, ovvero la classificazione e l'organizzazione dei concetti e delle categorie rilevanti nella XAI, insieme alla terminologia specifica utilizzata per descrivere i principi, le tecniche e i concetti fondamentali. Una comprensione approfondita di questa tassonomia e terminologia è essenziale per conoscere il contesto di rifermento del presente lavoro, e capire le ragioni per il quale è necessario costruire un framework metodologico per l 'applicazione della XAI.

Prospettive e target audience

Un aspetto fondamentale per lo sviluppo e l'integrazione della XAI è rappresentato sicuramente dall'analisi delle prospettive. Secondo [40] e [41] è possibile individuare cinque diverse prospettive con le quali declinare e caratterizzare un approccio di XAI:

- **Regulatoria**

  Dal punto di vista normativo, l'uso di sistemi di intelligenza artificiale (AI) nelle nostre vite quotidiane pone nuove sfide legislative, specialmente per decisioni con effetti legali. Il Regolamento Generale sulla Protezione dei Dati (GDPR) dell'Unione Europea esemplifica la necessità dell'AI Spiegabile (XAI) da una prospettiva normativa. Il GDPR introduce il "diritto di spiegazione", permettendo agli utenti di richiedere spiegazioni sulle decisioni influenti prese dagli algoritmi. Tuttavia, l'attuazione di tali regolamenti è complessa e senza una tecnologia che fornisca spiegazioni, questo diritto rimane puramente teorico.

- **Scientifica**

  Da una prospettiva scientifica, l'utilizzo di modelli di AI si concentra sullo sviluppo di una funzione approssimativa per risolvere problemi specifici. Dopo la creazione, un modello di AI, diventa una base di conoscenza estesa, superando, di fatto, la conoscenza intrinsecamente espressa dagli stessi dati originali. La XAI può essere utilizzata per rivelare la conoscenza scientifica estratta da questi modelli, contribuendo potenzialmente alla scoperta di nuovi concetti in vari settori scientifici.

- **Industriale**

  Da un punto di vista industriale, le normative e la sfiducia degli utenti nei confronti dei sistemi di AI rappresentano sfide nell'applicazione di sistemi AI complessi e accurati. In alcuni casi, modelli meno accurati ma più interpretabili possono essere preferiti a causa delle normative. Un vantaggio significativo della XAI è che può aiutare a mitigare il

compromesso comune tra interpretabilità del modello e prestazioni, affrontando queste sfide. Tuttavia, l'XAI può aumentare i costi di sviluppo e implementazione.

- **Sviluppo**

  Dal punto di vista dello sviluppo dei modelli, la XAI è cruciale per comprendere e mitigare i risultati inappropriati dei sistemi di AI, causati da fattori come dati di addestramento limitati o distorti, outlier, dati avversari e sovradimensionamento del modello. L'XAI aiuta a comprendere, correggere e migliorare i sistemi di AI per aumentarne la robustezza, la sicurezza e la fiducia degli utenti, minimizzando comportamenti errati, bias, ingiustizie e discriminazioni. Inoltre, l'XAI può facilitare la selezione tra modelli con prestazioni simili rivelando le caratteristiche utilizzate per prendere decisioni e agire come una funzione proxy per obiettivi ottimizzati incompleti.

- **End-user e prospettive sociali**

  Dal punto di vista degli utenti finali e prospettive sociali, risiede una forte preoccupazione sulla correttezza delle decisioni prese dai sistemi di AI sollevando forti dubbi sulla loro affidabilità, in quanto piccole modifiche, impercettibili per un'analisi umana, possono generare risultati completamente inaffidabili e diversi da loro. Inoltre, esiste la preoccupazione che questi sistemi possano produrre decisioni ingiuste se sviluppati con dati contenenti pregiudizi umani. Aumentare la spiegabilità e l'interpretabilità dei sistemi di AI può incrementare la fiducia degli utenti, permettendo loro di comprendere il ragionamento dietro le decisioni del modello e assicurando che il sistema serva al suo scopo progettato anziché semplicemente a quello per cui è stato addestrato.

Come si nota dalla descrizione delle prospettive, ogni applicazione possibile della XAI è fortemente correlata alla tipologia di stakeholder che

la rappresenta, il proprio target. Si dimostra di fondamentale importanza riuscire a caratterizzare le tipologie di attori presenti nel processo , al fine di indirizzare e costruire al meglio i desiderata e gli obiettivi della XAI.

Contesto di riferimento

Nell'ambito della XAI, le spiegazioni possono essere classificate, in prima istanza, in due categorie principali: local explanation e global explanation. Queste due tipologie di spiegazione offrono differenti livelli di comprensione su come un modello di machine learning prende decisioni. Una local explanation fornisce spiegazioni su singole decisioni o previsioni di un modello. Si concentra su una specifica istanza o input per capire perché il modello ha prodotto un certo output per quel caso particolare. È utile quando gli utenti o gli stakeholder necessitano di comprendere il processo decisionale del modello in un caso specifico. Ad esempio, capire perché un modello ha rifiutato una domanda di prestito o ha diagnosticato una particolare condizione medica. Una global explanation, invece, offre una visione d'insieme del comportamento del modello su tutti i dati o un vasto set di istanze. Mostra come il modello prende decisioni in generale, piuttosto che in casi specifici. È importante per capire il comportamento complessivo del modello, le sue regole decisionali generali e per identificare eventuali problemi sistemici come bias o incoerenze. In sintesi, mentre le local explanations sono focalizzate sulle decisioni a livello di singole istanze, fornendo intuizioni dettagliate e specifiche, le global explanations offrono una visione d'insieme, aiutando a capire le regole e i patterns generali che guidano il modello.

Secondo [42] è possibile caratterizza le due diverse tipologie di explanation secondo caratteristiche differenti.
Se consideriamo una prospettiva globale troviamo:

- **Completezza**: precisione nella descrizione del sistema fino alla capacità di anticipare i risultati del modello.

- **Potere Espressivo**: riguarda il linguaggio usato per le spiegazioni, che possono variare in forma e complessità.
- **Traslucenza**: misura quanto la spiegazione si basa sull'indagine degli interni del modello ML.
- **Portabilità**: valuta l'ampiezza dei modelli di ML coperti dal metodo XAI.
- **Complessità Algoritmica**: relativa alla complessità computazionale dei metodi di generazione delle spiegazioni.

Le spiegazioni individuali invece sono valutate in termini di:

- **Accuratezza**: si riferisce a quanto bene una spiegazione può prevedere dati non visti, parallela alla metrica di accuratezza in ML che conta il numero di predizioni corrette rispetto al totale dei campioni. Questo aspetto è importante nella XAI poiché le spiegazioni devono riflettere fedelmente il livello di accuratezza del modello ML a cui si riferiscono.
- **Consistenza**: si riferisce alla somiglianza tra le spiegazioni fornite da diversi modelli formati sullo stesso caso d'uso.
- **Stabilità**: valuta come variazioni minori nelle caratteristiche influenzano le spiegazioni in un modello specifico, essendo importante per la fiducia nel metodo di spiegazione.
- **Comprensibilità**: riguarda la capacità degli utenti di comprendere le spiegazioni generate, fondamentale nel contesto dell'interazione uomo-macchina.

Trasparenza dei modelli di AI

Nella letteratura, è presente una distinzione tra modelli intrinsecamente interpretabili e quelli che possono essere spiegati tramite tecniche esterne di XAI[38]. Questa dualità ha generato un tipo di classificazione che vede i

modelli di machine learning divisi secondo due categorie: modelli trasparenti e spiegabilità post-hoc. Per quanto concerne la prima categoria di modelli, nel contesto della trasparenza dell'AI, si considerano tre livelli: *trasparenza algoritmica, decomponibilità e simulabilità.* [33] La **simulabilità** indica la capacità di un modello di essere simulato o pensato in modo rigoroso da un essere umano, la **decomponibilità** indica la capacità di spiegare ciascuna delle parti di un modello (input, parametri e calcoli e può essere considerata anche come intelligibilità) [43]. L'ultimo livello di trasparenza, infine, è quella **algoritmica** e riguarda la capacità dell'utente di comprendere il processo seguito dal modello per produrre un determinato output sulla base di specifici dati di input. Per esempio, un modello lineare è considerato trasparente perché la sua superficie di errore può essere compresa e analizzata, permettendo all'utente di capire come il modello si comporterà in ogni situazione. Al contrario, questo non è possibile in architetture deep learning, dove il paesaggio delle perdite può essere opaco e la soluzione deve essere approssimata tramite ottimizzazione euristica. La principale limitazione per i modelli algoritmicamente trasparenti è che devono essere completamente esplorabili tramite analisi e metodi matematici. Tra i i modelli trasparenti troviamo: Linear/Logistic Regression, Decision Trees, K-Nearest Neighbors, Rule Based Learners, General Additive Models, BayesianModels.
Quando i modelli di machine learning non soddisfano i criteri di trasparenza, è necessario sviluppare un metodo separato per spiegare le loro decisioni. Questo è l'obiettivo delle tecniche di spiegabilità post-hoc, che mirano a comunicare informazioni comprensibili su come un modello già sviluppato produca le sue previsioni per un dato input. Queste tecniche possono essere categorizzate in due gruppi: 1) quelle progettate per l'applicazione a modelli ML di qualsiasi tipo (model agnostic) e 2) quelle specifiche per un particolare modello ML e quindi non direttamente trasferibili ad altri modelli.

Le tecniche model-agnostic per la spiegabilità post-hoc sono progettate per essere applicate a qualsiasi modello al fine di estrarre informazioni dal suo processo di predizione. Queste tecniche possono includere la semplificazione

per generare proxy che imitano i modelli originali, l'estrazione diretta di conoscenza dai modelli o la visualizzazione per facilitare l'interpretazione del loro comportamento. Seguendo la tassonomia introdotta in precedenza, le tecniche model-agnostic possono basarsi su:

- **Model simplification**: implica la creazione di modelli più semplici e comprensibili che approssimano il comportamento di un modello di machine learning più complesso. Questo approccio si basa sull'idea che un modello semplificato può essere più facilmente interpretato e compreso dagli umani. Questi modelli semplificati, che sono rappresentazioni approssimative del modello originale, vengono utilizzati per fornire spiegazioni sulle decisioni o previsioni del modello più complesso, mantenendo un equilibrio tra accuratezza e comprensibilità. Approcci come LIME [44] e le sue variazioni ([45],[46]), si concentrano sulla costruzione di modelli lineari costruiti localmente intorno le predizioni di un modello opaco. Queste tecniche, che includono anche l'estrazione di regole e la semplificazione del modello, mirano a rendere i modelli complessi interpretabili. Anche G-REX [47], sebbene non sia stato originariamente concepito per l'estrazione di regole da modelli opachi, la proposizione generica di G-REX è stata estesa per tener conto anche della spiegabilità dei modelli [48][49] Altri studi presenti in letteratura [50] propongono approcci differenti per l'apprendimento di regole in forme CNF (Conjunctive Normal Form) o DNF (Disjunctive Normal Form).

- **Feature relevance estimation**: le tecniche di spiegabilità basate sulla rilevanza delle caratteristiche (features) si concentrano sulla valutazione dell'importanza o rilevanza delle singole caratteristiche (features) nel modello di machine learning. Questi approcci analizzano quanto ogni caratteristica contribuisca alle previsioni del modello. L'obiettivo è di identificare quali caratteristiche sono più influenti nelle decisioni del modello. Questa comprensione aiuta a spiegare perché il modello ha fatto una certa previsione, offrendo intuizioni sulle relazioni tra

le caratteristiche e l'output del modello. SHAP [51] rappresenta un contributo significativo in questo campo e, basato sulla teoria dei giochi cooperativi, attribuisce a ogni caratteristica (feature) di un input il suo impatto sull'output del modello. Per ogni previsione, SHAP calcola il valore di ogni caratteristica, rappresentando quanto quella specifica caratteristica abbia contribuito al risultato, sia positivamente che negativamente. I risultati di SHAP sono spesso rappresentati visivamente (ad esempio, tramite grafici), permettendo agli utenti di identificare facilmente quali caratteristiche sono le più influenti per una data previsione.

- **Tecniche di visualizzazione**: si basano sull'utilizzo di rappresentazioni grafiche per spiegare le decisioni o i comportamenti di un modello di machine learning. Queste tecniche trasformano le informazioni complesse e astratte dei modelli in rappresentazioni visive più facilmente interpretabili. Le visualizzazioni possono includere grafici, mappe di calore o diagrammi che illustrano come le diverse caratteristiche influenzano le previsioni del modello, fornendo così un modo intuitivo per gli utenti per comprendere il funzionamento del modello. Lavori rappresentativi in quest'area si trovano in [52]che presenta un portfolio esteso di tecniche di visualizzazione per aiutare la spiegazione di un modello ML black-box, basandosi anche sull'integrazione di tecniche differenti. In aggiunta, [52] introduce nuovi metodi di SA (Sensitivity Analysis), basati sui dati, approccio Monte-Carlo e sui cluster, ed un'innovativa importance measure, la Average Absolute Deviation. Proseguendo lo studio, [53] introduce i grafici ICE (Individual Conditional Expectation) come strumento per visualizzare il modello stimato da qualsiasi algoritmo di apprendimento supervisionato. Le tecniche di spiegazione visiva nel contesto della spiegabilità post-hoc model-agnostic sono relativamente meno comuni rispetto ad altri metodi. Le ragioni di questa diffusione limitata risiedono sul fatto che creare visualizzazioni efficaci basandosi solo sugli input e output di un modello

opaco, senza accedere alla sua struttura interna, rappresenta una sfida di complessità elevata. Tali tecniche di visualizzazione sono spesso integrate con metodi che determinano la feature importance, fornendo le informazioni necessarie per creare rappresentazioni visive utili per l'utente finale.

Criteri di valutazione per la XAI

Come nella maggior parte dei sistemi IT, anche per la XAI possiamo identificare due grandi tipologie di criteri di valutazione, quelli oggettivi e quelli soggettivi. Un interessante lavoro [54] conduce un'analisi della letteratura per quanto concerne i criteri di valutazione della XAI, raggruppandoli e schematizzandoli attraverso tre misure distinte: obiettivo comune (esempio valutazione della performance di un modello) tipologia, se erano oggettivi o soggettivi, oggetto della loro misurazione (modello, XAI, utente). Al fine di comprendere al meglio l'analisi condotta dagli autori, si riporta in una versione estesa, contente anche le definizione di tutti i criteri analizzati, la tabella di sintesi dei criteri di valutazione della XAI:

Criteri	**Descrizione**	**Aspetto**
Performance	**La performance del modello**	**Modello**
Accuracy	Misura le prestazioni di un modello in termini di correttezza delle predizioni rispetto ai dati reali.	
Accountability	Riguarda la responsabilità e la tracciabilità delle decisioni prese dal modello.	
Confidentiality	Si riferisce alla protezione dei dati utilizzati dal modello	
Certainty	Indica il grado di sicurezza o affidabilità delle predizioni del modello.	

Confidence	i riferisce alla probabilità associata alle predizioni di un modello.	
Fairness *Bias (detected)*	**La capacità di un modello di trattare equamente tutti i gruppi e individui, evitando pregiudizi e discriminazioni**. Livelli di bias presente nel modello.	Modello
Privacy	**Quanto le informazioni sensibili sono protette nel modello**	Modello
Reliability *Robustness*	**La coerenza delle prestazioni di un modello su diversi set di dati e in diverse condizioni operative.** Indica la capacità di un modello di mantenere elevate prestazioni di fronte a variazioni o perturbazioni nei dati di input, come rumore, dati mancanti, o esempi non rappresentativi.	Modello
Identity	**listanze identiche devono avere spiegazioni identiche. Questo perché due istanze uguali dovrebbero avere predizioni uguali e, di conseguenza, valori di importanza delle caratteristiche uguali.**	XAI

Stability	Rappresenta quanto sono simili le spiegazioni per istanze simili. Alta stabilità indica che variazioni minime nelle caratteristiche di un'istanza non cambiano sostanzialmente la spiegazione, a meno che queste variazioni non influenzino anche fortemente la previsione.	
Separability	**Oggetti non identici non possono avere spiegazioni identiche. Anche se un modello prevede lo stesso risultato per diverse istanze, la spiegazione dovrebbe differire a causa dei diversi valori delle caratteristiche che hanno generato la previsione.**	XAI
Consistency	Si riferisce a quanto siano simili le spiegazioni fornite da modelli diversi addestrati sullo stesso compito. Se due modelli producono previsioni simili ma utilizzano relazioni o caratteristiche diverse, le loro spiegazioni dovrebbero essere diverse per riflettere questi aspetti distinti.	

Novelty	**Descrive se una spiegazione riflette il fatto che un'istanza, la cui previsione deve essere spiegata, provenga da una regione dello spazio delle caratteristiche lontana dalla distribuzione dei dati di addestramento. In tali casi, il modello potrebbe essere impreciso e la spiegazione inutile**	XAI
Representativeness	**Descrive quante istanze sono coperte dalla spiegazione. Le spiegazioni possono essere globali, coprendo l'intero modello (come l'interpretazione dei pesi in un modello di regressione lineare), oppure possono riguardare solo una singola previsione.**	XAI
Fidelity	**La coerenza delle prestazioni di un modello su diversi set di dati e in diverse condizioni operative.**	XAI

Description accuracy	Si riferisce alla precisione predittiva della spiegazione rispetto a dati non visti. In alcune situazioni una bassa accuratezza della spiegazione può essere accettabile se anche l'accuratezza del modello di machine learning è bassa. Ciò implica che l'importanza della spiegazione è valutata in relazione alla fedeltà con cui rappresenta le prestazioni effettive del modello.	
Feature importance/ relevance	La "Feature Importance" si riferisce a quanto una caratteristica (feature) specifica contribuisca alla performance di un modello di machine learning. Indica l'importanza relativa o l'impatto di ciascuna caratteristica nelle previsioni del modello. La "Feature Relevance", invece, riguarda la pertinenza o l'applicabilità di una caratteristica specifica al problema o al contesto di un modello. Questo termine si concentra su quanto una caratteristica sia significativa o utile per comprendere i risultati del modello in un determinato contesto o per un particolare compito di previsione	

BAM	Benchmarking Attribution Methods, un framework per valutare la correttezza delle attribuzioni delle features e la loro importanza relativa.	
Faithfulness	Valutare la correlazione tra l'importanza assegnata alle caratteristiche di un modello e l'effetto di ciascuna caratteristica sulle prestazioni del modello. Si misura rimuovendo progressivamente le caratteristiche importanti e osservando l'impatto sulle prestazioni.	
Monotonicity	Misura l'importanza o l'effetto di singole caratteristiche aggiungendole progressivamente in ordine di importanza per osservare le variazioni nelle prestazioni del modello. Le prestazioni del modello dovrebbero aumentare con l'aggiunta di caratteristiche più importanti.	
Appropriate Trust	**Quanto un utente è in grado di distinguere tra previsioni corrette ed errate e di agire in base ad esse.**	User

Calibrated Trust	Si riferisce alla corrispondenza tra il livello di fiducia che gli utenti ripongono nel sistema di AI e il livello effettivo di affidabilità e accuratezza del sistema stesso. Questo concetto enfatizza l'importanza di avere una fiducia ben bilanciata: gli utenti dovrebbero avere una fiducia proporzionata alle capacità reali del sistema, evitando sia una fiducia eccessiva che una insufficiente.	
Trust	Il grado in cui l'utente concorda con il modello.	
Reliance	Si riferisce al grado in cui gli utenti si affidano o si fidano dei risultati e delle spiegazioni fornite da un sistema di AI. Comprende la fiducia dell'utente nella capacità del sistema di eseguire le sue funzioni in modo accurato e affidabile, e quanto gli utenti sono disposti a basare le loro decisioni sulle raccomandazioni o conclusioni del sistema. In sostanza, riguarda l'equilibrio tra fiducia nel sistema e valutazione critica dei suoi output.	

Explanation Satisfaction	**La sensazione degli utenti di comprendere il sistema, le spiegazioni e l'interfaccia utente. Questo grado di comprensione incide sulla fiducia degli utenti nel sistema e sulla loro capacità di utilizzarlo efficacemente.**	User
Satisfaction Scale	Definizione di livelli misurabili per la soddisfazione degli utenti	
Comprehensibility	si riferisce alla facilità con cui gli utenti possono comprendere le spiegazioni fornite dal sistema. Questo include la chiarezza con cui le decisioni o i processi del modello di AI vengono comunicati agli utenti, assicurando che le spiegazioni siano presentate in un modo che possa essere facilmente inteso e interpretato da persone senza competenze tecniche specifiche	
Explanation Goodness	Costruzione di una checklist di domande per valutare la bontà delle spiegazioni	

Causability	Si riferisce alla misura in cui una spiegazione aiuta un utente a raggiungere una comprensione causale di una dichiarazione, con efficacia, efficienza e soddisfazione in un determinato contesto di utilizzo. Questo concetto si concentra sull'abilità di un sistema di spiegazione a fornire comprensione causale in modo efficace e soddisfacente per l'utente.	
System Causability Scale	Definizione di livelli misurabili per la valutazione della Causability	

Tabella 2.1: Criteri di valutazione della XAI

2.3 Considerazioni

L'obiettivo di questo capitolo è stato quello di fornire un'overview del contesto della XAI, corredato dalla costruzione di un linguaggio comune, allo scopo di far emergere un livello di complessità elevato dell'argomento e quindi della sua adozione, in particolare in contesti Enterprise. Gli studi che hanno sotteso la sintesi appena descritta, hanno evidenziato un quadro complessivo estremamente complesso. La sfida principale del framework proposto in 3.1, XAI4DSS, è quella di costruire un valido supporto per favorire l'adozione dell'AI all'interno dei processi decisionali, che si scontra con problematiche legate non solo alle performance dei modelli e tecnologie dell'AI, ma negli anni ha acquisito una forte sensibilità riguardo gli impatti che tali sistemi possono implicare nei processi decisionali. Impatti che possono generare conseguenze sia ad un livello estremamente operativo, come quello di avere decisioni non corrette, identificare la responsabilità sulle decisioni prese, ridurre il controllo umano, sia ad un livello molto più profondo e con conseguenze nel lungo

periodo, come in riferimento a problematiche di sicurezza, privacy, etica che possono generare complicazioni estremamente dannose. In sintesi, la sfida che vuole affrontare XAI4DSS si divide in una duplice sostanza, ovvero proporre una metodologia, delle linee guide in grado sia di applicare le migliori soluzioni tecniche nel miglior modo possibile, sia quella di affrontare e superare, nel contesto limitato dei processi decisionali aziendali, le sfide e le problematiche derivanti da un uso non responsabile dell'intelligenza artificiale.

CAPITOLO 3

XAI4DSS: COSTRUZIONE DI UN FRAMEWORK METODOLOGICO PER LA XAI

3.1 Introduzione al framework metodologico

La necessità di prevedere una metodologia efficace per l'introduzione di sistemi di Explainable AI nasce dall'obiettivo di creare sistemi AI che siano non solo avanzati dal punto di vista tecnologico, ma anche sicuri, affidabili e in linea con i valori umani e sociali. L'applicazione del framework metodologico proposto, va considerata in un contesto più ampio e complesso, come può essere un contesto Enterprise, dove va prevista una forte integrazione con due diverse tipologie di metodologie complementari: business process methodology e machine learning methodology. Se da un lato è opportuno inserire all'interno di workflow di processo gli step della XAI methodology, dall'altro è necessario integrare la presente metodologia all'interno di una metodologia di sviluppo più ampia per i sistemi Machine Learning che nasce proprio per garantire risultati affidabili, ridurre i rischi, ottimizzare l'uso delle risorse e, proprio

grazie all'integrazione della XAI, promuovere una pratica etica e trasparente. In particolare, una metodologia per lo sviluppo di sistemi Machine Learning, per garantire il successo di iniziative ad alta incertezza ed alto carattere innovativo, devo porsi l'obiettivo di:

1. **Garantire producibilità dei risultati**: una metodologia ben definita aiuta a garantire la riproducibilità dei risultati. Questo aspetto risulta cruciale perché in ogni momento chi interviene sui sistemi devono essere in grado di comprendere, replicare e verificare i risultati prodotti anche se sviluppati da altri.

2. **Gestire efficacemente il ciclo di vita dei modelli**: lo sviluppo di modelli di Machine Learning coinvolge diverse fasi, dalla raccolta dei dati alla valutazione del modello. Una metodologia fornisce una guida per ogni fase, contribuendo a gestire il ciclo di vita del modello in modo efficace.

3. **Ridurre il rischio**: aiutare a identificare e gestire i rischi associati a un progetto di Machine Learning. Ciò può includere rischi legati ai dati, alle prestazioni del modello, alla scalabilità e ad altri fattori critici.

4. **Ottimizzare le risorse**: supporto per ottimizzare l'uso delle risorse, tra cui tempo, dati e potenza di calcolo. Ciò è particolarmente importante dato che lo sviluppo di modelli di Machine Learning può essere intensivo in termini di risorse.

5. **Facilitare la collaborazione**: fornire una struttura comune che facilita la collaborazione tra membri del team, in particolare quando diverse persone lavorano su diverse parti di un progetto di Machine Learning.

6. **Valutare continuamente e migliorare le performance**: incorporare cicli di feedback e valutazione continua dei modelli. Questo permette di identificare aree di miglioramento e di apportare modifiche iterativamente per ottenere modelli più preformanti nel tempo.

7. **Comprendere aspetti etici e legali**: uno sviluppo etico dei modelli di Machine Learning è cruciale. Una metodologia di sviluppo può aiutare

a integrare considerazioni etiche e legali fin dall'inizio, evitando problemi futuri.

8. **Documentare efficacemente**: incoraggiare la documentazione chiara e completa di tutte le fasi del processo di sviluppo. Questo è essenziale per la comprensione del lavoro svolto e per la trasmissione del know-how.

Adottando e perseguendo alcuni di questi obiettivi, e allo stesso tempo promuovendo altri che sono altrettanto essenziali, un XAI methodology framework specializza la ML methodology e si concentra sull'interpretazione e sulla comprensione dei modelli di intelligenza artificiale in modo che siano trasparenti e comprensibili agli esseri umani. Infatti, gli obiettivi di una metodologia per la XAI si concentrano su assicurare che i sistemi di AI siano trasparenti, comprensibili e responsabili. Secondo una sintesi semplificata, gli obiettivi di uno XAI methodology framework si possono dividere in:

- **Trasparenza**: rendere chiaro il funzionamento interno dei modelli AI e consentire agli utenti di comprendere come il sistema prende le sue decisioni, quali dati utilizza e quali logiche segue.
- **Comprensibilità**: assicurare che le spiegazioni fornite siano facilmente comprensibili per il pubblico target, rendendo le decisioni AI accessibili non solo agli esperti tecnici, ma anche agli utenti non tecnici, migliorando la fiducia e l'accettazione del sistema.
- **Responsabilità**: fornire spiegazioni che permettano di attribuire la responsabilità delle decisioni prese dal sistema Ai allo scopo di facilitare la responsabilizzazione in caso di errori o risultati inaspettati e supportare la conformità legale ed etica.
- **Fiducia**: costruire fiducia tra gli utenti e i sistemi AI, garantendo che gli utenti si sentano sicuri nell'utilizzo dei sistemi AI, comprendendo i loro limiti e capacità.

- **Personalizzazione**: adattare le spiegazioni alle esigenze specifiche dei diversi utenti, fornendo informazioni pertinenti e utili per vari gruppi di stakeholder, rispettando il loro livello di conoscenza e interesse.

- **Conformità Etica e Legale**: garantire che le spiegazioni rispettino le normative vigenti e i principi etici, assicurando che il sistema rispetti la privacy, la non discriminazione e altre normative importanti (come ad esempio il GDPR).

- **Miglioramento dei Processi Decisionali**: migliorare la qualità delle decisioni prese con il supporto dell'AI, utilizzando le spiegazioni per perfezionare i modelli, ridurre il bias e aumentare l'accuratezza delle decisioni.

Per perseguire tali obiettivi è necessario che il framework contempli, in prima istanza alcune di queste caratteristiche:

- **Comprendere il contesto e i requisiti**:
  - **Scopo dell'interpretazione:** identificare il motivo per cui l'interpretazione è necessaria. Ad esempio, potrebbe essere per la spiegazione dei risultati, l'aderenza alle normative, o la fiducia degli utenti.
  - **Requisiti degli stakeholder**: coinvolgere gli utenti finali, gli esperti del dominio e gli altri stakeholder per capire quali informazioni sono importanti per loro.

- **Selezione del modello e dei metodi di spiegazione**
  - **Scelta del modello**: utilizzare modelli di machine learning che sono intrinsecamente interpretabili o utilizzare tecniche di spiegazione per modelli complessi.
  - **Metodi di spiegazione**: selezionare le tecniche di spiegazione più adatte al contesto, come feature importance, surrogate models, tecniche di visualizzazione, o metodi basati su regole.

- **Preparazione e pre-processamento dei dati**
  - **Trasparenza dei dati**: garantire la qualità dei dati, documentarne l'origine, l'elaborazione e l'eventuale preprocessing per migliorarne la trasparenza.
- **Generazione di spiegazioni**
  - **Interpretazione post-hoc**: applicare tecniche di spiegazione dopo che il modello è stato addestrato per comprendere come prende decisioni. Questo può coinvolgere l'analisi delle feature, la generazione di spiegazioni locali o globali.
  - **Visualizzazione**: creare visualizzazioni chiare ed efficaci per comunicare le spiegazioni ai diversi utenti.
- **Valutazione e miglioramento**
  - **Valutazione delle spiegazioni**: misurare l'efficacia delle spiegazioni rispetto ai requisiti degli utenti.
- **Documentazione e comunicazione**
  - **Documentazione**: registrare e documentare le spiegazioni e le decisioni prese durante il processo di spiegazione e predisporre gli utenti finali alla comprensione dei risultati ottenuti.

Partendo da questi macro-obiettivi e macro attività a supporto, è stato sviluppato XAI4DSS, un framework metodologico costituito di 6 fasi distinte cicliche (come mostrato in Figura 3.1 che ispirandosi ad un flusso agile che prevede ricicli e miglioramento continuo) comprende:

1. **Context Understanding**: analizzare il contesto ed individuare le priorità tra le differenti possibili spiegazioni considerando prospettive diverse e fattori diversi.

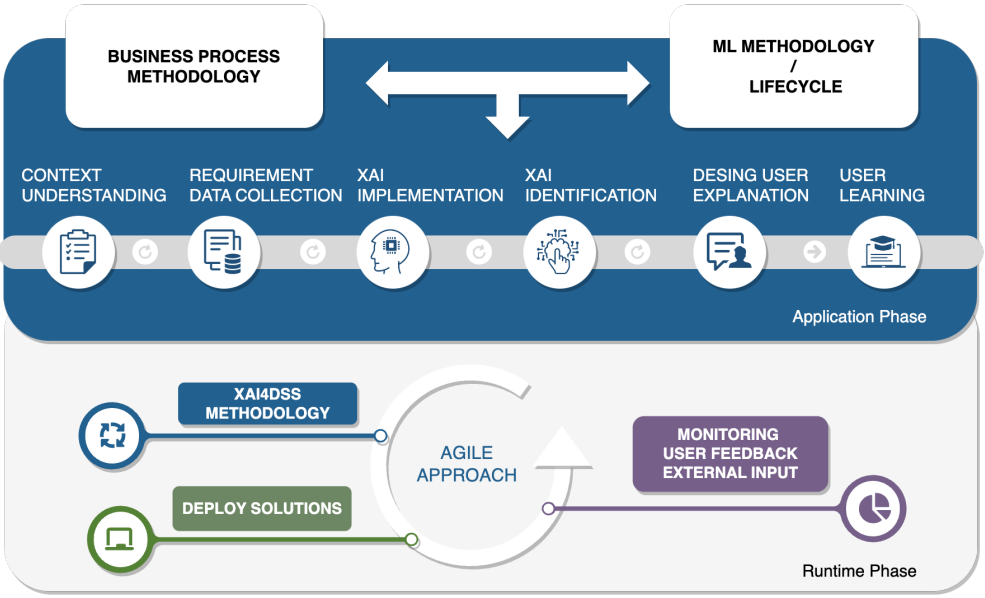


Figura 3.1: XAI4DSS: framework metodologico per la XAI

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.

Per ogni fase distinta sono stati individuati task e deliverable associati.

3.2 Descrizione del framework: fasi, task, deliverable

Partendo dalle sei fasi individuate nel paragrafo precedente, si procede ora con la descrizione dettagliata del framework metodologico identificando task e deliverable associati. L'applicazione del framework va contemplata sia in una modalità integrativa, con metodologie di machine learning e metodologie di process management, sia in una modalità continous agile, cosi da poter supportare la natura dinamica e rapidamente evolutiva di applicazione "data e AI based". L'ambito dell'AI e del XAI è in continua evoluzione, con nuovi modelli, approcci, tecniche e strumenti che emergono regolarmente. Un approccio agile permette, quindi, di adattarsi rapidamente a questi cambiamenti, integrando nuove tecnologie o aggiornando le strategie esistenti. Inoltre l'AI può presentare incertezze, sia nei dati sia nei modelli, e solo un approccio agile consente di identificare e gestire questi rischi in modo più efficace, adattando le strategie di spiegazione in base alle nuove informazioni. Di fondamentale importanza, inoltre, è la capacità del framework di assorbire dunque cambiamenti, aspetti cruciali per migliorare la qualità e l'efficacia delle spiegazioni, anche nel caso in cui essi si presentino sotto forma di risultati di monitoraggio o feedback diretti degli utenti, facilitando di fatto l'iterazione rapida e consentendo miglioramenti continui. Infine, un approccio agile, incoraggia l'innovazione e l'esplorazione di nuove idee nel campo del XAI, permettendo di sperimentare e implementare soluzioni creative. Per rendere dunque sia l'applicazione che l'esecuzione del framework metodologico efficace e resiliente ai cambiamenti, come si evince dalla Figura 3.1, XAI4DSS prevede due macro fasi:

- **Application Phase**: in questa macrofase, si applicano tutte le fasi della metodologia per riuscire a costruire sistemi XAI efficaci.
- **Runtime Phase**: dopo aver costruito i sistemi di XAI, output della metodologia, si procede con il deploy completo delle soluzioni

integrandole nei processi e si applica un monitoraggio costante, secondo le modalità individuate sempre dalla metodologia, fatto di misurazioni dei KPI (Key Perofomance Indicator), raccolta periodica dei feedback degli utenti, continuo monitoraggio di input esterni che potrebbero far ripartire la metodologia (nuovi o modifiche ai casi d'uso, nuove tecnologie disponibili, in generale nuovi requisiti).

In conclusione, l'agilità che caratterizza il XAI4DSS è fondamentale per garantire che i sistemi siano al passo con i rapidi sviluppi tecnologici, siano personalizzabili e reattivi ai bisogni degli utenti, e siano in grado di gestire efficacemente incertezze e rischi. Per quanto concerne la macro fase di Application, nella Tabella 3.1 sono sintetizzate le fasi e i task previsti per ognuna di essa.

FASI	**TASK**
1.CONTEXT UNDERSTANDING	1.1 Analisi del Contesto 1.2 Definizione degli Obiettivi 1.3 Valutazione delle Prospettive Diverse 1.4 Considerazione dei Fattori Multipli 1.5 Selezione delle Spiegazioni Prioritarie
2.REQUIREMENTS FOR DATA COLLECTION	2.1 Identificazione di Dati Rappresentativi 2.2 Valutazione e definizione dei criteri di Qualità dei Dati 2.3 Definizione e valutazione dei rischi nell'utilizzo dei dati 2.4 Modalità creazione e proprietà caratterizzanti dei dati sintetici
3.XAI IMPLEMENTATION	3.1 Integrazione strategie di implementazione dei modelli AI 3.2 Implementazione dei modelli di explainability

4.XAI IDENTIFICATION	4.1 Applicazione dei sistemi XAI ai diversi casi d'uso 4.2 Integrazione nei processi decisionali
5.DESIGN AND DEVELOPMENT USER EXPLANATION	5.1 Desing User Explanation 5.2 Sviluppo, Testing, Integrazione
6.USER LEARNING	6.1 Formazione sui temi di AI e le decisioni autonome 6.2 Formazione sugli strumenti end user

Tabella 3.1: XAI4DSS:Fasi e task

3.2.1 Fase 1: Context Understanding

La fase 1 è fondamentale per comprendere il contesto in cui una o più soluzione AI based deve operare. In particolare, si rende necessario e propedeutico, determinare quali aspetti dei processi interessati sia di sviluppo che decisionali, richiedono spiegazioni chiare e dettagliate sui comportamenti dei sistemi in esame. La fase di individuazione del contesto è suddivisa in varie sotto fasi, costruite considerando: priorità tra le diverse spiegazioni possibili, prospettive diverse e fattori multipli.

Task

1.1 **Analisi generale del contesto**

Identificare l'ambiente in cui il sistema AI è utilizzato, comprendendo il settore, i processi aziendali e le esigenze specifiche. Identificare gli attori chiave, chi sono gli utenti o le parti interessate che richiedono spiegazioni, ovvero gli stakeholder principali, inclusi utenti finali, esperti di dominio, regolatori e altri gruppi influenti. Per ognuno di esso comprendere le loro esigenze, aspettative e livelli di competenza tecnica. Nel comprendere il contesto operativo in cui il modello di AI è implementato è necessario anche valutare i vincoli operativi, come risorse disponibili, tempo, e condizioni ambientali.

1.2 **Definizione degli obiettivi**

Identificare gli obiettivi specifici della spiegazione: descrivere cosa si sta cercando di ottenere attraverso le spiegazioni, come ad esempio, migliorare la fiducia degli utenti, individuare eventuali bias nel modello o rispettare requisiti normativi.

1.3 **Valutazione delle prospettive**

Considerare le diverse prospettive degli stakeholder: diversi attori possono avere esigenze diverse in termini di spiegazioni. Ad esempio, un utente finale potrebbe essere interessato a spiegazioni più intuitive, mentre uno sviluppatore potrebbe richiedere dettagli più tecnici.

Identificare, inoltre, le variabili critiche: determinare quali variabili o feature del modello sono particolarmente rilevanti per l'utente in questione. Questo può aiutare a focalizzare le spiegazioni sulle informazioni più rilevanti. Vediamo alcuni esempi di prospettive caratterizzati e descritti considerando obiettivi e approcci da utilizzare.

Prospettiva	**Obiettivo**	**Approccio**
Regolamenti	Garantire che il modello rispetti le leggi e le normative vigenti.	Concentrarsi sulla trasparenza del processo decisionale, spiegare come il modello adempie ai requisiti normativi e fornire spiegazioni comprensibili alle autorità di regolamentazione.
End-User	Fornire spiegazioni comprensibili e utili agli utenti finali per aumentare la fiducia e facilitare l'accettazione del sistema.	Utilizzare linguaggio intuitivo, grafici comprensibili e fornire spiegazioni che siano rilevanti per le decisioni che gli utenti devono prendere. Assicurarsi che le spiegazioni siano in linea con le aspettative degli utenti.

Impatto Sociale ed Etico	Affrontare le implicazioni etiche e sociali del modello, evitando bias e assicurando che le decisioni del modello siano giustificabili, eque e trasparenti dal punto di vista sociale.	Identificare e mitigare i bias nei dati e nei modelli, garantendo che le spiegazioni siano trasparenti e comprensibili per il pubblico generale. Coinvolgere le diverse tipologie di utenti nell'elaborazione delle linee guida etiche.
Industriale	Massimizzare l'efficienza e l'efficacia dell'implementazione del modello nel contesto industriale.	Fornire spiegazioni che siano rilevanti per gli obiettivi operativi e commerciali, garantendo che il modello sia integrato in modo efficiente nei processi aziendali. Ottimizzare la spiegazione in modo che supporti le esigenze decisionali dell'industria.
Scientifica	Consentire la comprensione approfondita del funzionamento interno del modello da parte degli esperti scientifici.	Fornire spiegazioni tecniche dettagliate, comprese le caratteristiche principali e le interazioni all'interno del modello. Assicurarsi che gli scienziati possano valutare la validità del modello e contribuire alla sua evoluzione.

Tabella 3.2: Esempi di prospettive

Ciascuna di queste prospettive svolge un ruolo importante nel garantire il successo e l'accettazione di un sistema di AI spiegabile. Integrare queste prospettive nella fase di valutazione delle spiegazioni contribuirà a creare un sistema che sia trasparente, etico, socialmente responsabile e utile per una vasta gamma di stakeholder. Le prospettive riportate in tabella sono da considerarsi integrabili e adattabili al contesto in cui viene istanziata la metodologia.

1.4 Considerazione dei Fattori Multipli

La considerazione dei fattori multipli è fondamentale per garantire che le spiegazioni siano adattate alle esigenze specifiche del sistema e degli utenti coinvolti. Alcune variabili potrebbero essere più critiche di altre nella spiegazione del processo decisionale. Assegnare pesi alle variabili in base alla loro importanza può aiutare a stabilire priorità. Doveroso inoltre considerare fattori esterni che potrebbero influenzare le priorità, come requisiti normativi, esigenze degli utenti o impatti sociali ed etici. Vediamo come ciascuno dei fattori elencati contribuisce a questa considerazione:

Fattore	**Obiettivo**	**Approccio**
Dominio	Considerare quanto il modello e le relative spiegazioni siano rilevanti per il dominio specifico in cui operano.	Adattare il linguaggio delle spiegazioni e il livello di dettaglio tecnico in base al contesto del dominio. Le spiegazioni dovrebbero essere comprensibili e utili agli stakeholder del dominio specifico.

Casi d'uso	Adattare le spiegazioni alle esigenze specifiche di ogni caso d'uso supportato dal modello di AI.	Identificare le informazioni più rilevanti per ciascun caso d'uso e presentare spiegazioni in modo contestualmente significativo. Ad esempio, un caso d'uso clinico potrebbe richiedere spiegazioni diverse da un caso d'uso finanziario.
Tipo di dati	Riconoscere come il tipo di dati utilizzati influenzi le spiegazioni richieste.	Personalizzare il processo di spiegazione in base al tipo di dati. Ad esempio, per dati strutturati come tabelle, le spiegazioni possono concentrarsi su attributi specifici, mentre per dati non strutturati come immagini, possono essere visualizzate le caratteristiche più rilevanti.

Impatto AI sulle persone	Valutare come le decisioni del modello influenzano le persone coinvolte.	Fornire spiegazioni che consentano agli utenti di comprendere l'impatto delle decisioni del modello sulle loro vite. Questo è particolarmente importante quando il modello può avere conseguenze significative.
Target Audience	Riconoscere chi sono gli utenti finali o gli stakeholder che utilizzeranno le spiegazioni. Esperti di dominio, utenti coinvolti nelle decisioni, certificatori, data scientist, developer, Manager.	Adattare il contenuto e il formato delle spiegazioni in base alla conoscenza e alle aspettative specifiche dell'audience. Le spiegazioni per un pubblico tecnico potrebbero differire da quelle per un pubblico non tecnico.
Esperti di Dominio	Coinvolgere gli esperti del dominio nella definizione delle spiegazioni.	Collaborare con gli esperti di dominio per identificare le informazioni più critiche e rilevanti. Questo assicura che le spiegazioni siano in sintonia con le conoscenze e le aspettative degli esperti.

Utenti coinvolti nelle decisioni	Considerare gli utenti o le parti interessate che sono direttamente coinvolti nelle decisioni basate sul modello.	Assicurarsi che le spiegazioni siano chiare e pertinenti per gli utenti coinvolti, facilitando così la comprensione delle decisioni del modello e la loro accettazione.

Tabella 3.3: Esempi di Fattori da considerare

Considerare attentamente questi fattori multipli consente di personalizzare le spiegazioni in modo efficace, migliorando la trasparenza e la comprensibilità del modello di AI per una vasta gamma di utenti e contesti.

1.5 **Selezione delle Spiegazioni Prioritarie**

Sulla base degli output dei precedenti task si procede con un'analisi delle esigenze specifiche valutando le richieste e le necessità di spiegazione da parte di vari stakeholder procedendo con una loro priorizzazione in base al contesto. Si devono determinare quali spiegazioni sono più importanti in base al contesto operativo e agli obiettivi del sistema AI, considerando di fatto che in alcuni domini le spiegazioni prioritarie potrebbero riguardare le decisioni critiche. Per determinare le diverse priorità occorre valutare in primis l'impatto delle spiegazioni. È necessario considerare come diverse spiegazioni possono influenzare il processo decisionale, la fiducia degli utenti e la conformità alle normative. In questa fase potrebbe essere utile considerare e determinare alcune linee guida nella scelta delle tecniche di explainable previste nelle fasi successive. Decidere quali tecniche di spiegazione (locali, globali, basate su modello o post-hoc) sono più adatte per comunicare efficacemente le priorità identificate. Raccolto e razionalizzato il maggior numero di informazioni riguardante le caratteristiche che potrebbero influenzare e determinare le proprietà

nelle spiegazioni, per formalizzare un risultato concreto nella selezione delle spiegazioni prioritarie è necessario utilizzare un approccio basato su criteri. Tali criteri vanno costruiti secondo una modalità che li renda oggettivi e misurabili, considerando, ad esempio, aspetti come la rilevanza per l'utente, la comprensibilità e la coerenza con gli obiettivi generali del sistema di AI.

In sintesi, questa fase è progettata per garantire che le spiegazioni fornite dal sistema di AI siano contestualmente rilevanti, comprensibili e allineate agli obiettivi generali del sistema, considerando le esigenze e le prospettive di diversi stakeholder.

Deliverables

Nella fase di individuazione del contesto nell'ambito dell'Explainable AI (XAI), è possibile definire diversi deliverables per i diversi task individuati.

Task	**Deliverable**
1.1 Analisi del contesto	• Documento che identifica gli attori chiave, il contesto operativo e i requisiti normativi.
1.2 Definizione degli obiettivi	• Documento che delinea gli obiettivi specifici della spiegazione per ciascun gruppo di stakeholder.
1.3 Valutazione delle prospettive diverse	• Report che riassume le diverse prospettive considerate e le relative raccomandazioni per la personalizzazione delle spiegazioni.
1.4 Considerazione dei fattori multipli	• Report che descrive i diversi Fattori individuati. • Matrice che assegna pesi e priorità ai diversi fattori considerati durante la personalizzazione delle spiegazioni.

1.5 Selezione delle spiegazioni prioritarie	• Guida o documento che presenta le spiegazioni prioritarie selezionate per ciascun caso d'uso o gruppo di utenti. • Strumento per il calco delle priorità.

Tabella 3.4: Possibili deliverables Fase 1

3.2.2 Fase 2: Requirements for Data Collection

La fase di raccolta dei requisiti per la data collection nell'ambito di una metodologia di Explainable AI (XAI) è cruciale per assicurare che i dati utilizzati per addestrare il modello e fornire spiegazioni siano di alta qualità e siano in grado di supportare efficacemente la trasparenza del processo decisionale. Questa fase è orientata a definire i requisiti per la raccolta dei dati, con particolare attenzione all'etichettatura e alla selezione dei dati di input. Partendo dagli output della fase 1, che determinano in primis contesto, stakeholder e priorità, si definiscono in maniera organica e completa i requisiti per la fase di data collection , considerando gli aspetti di explainable e finalizzati ad attuare un processo trasparente sin dall'inizio del processo di costruzione di sistemi di AI.

Task

2.1 **Identificazione di dati rappresentativi**

Il task prevede, in primis, l'analisi di quali tipi di dati (quantitativi, qualitativi, temporali, geospaziali, etc.) sono stati individuati come necessari per i modelli di AI e le relative spiegazioni. Bisogna garantire, poi, che il set di dati utilizzato per addestrare il modello rifletta in modo adeguato la diversità e la complessità del contesto in cui il modello verrà utilizzato. Questo contribuirà a garantire che le spiegazioni siano robuste in diverse situazioni. resta necessario, inoltre, identificare i dati specifici che possono essere utilizzati per spiegare razionalmente le decisioni del modello. Questi dati possono includere esempi significativi, casi di test specifici o scenari che mettono in luce il ragionamento del modello. Nell'identificazione dei dati da collezionare, già in questa fase, è possibile valutare e mitigare, a design time, i bias nei dati che potrebbero influenzare le spiegazioni. La presenza di bias può compromettere la fiducia degli utenti e la qualità complessiva delle spiegazioni. È necessario inoltre tracciare per ogni dataset analizzato le diverse fonti dati, le modalità con le quali sono stati raccolti, i criteri e i processi di valutazione

per determinare la loro qualità e le relative misure adottate per il loro miglioramento. Infine, prestare particolare attenzione all'etichettatura dei dati, garantendo che le annotazioni riflettano accuratamente le decisioni del modello. Questo è essenziale per fornire spiegazioni corrette e comprensibili.

2.2 Valutazione e definizione dei criteri di qualità dei dati

In aggiunta alle metodologie classiche per la determinazione della qualità dei dati, in questo task si definiscono i criteri fondamentali sulla raccolta dati affinché le spiegazioni possano essere di valore. In particolare, è necessario verificare ed elicitare che i dati utilizzati siano:

- **Rappresentativi**

  I dati devono rappresentare accuratamente la diversità e la complessità del contesto in cui il modello verrà utilizzato. Devono catturare le variazioni significative e riflettere tutte le possibili condizioni in cui il modello dovrà operare. E' necessario esaminare la distribuzione dei dati per assicurarsi che sia proporzionata e rifletta in modo accurato le variazioni che il modello potrebbe incontrare nella fase di implementazione.

- **Aggiornati**

  I dati devono essere attuali e rappresentare la situazione corrente del dominio di applicazione. Questo è particolarmente importante se ci sono cambiamenti nel tempo che possono influenzare il comportamento del modello. In questa attività si definiscono i criteri per monitorare la temporalità dei dati e assicurarsi che siano aggiornati regolarmente. Verificare, inoltre, se i dati rispecchiano eventuali cambiamenti significativi nel contesto del dominio.

- **Appropriati**

  I dati devono essere appropriati per il problema specifico che il modello sta affrontando. Ciò significa che devono coprire tutte le variabili rilevanti e fornire informazioni significative per il processo decisionale.

Bisogna analizzare, inoltre, la completezza dei dati rispetto alle variabili rilevanti e confermare che non ci siano lacune significative nelle informazioni che potrebbero influenzare la capacità del modello di generalizzare e spiegare le decisioni.

- **Adeguati**

  I dati devono essere sufficientemente dettagliati e accurati per supportare il processo decisionale del modello. Devono contenere informazioni pertinenti e affidabili per garantire che le spiegazioni siano significative. Definire le modalità per valutare la precisione e la qualità dei dati. Esaminare la presenza di errori, outliers e assicurarsi che i dati siano sufficientemente dettagliati per fornire spiegazioni comprensibili e coerenti.

La verifica di questi aspetti richiede spesso un approccio combinato di analisi esplorative dei dati, confronto con il dominio di applicazione, e monitoraggio continuo durante l'intero ciclo di vita del modello. La qualità dei dati è essenziale per ottenere spiegazioni accurate e affidabili, contribuendo così al successo dell'approccio di Explainable AI.

2.3 Definizione e valutazione dei rischi nell'utilizzo dei dati

Nell'ambito della fase di definizione dei requisiti del dato, è fondamentale comprendere e gestire i rischi associati all'utilizzo dei dati. Ecco alcuni dei principali rischi da considerare che possono emergere durante questa fase:

- **Bias nei dati**

  I dati potrebbero riflettere pregiudizi o distorsioni presenti nella raccolta originale, introducendo bias nel modello. Il modello può imparare da dati distorti, portando a decisioni o spiegazioni che riflettono pregiudizi esistenti nel dataset. In questo task si effettua un'analisi approfondita dei dati per identificare e correggere eventuali bias, utilizzando tecniche di bilanciamento e correzione per mitigare

il problema.

- **Mancanza di rappresentatività**

  I dati potrebbero non rappresentare accuratamente la diversità del contesto in cui il modello verrà implementato. Il modello potrebbe non essere in grado di generalizzare efficacemente situazioni non coperte dai dati, portando a spiegazioni poco rappresentative. E' necessario assicurarsi che il dataset sia sufficientemente diversificato e rappresentativo del contesto operativo del modello.

- **Dati inconsistenti, errati o obsoleti**

  La presenza di dati errati o inconsistenti nel dataset può determinare che il modello può apprendere relazioni errate o produrre spiegazioni incoerenti a causa di dati inaccurati. Inoltre, i dati potrebbero diventare obsoleti nel tempo, non riflettendo più accuratamente la situazione attuale Determinare quali controlli di qualità dei dati sono da effettuare per identificare e correggere eventuali errori, utilizzare tecniche di pulizia dei dati e monitorare la temporalità dei dati e aggiornare regolarmente il dataset per riflettere cambiamenti nel contesto.

- **Overfitting ai dati di training**

  Il modello potrebbe adattarsi eccessivamente ai dati di addestramento, perdendo la capacità di generalizzare e potrebbe quindi fornire spiegazioni che sono troppo specifiche per i dati di addestramento, non generalizzando bene a nuove situazioni. In linea con la ML metodology è possibile utilizzare tecniche di regolarizzazione e assicurarsi che il dataset contenga una varietà sufficiente di esempi.

- **Dati sensibili o riservati**

La presenza di dati sensibili o riservati potrebbe comportare rischi legati alla privacy e alla sicurezza e la divulgazione non autorizzata di informazioni sensibili potrebbe violare normative sulla privacy e danneggiare gli utenti. Questo tipo di dati potrebbe non essere coinvolto nella costruzione dei servizi o di un particolare modello di AI ma potrebbe essere utilizzato nelle tecniche di spiegabilità. Infatti, per spiegare le decisioni prese potrebbe essere necessario correlare dati aggiuntivi rendendo fruibile, di fatto, l'accesso a informazioni riservate. Bisogna quindi valutare la possibilità di anonimizzare o criptare dati sensibili, adottare misure di sicurezza adeguate e rispettare le normative sulla privacy.

- **Etichettatura dei Dati**

  Prestare particolare attenzione all'etichettatura dei dati, garantendo che le annotazioni riflettano accuratamente le decisioni del modello. Questo è essenziale per fornire spiegazioni corrette e comprensibili.

2.4 Modalità creazione e proprietà caratterizzanti dei dati sintetici

La creazione di dati sintetici è una pratica utile in molteplici scenari, inclusi quelli legati alla XAI. La generazione di dati sintetici oltre a mitigare la scarsità di dati, costruendo una maggiore rappresentatività dei dati può essere utilizzata per dare maggiore equità e trasparenza ai sistemi di AI. Infatti il suo impiego in un processo caratterizzato da modelli e sistemi di XAI ha un impatto significativo in diverse aree chiave:

- **Costruzione, validazione e certificazione dei modelli**

  I dati sintetici possono servire come sostituti quando i dati di addestramento sono sensibili e non possono essere condivisi ulteriormente. Possono abilitare, quindi, la costruzione, la validazione e anche la certificazione, di un modello rendendo il processo più sicuro e rispettoso della privacy.

- **Privacy e ampio coinvolgimento degli stakeholder**

  Poiché i dati sintetici non sono soggetti a rischi legati alla privacy, consentono di coinvolgere un gruppo molto più ampio di stakeholder e comunità nel processo di spiegazione e validazione del modello. Questo favorisce un auditing algoritmico trasparente e contribuisce a garantire la sicurezza dei sistemi alimentati da ML attraverso la pratica dell'XAI.

- **Riduzione del bias e maggiore diversità dei dati**

  I dati sintetici possono ridurre il bias garantendo agli utenti una diversità di dati che rappresenti in modo fedele il mondo reale. Poiché i dataset sintetici sono etichettati automaticamente e possono includere casi rari ma cruciali, a volte sono preferibili ai dati del mondo reale.

- **Performance e generalizzazione**

  I modelli di apprendimento automatico addestrati con dati sintetici possono avere performance diverse da un modello addestrato su dati reali. In alcuni casi superiori, in altre inferiori. In quest'ultimo caso può essere necessario valutare il giusto trade off tra performance e requisiti di spieglabilità e/o data privacy.

Al fine di costruire un sistema performante ma soprattutto trasparente di XAI è necessario tracciare tutte le attività atte a determinare la modalità di creazione e le proprietà caratterizzanti dei dati sintetici. Nell'esplorare e selezionare il modello di generazione più adatto alle caratteristiche del dominio e del dataset individuato, oltre a scegliere un modello che possa replicare fedelmente la distribuzione dei dati reali, mantenendo le relazioni e le strutture importanti, alcuni requisiti espressi nella Fase 1 potrebbero indirizzare la scelta del modello verso approcci che rendono più chiaro e comprensibile il processo generativo. Infatti, quando si esaminano le diverse tecniche di generazione di dati sintetici, come Generative Adversarial Networks (GANs), Variational Autoencoders

(VAEs), o altre tecniche basate su regole, oltre a valutare la capacità di ciascuna tecnica di replicare le caratteristiche rilevanti del dataset originale, potrebbe essere importante comprendere a diversi stakeholder la logica sottesa al processo generativo. Alcune tecniche come le GAN potrebbero risultare più performanti da un punto di vista di qualità inteso come performance del modello finale di AI, ma potrebbero risultare non compliance ai requisiti di trasparenza perché più complessi da spiegare e comprendere. Per adattare, quindi, il modello di generazione in base alle peculiarità del dominio e alle specifiche esigenze di spiegazione e garantire che il modello sintetico mantenga le proprietà chiave necessarie per il contesto di utilizzo, è necessario comprendere ed esplicitare un insieme di proprietà caratterizzanti dei dati sintetici:

- Preservazione delle distribuzioni statistiche: verificare che le distribuzioni statistiche degli attributi nei dati sintetici siano coerenti con quelle dei dati reali.
- Conservazione delle relazioni: accertarsi che le relazioni tra le caratteristiche(feature) nei dati sintetici riflettano accuratamente quelle dei dati reali.
- Inclusione di anomalie e casi estremi: assicurarsi che i dati sintetici contengano anomalie e casi estremi presenti nei dati reali per garantire che il modello sintetico sia in grado di catturare comportamenti inusuali o scenari rari.
- Rappresentazione adeguata delle categorie: verificare che le categorie e le classi dei dati sintetici riflettano accuratamente quelle dei dati reali. Le categorie e le classi devono essere ben rappresentate nei dati sintetici per garantire una generalizzazione appropriata.
- Considerazione dei contesti temporali o spaziali: se i dati reali includono aspetti temporali o spaziali, assicurarsi che il modello sintetico rifletta tali dinamiche.

- Rispetto delle limitazioni etiche e di privacy: assicurarsi che la generazione di dati sintetici rispetti le limitazioni etiche e di privacy del contesto applicativo per evitare la generazione di dati sintetici che possano compromettere la riservatezza o introdurre bias etici.
- Verifica con stakeholder: coinvolgere gli stakeholder chiave per valutare la qualità e la rappresentatività dei dati sintetici rispetto alle esigenze specifiche al fine di soddisfare le aspettative degli utenti e degli esperti di dominio.

Deliverables

Task	**Deliverable**
2.1 Identificazione di dati rappresentativi	• Documentazione che contempla i data requirement e specifica i requisiti per i dati utilizzati nel contesto della XAI, la governance dei dati, tipologia di dati, mappatura delle fonti, procedure di raccolta, proprietà desiderate, lo schema dei dati, il vocabolario dei metadati. • Linee guida per l'etichettatura dei dati, un documento che descrive linee guida chiare per l'etichettatura dei dati, garantendo che siano annotate in modo corretto e significativo.

2.2 Valutazione e definizione dei criteri di qualità dei dati	• Documentazione delle procedure e delle metriche utilizzate per garantire la qualità dei dati, come verifica dell'integrità, pulizia dei dati, e identificazione di outliers. • Report di valutazione della qualità, un rapporto dettagliato che documenta le valutazioni della qualità dei dati, identificando punti di forza e debolezza nel dataset. • Piano di miglioramento della qualità dei dati, un piano che descrive le azioni specifiche che saranno intraprese per migliorare la qualità dei dati, con indicazione di chi è responsabile di ciascuna azione. • Registro delle misure adottate per la risoluzione dei problemi di qualità, un registro che traccia le misure adottate per risolvere i problemi di qualità dei dati, come l'aggiunta di dati mancanti, la correzione di errori, o la rimozione di dati incoerenti.
2.3 Definizione e valutazione dei rischi nell'utilizzo dei dati	• Matrice dei rischi associati ai dati, un documento che elenca e valuta i potenziali rischi legati ai dati e al definizione dei criteri utilizzati per valutare il rischio.
2.4 Modalità creazione e proprietà caratterizzanti dei dati sintetici	• Documentazione che specifica tutte le caratteristiche e le proprietà caratterizzanti dei dati sintetici utilizzati.

Tabella 3.5: Possibili deliverables Fase 2

3.2.3 Fase 3: XAI Implementation

Concluse le fasi 1 e 2 che hanno l'obiettivo di caratterizzare i casi d'uso attraverso l'emersione di requisiti di diverse tipologie, la fase di implementazione è fondamentale per mettere in pratica le tecniche di spiegabilità in rispetto, appunto, dei requisiti definiti in precedenza. Questa fase rappresenta un forte punto di contatto, imprescindibile, con le metodologie AI preesistenti, dal momento che i requisiti individuati nelle fasi precedenti, se da un lato indirizzano e definiscono il periodo per lo sviluppo di moduli XAI, dall'altro potrebbero far emergere dei requisiti che impattano anche nello sviluppo dei modelli di AI. Potrebbe determinarsi ad esempio, l'impossibilità di utilizzare alcuni tipi di tecnologie AI per una mancanza di supporto ai requisiti elicitati nelle fasi precedenti, obbligando, di fatto, l'utilizzo di soluzioni alternative a scapito delle performance. Partendo, quindi, dall'analisi dei modelli AI sviluppati o da sviluppare, in questa fase si determinano le migliori soluzioni di XAI comprensive anche di un sistema di monitoraggio e valutazione dell'efficacia.

TASK

3.1 **Integrazione strategie di implementazione dei modelli AI**

Dalle fasi precedenti sono stati consolidati e documentati i requisiti di spiegabilità che si riferiscono a diverse prospettive e fattori che concorrono alla specializzazione dei possibili casi d'uso da soddisfare. Tali requisiti possono determinare una riconsiderazione delle priorità nello sviluppo di modelli di ML, dove la chiarezza e la trasparenza possono prendere il sopravvento sulla massimizzazione delle prestazioni. Questo può portare ad un cambiamento significativo sia nella scelta del modello che nelle strategie di sviluppo e implementazione. Infatti se le metodologie di sviluppo Machine Learning (più in generale di AI) preesistenti guidano lo sviluppo di modelli guardando soprattutto alla efficacia ed efficienza di soluzioni performanti per rispondere alle richieste dei diversi casi d'uso, questo task si pone come punto di contatto e di

guida affinché lo sviluppo del modello tenga conto dei vincoli stabiliti nelle precedenti fasi. È utile, infatti, procedere con un'analisi accurata dei modelli da rendere spiegabili per definire alcuni requisiti per i task successivi. I requisiti di XAI sulla scelta del modello di ML sono significativi e possono influenzare vari aspetti del processo di sviluppo e implementazione di sistemi AI. Se alcune attività non erano considerate e valutate, con l'integrazione dei sistemi di XAI va rivisto il processo di sviluppo estendendo ed integrando task non previsti. Di seguito si descrivono alcuni dei principali impatti sullo sviluppo di sistemi ML based:

- **Selezione del modello**: modelli ML più complessi come le reti neurali profonde possono essere molto performanti ma meno interpretabili. I requisiti di XAI possono indurre alla scelta di modelli più semplici e intrinsecamente interpretabili come alberi decisionali o modelli lineari, specialmente in settori regolamentati come la finanza o la sanità.
- **Complessità vs. interpretabilità**: i requisiti di XAI possono richiedere un trade-off tra complessità e interpretabilità. Potrebbe essere necessario sacrificare una certa misura di performance per garantire che il modello sia sufficientemente trasparente e le sue decisioni comprensibili. Infatti, la necessità di spiegazioni può richiedere di scendere a compromessi sulla precisione o sulle prestazioni del modello per garantire una maggiore comprensibilità.
- **Documentazione e conformità**: i modelli selezionati devono essere accompagnati da una documentazione che spieghi il loro funzionamento e le decisioni in modo che sia conforme con le aspettative normative e di settore. La documentazione deve essere dettagliata, includendo le ragioni dietro ogni iterazione del modello e come questo impatta le spiegazioni riportando anche le modifiche apportate per migliorare l'interpretabilità. Va considerato che tali

attività richiedono tempo e risorse aggiuntive che impattano sul processo di sviluppo. Infine, la conformità con le normative vigenti, come il GDPR, che richiede la trasparenza nelle decisioni algoritmiche, può essere un fattore determinante nella scelta del modello.

- **Sviluppo e validazione**: i requisiti di XAI possono prolungare le fasi di sviluppo e validazione dei modelli per includere test addizionali sull'interpretabilità e sulla chiarezza delle spiegazioni fornite dal modello.
- **Processo di addestramento**: durante l'addestramento, si potrebbe richiedere di utilizzare tecniche di regolarizzazione o di ottimizzazione che promuovono la spiegabilità, come la feature importance o l'analisi dei contributi delle variabili.
- **Risorse e capacità computazionali**: le tecniche di XAI possono richiedere capacità computazionali aggiuntive o l'uso di strumenti esterni che supportano l'interpretazione dei modelli, influenzando la selezione del modello in base alle risorse disponibili.
- **Engagement degli stakeholder**: la necessità di spiegare i modelli agli stakeholder non tecnici può influenzare la scelta verso modelli che, pur essendo meno performanti, permettono una maggiore facilità di spiegazione e quindi un maggiore engagement.
- **Processo di testing**: le procedure di testing del modello potrebbero dover essere aggiornate per includere la valutazione dell'efficacia delle spiegazioni, oltre alle metriche di performance tradizionali. Oltre a documentare in dettaglio come i modelli solo validati è necessario descrivere come le spiegazioni vengono generate e validate, inclusi i casi di test e i risultati, per soddisfare i requisiti di trasparenza e conformità. Anche il valutare metriche di performance standard come accuratezza, precisione, recall e AUC, il testing può non soddisfare i requisti, determinando, di fatto, la necessita di includere la valutazione dell'interpretabilità e della chiarezza delle spiegazioni

generate dal modello. Inoltre, bisogna assicurarsi che le spiegazioni fornite dal modello riflettano fedelmente il suo processo decisionale. Questo può richiedere test aggiuntivi per verificare la corrispondenza tra le spiegazioni e l'output del modello. Il testing avanzato per l'XAI può richiedere risorse aggiuntive e formazione specifica per i tester, aumentando così il tempo e il costo del processo di testing. Integrando l'XAI nel processo di testing, si aumenta la robustezza e la trasparenza del modello di ML, ma si aggiungono anche complessità e richieste di risorse. Queste considerazioni devono essere bilanciate con le esigenze operative e gli obiettivi generali del progetto.

- **Monitoraggio e versioning dei modelli**: è necessario adottare sistemi affidabili di monitoraggio e versioning dei modelli per assicurare il corretto tracciamento dei modelli utilizzati al fine di garantire i requisiti richiesti. Il monitoraggio deve ora tener conto non solo delle prestazioni del modello ma anche della qualità e coerenza delle spiegazioni nel tempo. Ciò può richiedere l'implementazione di metriche e dashboard aggiuntive per tracciare l'interpretabilità. L'introduzione dei sistemi XAI impatta fortemente sui sistemi di versioning che devono essere più granulari, registrando non solo le modifiche al modello ma anche le modifiche alle tecniche di spiegazione e ai parametri pertinenti. Ogni versione del modello deve essere tracciata con un record delle spiegazioni corrispondenti per garantire che i cambiamenti non degradino la qualità delle spiegazioni fornite. La necessità di mantenere modelli altamente interpretabili può comportare aggiornamenti più frequenti, poiché le spiegazioni potrebbero richiedere di essere adeguate in risposta a nuovi dati o feedback degli utenti. Ogni aggiornamento del modello richiede una valutazione per determinare se sono stati introdotti nuovi bias e se le spiegazioni rimangono valide e non discriminatorie.

3.2 Implementazione dei modelli di explainability

In questo task, vengono selezionate e adottate le strategie, gli strumenti e i modelli di explainability che meglio rispondono ai requisiti raccolti e alle metriche di valutazione stabilite nelle fasi precedenti. Per procedere ad una razionalizzazione delle migliori scelte da adottare, è possibile suddividere tale task in diversi sotto-task cosi da focalizzarne le attività.

3.2.1 Selezione delle Strategie, Strumenti e dei Modelli di Explainability

Si valutano diverse strategie di explainability come modelli intrinsecamente interpretabili, tecniche post-hoc, o combinazioni di entrambi. Questo include l'analisi delle loro capacità di fornire spiegazioni chiare e la loro aderenza ai requisiti specifici. Definite le strategie si procede con la selezione degli strumenti e i modelli specifici da utilizzare, basandosi sulla loro compatibilità con il contesto di applicazione e le metriche di valutazione che di fatto si allineano, nel miglior modo, ai requisiti del progetto e al contesto di applicazione, considerando fattori come la natura dei dati, la complessità del modello e il pubblico target delle spiegazioni. Nella definizione progettuale è possibile prevedere l'integrazione di diverse tecniche di XAI per fornire spiegazioni complete e multi livello. Scelti strumenti e modelli è necessario determinare come le tecniche di XAI scelte possono essere integrate efficacemente con il modello di ML esistente, assicurandosi che le spiegazioni siano generate in modo efficiente e accurato. Questo task risulta fondamentale per assicurare che le scelte fatte in termini di explainability siano le più adatte alle esigenze del progetto e utili per gli utenti finali.

3.2.2 Definizione delle metriche

Considerati requisiti, strategie, strumenti e dei modelli di Explainability, in questo sotto task si definiscono le metriche che saranno utilizzate per valutare l'efficacia delle spiegazioni fornite dai modelli di XAI. Tali metriche possono essere organizzate secondo un

livello crescente che parte dai criteri per valutare prototipazione e sperimentazione delle prime implementazioni, fino a definire i criteri con i quali tutti gli stakeholder previsti da requisiti devono valutare la qualità delle spiegazioni prodotte. Di seguito sono riportate alcune delle metriche più comuni:

Metrica	**Descrizione**
Metriche di Comprensibilità	**Misura della comprensione**: valuta quanto facilmente gli utenti possono comprendere le spiegazioni fornite dal sistema. **Misura della chiarezza**: valuta la chiarezza delle spiegazioni in termini di linguaggio, struttura e presentazione.
Metriche di qualità delle spiegazioni	**Fedeltà**: misura quanto accuratamente le spiegazioni riflettono il processo decisionale del modello sottostante. **Completezza**: valuta se le spiegazioni coprono tutti gli aspetti rilevanti del processo decisionale del modello. **Coerenza**: misura la coerenza delle spiegazioni in diverse istanze o condizioni operative.
Metriche tecniche	**Robustezza**: Valuta la stabilità delle spiegazioni al variare degli input. **Trasparenza**: Grado in cui il processo decisionale del modello è visibile e comprensibile.
Metriche di Affidabilità	**Affidabilità**: Capacità delle spiegazioni di essere consistenti e affidabili su diversi set di dati. **Generalizzabilità**: Misura la capacità delle spiegazioni di rimanere valide quando il modello è applicato a nuovi dati.

Metriche di impatto sulle prestazioni	**Efficienza computazionale**: Considera l'overhead computazionale introdotto dal processo di spiegazione. **Impatto sull'accuratezza del modello**: Valuta se l'introduzione di spiegazioni ha impatto sull'accuratezza del modello.
Metriche di conformità e aspetti etici	**Conformità Normativa**: Misura il grado di aderenza del sistema XAI alle normative vigenti. **Equità delle Spiegazioni**: Valuta se le spiegazioni e i risultati del modello mostrano bias verso certi gruppi.
Metriche di Usabilità	**Facilità di Implementazione**: Valuta quanto sia semplice integrare la tecnica di spiegazione nel sistema AI. **Integrazione nel Flusso di Lavoro**: Misura l'efficacia con cui le spiegazioni si adattano e supportano i processi decisionali degli utenti.

Tabella 3.6: Possibili metriche per la XAI

Queste metriche aiutano a garantire che i sistemi XAI non solo forniscono spiegazioni tecnicamente accurate, ma sono anche utili, comprensibili, etiche e conformi ai requisiti degli utenti e alle normative vigenti.

3.2.3 Progettazione, Sviluppo e sperimentazione

Il presente sotto task è dove si concretizzano le idee e si verifica l'efficacia delle tecniche di spiegabilità selezionate. Questa attività si concentra sulla progettazione e realizzazione pratica delle soluzioni di XAI e sulla loro valutazione preliminare attraverso la sperimentazione con set di dati limitati per valutare la qualità delle spiegazioni. Lo sviluppo segue metodologie e approcci proprie dei sistemi di AI

con una metodologia AGILE incrementale supportata, appunto, da rapide prototipazioni e sperimentazioni constanti. Si procede quindi preliminarmente con la progettazione dell"architettura completa del sistema XAI, considerando in primis l'integrazione efficace con i modelli di ML esistenti o previsti. È necessario assicurare, inoltre, che il sistema di XAI sia progettato per integrarsi senza problemi con i flussi di lavoro e i processi esistenti, minimizzando la necessità di modifiche radicali alle pratiche di lavoro e stabilire come i dati saranno gestiti all'interno del sistema, inclusa la raccolta, l'elaborazione e l'utilizzo dei dati per generare spiegazioni. Completata la progettazione, si procede con lo sviluppo della soluzione, che come descritto in precedenza segue un approccio a step crescenti. Infatti, i moduli sviluppati sono sperimentati in maniera incrementale su diversi set di dati per valutare la qualità delle spiegazioni generate. Questo include la valutazione delle metriche individuate nel sotto task precedente prevedendo anche la raccolta di feedback dagli stakeholder sull'efficacia delle spiegazioni. Sulla base dei risultati dei test e del feedback, si apportano modifiche e ottimizzazioni dei prototipi di XAI per migliorarne le prestazioni, la precisione delle spiegazioni e l'usabilità. Tutte le attività eseguite devono essere documentate in dettaglio, inclusi i metodi utilizzati, i set di dati impiegati, le difficoltà incontrate e le soluzioni adottate. Raggiunti i risultati soddisfacenti si preparano i prototipi ottimizzati per l'implementazione effettiva, includendo la pianificazione dell'integrazione nei sistemi esistenti e la preparazione per ulteriori test di validazione. Questa fase di prototipazione e sperimentazione è essenziale per assicurare che i modelli di XAI non solo funzionino come previsto in termini tecnici, ma siano anche efficaci dal punto di vista dell'utente finale e aderiscano agli standard di qualità e trasparenza richiesti. Consolidata la maturità dei moduli costruiti si procede, quindi, con lo sviluppo dell'integrazione pianificata di tali moduli all'interno

delle pipeline AI pre-esistenti, al fine di predisporre una soluzione completamente integrata ed ottimizzata. Inoltre, questo task, prevede la progettazione e lo sviluppo della soluzione di monitoraggio per la verifica costante delle prestazioni dei sistemi XAI e la loro aderenza ai KPI e le metriche stabilite.

Deliverables

Task	**Deliverable**
3.1 Integrazione strategie di implementazione dei modelli AI	• Report che descrive con accuratezza gli impatti dei requisiti XAI sulla scelta del modello.
3.2 Implementazione dei modelli di explainability	• Elenco delle strategie di explainability selezionate con giustificazioni basate sui requisiti. • Documento che raccoglie le scelte fatte descrivendo gli strumenti e i modelli scelti con una valutazione comparativa. • Documento di specifica delle diverse metriche associate ai modelli. • Documento di progettazione del sistema XAI. • Report di sperimentazione che include esempi di spiegazioni generate e valutazioni preliminari. • Report dei risultati dei test iniziali e valutazioni delle spiegazioni e piano di miglioramento basato sui feedback ricevuti. • Soluzione software completa di moduli XAI, integrazione, monitoring.

Tabella 3.7: Possibili deliverables Fase 3

3.2.4 Fase 4: XAI Identification

Task

Questa fase rappresenta il processo in cui si applicano i sistemi XAI appena sviluppati a tutti i casi d'uso previsti. Questo step serve a garantire che i modelli di Machine Learning siano non solo tecnicamente validi, ma anche comprensibili e giustificabili agli occhi degli utenti e degli altri stakeholder.

4.1 **Applicazione dei sistemi XAI ai diversi casi d'uso**

Si analizzano i diversi casi d'uso identificati nelle fasi precedenti e si applicano le strategie XAI previste, per identificare le correlazioni e i fattori che influenzano maggiormente le decisioni del modello. Per ogni dataset oggetto dei casi d'uso si procede con un'analisi completa per identificare eventuali bias. Si applicano quindi tutte le metriche previste per valutarne l'aderenza su tutte le casistiche e con tutti i dataset completi. E' necessario, inoltre, confrontare le performance e le spiegazioni del modello con i principi e gli obiettivi XAI predefiniti per assicurarsi che ci sia aderenza, grazie anche al coinvolgimento e supporto degli esperti di dominio utile anche a valutare la coerenza e l'accuratezza delle spiegazioni fornite dai sistemi XAI rispetto alla realtà del dominio specifico. Per ogni caso d'uso è fondamentale che dopo aver estratto la logica razionale di ogni singolo modello di AI, si deve integrare l'output generato all'interno del processo decisionale. La sfida è rappresentata dal ricondurre i fattori ritenuti determinanti ad una visione completa e direttamente comprensibile ai diversi stakeholder. I sistemi di XAI possono supportare sicuramente l'emersione delle dipendenze statistiche, ma la trasformazione di tali caratteristiche in una spiegazione più completa, necessita di uno sforzo maggiore basato sicuramente sul contesto dove si opera. Infatti, un approccio che tiene conto del contesto specifico, facilita l'individuazione di attributi e dinamiche significative che possono avere un impatto concreto sulla situazione in esame. Questo processo è essenziale perché si concentra

direttamente sulla persona al centro della decisione. È necessario, inoltre, esaminare attentamente le correlazioni pertinenti per assicurarsi che contribuiscano ad una spiegazione valida. Questo implica verificare che tali correlazioni non siano false o influenzate da variabili non considerate e valutare l'adeguatezza delle inferenze statistiche rispetto alle condizioni particolari della persona coinvolta. Questo task assicura, quindi, che l'implementazione di XAI non si limiti solo alla funzionalità tecnica, ma si estenda anche all'effettiva utilità pratica delle spiegazioni generate, consentendo una maggiore fiducia e trasparenza nei modelli di AI utilizzati nei sistemi di supporto alle decisioni.

4.2 Integrazione nei processi decisionali

Questo task si focalizza su come trasformare le informazioni tecniche e statistiche ottenute dai modelli di AI in spiegazioni comprensibili, che possono essere facilmente integrate nel processo decisionale. È importante riflettere su come la particolare combinazione di elementi che ha influenzato l'esito del modello, insieme alle specificità della situazione del destinatario della decisione, possano fornire una solida base alla valutazione finale della decisione da intraprendere. Determinati quali elementi statistici sono particolarmente rilevanti per il processo decisionale in questione, è necessario comprendere come tradurre le informazioni tecniche in un linguaggio chiaro e accessibile. Questa attività si basa spesso sulla semplificazione dei concetti complessi senza, però, perdere l'essenza informativa. Per raggiungere un risultato ottimale, è necessario che chi utilizza uno o piu risultati dei sistemi AI dovrebbe avere la possibilità di capire senza difficoltà come i risultati statistici siano stati applicati al proprio caso specifico e quali motivazioni abbiano portato il sistema a tale valutazione. Tale obiettivo può essere conseguito mediante una spiegazione in termini chiari e semplici o attraverso qualunque altro metodo che faciliti la comprensione della decisione presa. In base alle diverse prospettive emerse nelle fasi precedenti si rende necessario adattare le spiegazioni in modo che si

allineino con il contesto specifico del processo decisionale. Questo può includere la personalizzazione delle spiegazioni in base alle esigenze e alle circostanze del destinatario della decisione. In questo task, in maniera preliminare, può essere già utile fornire indicazioni sull'utilizzo, o eventualmente lo sviluppo, di supporti visivi come grafici, diagrammi e tabelle riassuntive per accompagnare le spiegazioni testuali, rendendo le informazioni più facili da comprendere e da ricordare. In sintesi, il presente task deve produrre le linee guida su come utilizzare le spiegazioni nel processo decisionale, assicurando che siano facilmente accessibili ed utilizzabili.

Deliverables

Task	**Deliverable**
4.1 Applicazione dei sistemi XAI ai diversi casi d'uso	• Report dettagliato con le analisi delle correlazioni per ciascun caso d'uso. • Report dettagliato con le valutazioni di tutte le metriche per ciascun caso d'uso. • Documentazione dei bias rilevati con raccomandazioni per la loro mitigazione. • Feedback degli esperti e report di verifica della coerenza. • Confronto tra principi XAI e comportamento del modello, con valutazione di aderenza. • Un report che sintetizza i risultati statistici in termini comprensibili per un pubblico non tecnico.
4.2 Integrazione nei processi decisionali	• Linee guida e documentazione per l'uso delle spiegazioni nel processo decisionale.

Tabella 3.8: Possibili deliverables Fase 4

3.2.5 Fase 5: Design and Development User Explanation

Questa fase si concentra sulla costruzione e presentazione delle informazioni in modo chiaro e accessibile, tenendo conto dei requisiti emersi nelle fasi precedenti e personalizzati per le diverse tipologie di utenti (stakholder)

Task

5.1 **Desing User Explanation**

Nel task di design si parte identificando quali informazioni sono rilevanti e necessarie per ogni gruppo di stakeholder, che possono variare da utenti tecnici a non tecnici, da esperti di dominio a clienti finali. Partendo dai deliverable precedenti e integrando i risultati della fase precedente si procede con una mappatura, fortemente dettagliata, delle informazioni necessarie per ciascun gruppo di stakeholder, scegliendo, di fatto, il metodo di presentazione delle informazioni che meglio si adatta al modo in cui vengono prese le decisioni. Avendo elicitato il mapping tra informazioni e stakeholder è possibile costruire e progettare un modello di spiegazione stratificato che consideri le diverse prospettive e priorità delle spiegazioni identificate e offra livelli di dettaglio diversi a seconda delle esigenze degli stakeholder. E' necessario inoltre progettare un sistema che permetta un confronto interattivo tra gli utenti a più livelli, con possibilità di feedback e interazione tra i vari stakeholder. Il tutto deve essere opportunamente progettato per essere integrato in maniera efficiente all'interno dei processi decisionali considerati. Le diverse modalità di presentazione che caratterizzano una strategia di presentazione personalizzata per ciascun tipo di decisione e stakeholder, possono dividersi tra report, dashboard, visualizzazioni, spiegazioni interattive, sviluppo di applicazioni web. È necessario, quindi, sviluppare la progettazione sia dell'interfaccia utente (UI) che dell'esperienza

utente (UX) affinché si possano costruire strumenti end user intuitivi, user-friendly e personalizzati in base alle esigenze individuate. Questo può includere la creazione di mockup, wireframe e prototipi. Tutti gli strumenti devono essere correlati con sistemi di feedback e definizione di metriche per valutarne l'efficacia.

5.2 **Sviluppo, test, integrazione della User Explanation**

Partendo dai deliverable di design del task precedente, si procede con il costruire le funzionalità personalizzate che rispondono alle esigenze specifiche degli utenti, come dashboard personalizzabili, sistemi interattivi, reporitstica avanzata, integrazione di interfacce su sistemi pre-esistenti (etc..) integrando il tutto con i sistemi di XAI sviluppati nella fase precedente, assicurando che le spiegazioni e i dati forniti siano accurati, aggiornati e pertinenti. Sulla base delle diverse tipologie di strumenti sviluppati si procede ad una fase di test approfonditi, inclusi test di usabilità, test funzionali e test di performance, per assicurare che il tutto soddisfi i requisiti e le aspettative degli utenti. Consolidata la fase di testing si procede con l'integrazione e rilascio degli strumenti nei processi decisionali e si avvia la raccolta di feedback continuo dagli utenti per iterazioni e miglioramenti futuri.

Deliverables

Task	**Deliverable**

5.1 Desing User Explanation	• Report dettagliato con la mappatura delle informazioni necessarie per ciascun gruppo di stakeholder. • Descrizione della strategia di presentazione personalizzata per ciascun tipo di decisione e stakeholder. • Definizione del modello di spiegazione multi-livello, con livelli di dettaglio variabili. • Progettazione di tutti gli strumenti end user da integrare, sviluppare, personalizzare incluso gli strumenti di feedback. • Piano di testing. • Piano di integrazione, progettazione piano di feedback e definizione delle metriche per ogni strumento individuato.
5.2 Sviluppo, Testing, Integrazione	• Sviluppo strumenti end user. • Testing degli strumenti. • Integrazione degli strumenti nei processi e avvio valutazione continua di feedback e metriche.

Tabella 3.9: Possibili deliverables Fase 5

3.2.6 Fase 6: User Learning

Quando sistemi di AI sono fortemente integrati nei decision support system e quindi troviamo processi dove decisori umani sono coinvolti in modo significativo in un risultato assistito, è fondamentale che tali decisori debbano essere adeguatamente formati e preparati a utilizzare i risultati del modello in modo responsabile ed equo.

Task

6.1 **Formazione sui temi di AI e le decisioni autonome**

La formazione deve comprendere, in primis, il trasferimento di conoscenze di base sulla natura dell'apprendimento automatico e sui limiti dell'AI e delle tecnologie di supporto alle decisioni automatizzate. Si deve fornire, dunque, un piano formativo che includa una panoramica di come funzionano i modelli di AI, i loro limiti e le potenziali sfide. Il tutto va progettato e programmato esplicitando, in particolare, l'incertezza e affidabilità delle decisioni automatizzate, comprese le questioni di bias che potrebbero verificarsi. L'obiettivo è quindi insegnare agli utenti come interpretare correttamente i risultati forniti dai sistemi di AI e come integrarli in modo responsabile nel processo decisionale.

6.2 **Formazione sugli strumenti end user**

La seconda parte prevede un piano formativo efficace per gli strumenti end user di XAI sviluppati nella fase di "Design and Development of User Explanation", è importante assicurarsi che gli utenti comprendano pienamente come utilizzare questi strumenti e interpretare le spiegazioni fornite. Il piano formativo inizia con sessioni dedicate di presentazione degli strumenti sviluppati, dove viene spiegato in modo chiaro il loro scopo, le funzionalità chiave e i contesti specifici di utilizzo. Questo passaggio è essenziale per assicurare che gli utenti comprendano a fondo come e perché utilizzare gli strumenti XAI. Tale presentazione dovrà fornire dimostrazioni pratiche e tutorial interattivi, per guidare gli utenti

attraverso i vari aspetti degli strumenti, mostrando loro come possono essere utilizzati in scenari reali, rendendo l'apprendimento coinvolgente ed attivo.

Per procedere con una formazione più approfondita, possono essere organizzati workshop interattivi, dove gli utenti hanno l'opportunità di mettere in pratica ciò che hanno imparato, utilizzando gli strumenti XAI con esempi reali o simulati e ricevendo feedback immediati. Questo approccio pratico è fondamentale per consolidare le conoscenze e le competenze acquisite.

Un altro aspetto importante della formazione è educare gli utenti su come interpretare i risultati proposti delle spiegazioni fornite dagli strumenti XAI in relazione alle limitazioni e alle implicazioni delle informazioni presentate, precedentemente affrontate nel task 4.1, assicurando così che gli utenti siano pienamente consapevoli di come utilizzare in modo efficace e responsabile le informazioni ottenute.

Infine, è fondamentale valutare regolarmente le competenze degli utenti nell'utilizzo degli strumenti XAI e raccogliere i loro feedback per apportare miglioramenti continui, considerando, ad esempio, questionari di valutazione, sondaggi, sessioni di raccolta feedback e sessioni di revisione. Questo processo di feedback e valutazione garantisce che il materiale formativo rimanga aggiornato e rilevante, soprattutto considerando la rapidità nelle evoluzioni degli sviluppi nel campo dell'AI e degli strumenti XAI .

Deliverables

Task	**Deliverable**
6.1 Formazione sui temi di AI e le decisioni autonome	• Piano di formazione e sviluppo dei moduli. • Erogazione del piano di formazione.

6.2 Formazione sugli strumenti end user	• Piano di formazione e sviluppo dei moduli. • Erogazione del piano di formazione. • Monitoraggio continuo delle competenze acquisite.

Tabella 3.10: Possibili deliverables Fase 6

CAPITOLO 4

APPLICAZIONE DI TECNICHE E MODELLI DI EXPLAINABLE AI

4.1 Introduzione

In questo capitolo, verranno esaminate l'applicazione di tecniche e modelli di Explainable AI (XAI) in tre differenti scenari. 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. Nel secondo scenario, si tratta della segmentazione di testo in documenti legali, impiegando reti neurali di tipo transformer, e come la XAI può essere applicata per migliorare sia il processo di etichettatura che la robustezza del modello e la sua spiegabilità per l'utente finale. Infine, il terzo scenario emerge 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 sistemi di riconoscimento ottico dei caratteri (OCR), l'approccio proposto 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'intuizione fornita dal framework 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[51]) 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.

4.2 PUMP & DUMP: le frodi nel mondo delle Criptovalute

4.2.1 Introduzione

Questa sezione si dedica all'impiego di tecniche di Explainable AI (XAI) nei sistemi di intelligenza artificiale e di elaborazione del linguaggio naturale (NLP), sviluppati per individuare le frodi di Pump & Dump nel mercato delle criptovalute [55]. La crescente popolarità delle criptovalute ha portato a un incremento sia delle opportunità di investimento che delle truffe, in particolare quelle legate al Pump & Dump. Questa pratica fraudolenta consiste nel manipolare il mercato con informazioni ingannevoli per innalzare artificialmente il valore di una criptovaluta. Nel metodo proposto, l'NLP gioca un ruolo chiave nell'analizzare il linguaggio usato sui social media e nelle chat online, che sono spesso teatro di queste truffe. L'analisi linguistica può svelare indizi di manipolazione, facilitando la rilevazione tempestiva di queste attività illecite. Più precisamente con il P&D, gli speculatori manipolano il mercato diffondendo informazioni false per innalzare il valore di una criptovaluta e poi vendere le loro quote a un prezzo elevato. Come mostrato in Figura 4.1, questa pratica si articola in diverse fasi: accumulo di token, "pumping" (diffusione di notizie false per aumentare il prezzo) e "dumping" (vendita delle monete a prezzi gonfiati).

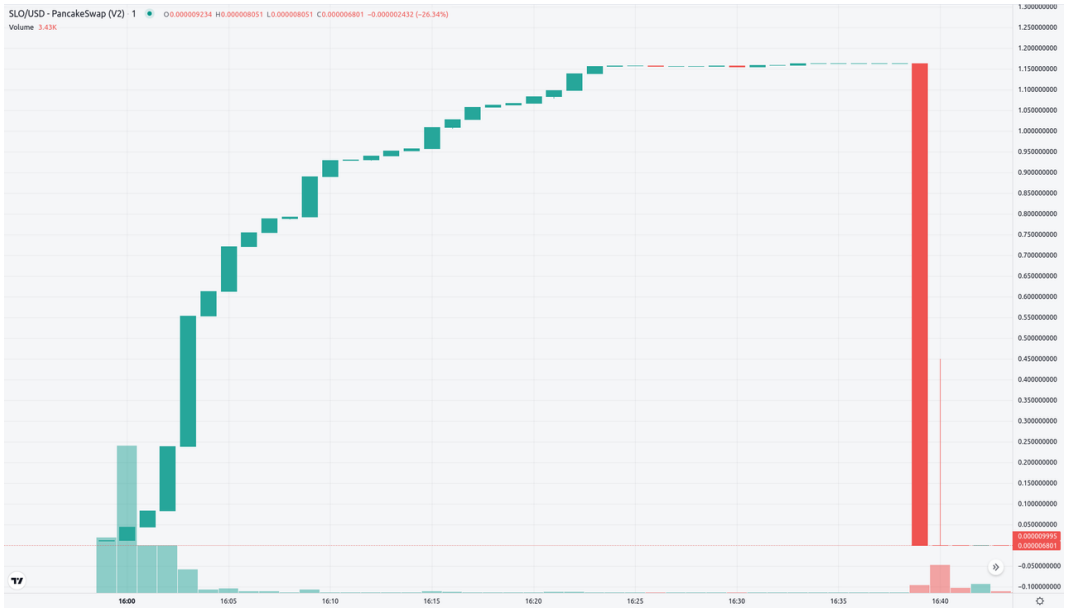


Figura 4.1: Pump and Dump. [56]

Per promuovere la disinformazione sulle criptovalute, vengono formati e sfruttati gruppi pubblici sui social media, come ad esempio su piattaforme come Telegram. L'intento è di puntare su criptovalute meno conosciute, che risultano più semplici da manipolare. Metodi quali la diffusione di notizie false, la creazione di alleanze inesistenti o l'uso di sponsorizzazioni fittizie da parte di celebrità vengono impiegati per ingannare gli utenti. Nella fase di "pumping", a questi utenti viene richiesto di divulgare ulteriormente queste informazioni ingannevoli per influenzare altri investitori.[57]. In uno schema tipico, i leader del gruppo social annunciano in anticipo la manipolazione, spingendo i partecipanti ad acquistare la criptovaluta prima della sua promozione attiva. Questa fase di acquisto anticipato è spesso rafforzata da informazioni false per attirare altri investitori [58]. Parallelamente, un ulteriore analisi della letteratura ha rivelato studi significativi sull'identificazione di anomalie nei social media. Ad esempio, Neda Soltani [59] ha esplorato il rilevamento di anomalie e utenti sospetti usando l'analisi della rete mentre un altro studio interessante [60] ha applicato il modello di Fama French per esaminare l'influenza dei social media sugli investimenti, analizzando la strategia di Wallstreetbets. Non mancano inoltre studi riguardo l'applicazione di tecniche

di Machine Learning per l'identificazione della manipolazione del prezzo del Bitcoin [61], attraverso l'analisi delle emozioni e dei sentimenti espressi dagli utenti sui social media. Questi approcci offrono, di fatto, nuove prospettive sulla comprensione del comportamento del mercato influenzato dai social media.

I truffatori, per manipolare il valore delle valute, promuovono la diffusione di informazioni false, spesso utilizzando bot per amplificare il loro impatto [62]. Dal punto di vista finanziario, le manipolazioni di mercato sono state classificate in tre categorie principali: quelle basate sull'informazione, quelle basate sull'azione e quelle basate sul commercio [63]. I Pump & Dump rappresentano una forma di manipolazione che combina elementi sia dell'informazione che del commercio. Un lavoro significativo in questo ambito è stato condotto da Kamps[58], che ha sviluppato un approccio per rilevare i Pump and Dump attraverso l'uso di una soglia adattativa.

4.2.2 Metodologia proposta e risultati

La metodologia proposta si basa sull'utilizzo innovativo e combinato di due diverse tipologie di informazioni inerenti l'analisi delle criptovolaute: dati finanziari e conversazioni social. I dati finanziari per lo studio fanno riferimento alle transazioni dell'exchange Binance e sono stati inizialmente resi pubblici da La Morgia [64]. Per la costruzione di questi dataset, sono state etichettate le transazioni di diverse criptovalute che erano state identificate come casi noti di Pump and Dump, partendo dai flussi conversazioni social che facevano riferimento a quelle particolari cripto coinvolte. Più precisamente, gli autori, sono partiti analizzando i canali telegram che promuovevano l'acquisto delle criptovalute ed hanno verificato le segnalazioni di pump presenti, forniti dagli stessi amministratori dei gruppi, riuscendo a collezionare 104 occorrenze di dati P&D in due anni. Partendo poi dai timestamp dalle segnalazioni, sono state raccolte le transazioni di bitcoin, moneta utilizzata per acquistare le criptovalute coinvolte, correlate ai pump, attraverso le API di Binance. Dopo aver ottenuto i dati grezzi dall'API di Binance, sono stati sottoposti a un

processo di pre-elaborazione. Le transazioni sono state aggregate in intervalli temporali di cinque, quindici e venticinque secondi, creando tre diversi set di dati che presentano le seguenti caratteristiche:

- PumpIndex, Symbol: l'indice di base 0 della pump, identificandola come una delle 104 pump disponibili, e il simbolo della criptovaluta coinvolta nella pump;
- StdRushOrder, AvgRushOrder: la variazione percentuale media del numero di ordini urgenti e la deviazione standard mobile;
- StdTrades: il numero di deviazioni standard mobili delle operazioni di acquisto e vendita;
- StdVolume, AvgVolume: la variazione percentuale media del volume degli ordini e la deviazione standard mobile;
- StdPrice, AvgPrice, AvgPriceMax: la variazione percentuale media, la deviazione standard e la variazione percentuale massima del prezzo dell'attività.

Come descritto in precedenza, in aggiunta ai dati delle transazioni, la presente metodologia, basa la sua intuizione ed innovazione sull'utilizzo delle informazioni social come ulteriore features da considerare nella valutazione della possibilità che avvenga un fenomeno di P&D. Se l'approccio precedente, partiva dai social solo per identificare ed etichettare l'evento di pump, il metodo proposto, ribalta il suo utilizzo facendone diventare parte attiva per la classificazione. Infatti, per ampliare il dataset con informazioni provenienti dai social, è stata ricondotta una raccolta di messaggi dal social network Telegram, focalizzandosi sugli orari in cui avvenivano i pump e sui link dei gruppi Telegram interessati. L'accesso diretto a questi gruppi ha permesso di scaricare i messaggi attraverso le API di Telegram, coprendo un periodo di due giorni prima e dopo ciascun pump. Tuttavia, le restrizioni nell'accesso alla cronologia completa dei messaggi hanno limitato la portata del dataset raccolto. Successivamente, i messaggi raccolti sono stati sottoposti

a un processo di pre-elaborazione utilizzando tecniche di Natural Language Processing, seguendo una serie di fasi ben definite:

- Bonifica del testo: rimozione di emoji, immagini, stop-word e link esterni;
- Lemmatizzazione: raggruppare le diverse forme inflesse di una parola come unico elemento di base (lemma);
- Pos Tagging: processo di marcatura di una parola in un testo come corrispondente a una particolare parte del discorso si basa sulla sua definizione e sul suo contesto, escludendo le parole che si riferiscono a sostantivi e nomi propri. Questo approccio consente un'analisi più accurata del testo.

Dopo aver raccolto i messaggi dalla chat pertinente per ciascun indice di pump, si è passati alla fase di estrazione delle caratteristiche.

In conclusione, per ogni set di dati creati da [64], per ogni transazione finanziaria e indice di pump, è stata aggiunta una nuova caratteristica. Questa assume valore 1 se il simbolo della valuta è presente nei messaggi scambiati 5 minuti prima della transazione, altrimenti 0. Il set di dati finale, composto da 89 pump, integra questa caratteristica categoriale alle informazioni finanziarie già presenti.

Per valutare l'effetto della nuova caratteristica estratta dal social Telegram, è stato utilizzato un classificatore basato su Random Forest, in continuità con quanto fatto La Morgia[64]. Dal momento che il dataset risulta essere fortemente sbilanciato, in aderenza al caso d'uso, con soli 89 eventi di P&D, non è stato diviso in set di training e test, ma è stata applicata una K-fold cross validation, con K=5, per procedere alla valutazione delle sue performance.

La fase sperimentale ha mostrato un miglioramento statisticamente significativo rispetto all'approccio precedente grazie, appunto, l'integrazione della caratteristica sociale nel classificatore ed evidenzia l'efficacia del nuovo approccio basato sull'analisi delle feature derivate dai dati sociali. I risultati

ottenuti con il classificatore su tutti i diversi dataset, sono descritti nella Tabella 4.2, e confrontati con i precedenti risultati riportati in Tabella 4.1.

Chunk-Size	Precision	Recall	F1
5 Sec	89.1%	83.1%	86.04%
15 Sec	97.5%	89.8%	93.5%
25 Sec	97.5%	91.1%	94.1%

Tabella 4.1: Risultati 5-Folds Random Forest considerando solo caratteristiche finanziarie.

Chunk-Size	Precision	Recall	F1
5 Sec	94.2%	91%	92.5%
15 Sec	98.8%	94.3%	96.5%
25 Sec	97.6%	94.3%	95.9%

Tabella 4.2: Risultati 5-Folds Random Forest considerando caratteristiche finanziarie e social.

4.2.3 Applicazione di tecniche di XAI

L'integrazione dei modelli e tecniche di XAI sono state pensate e personalizzate per supportare la validità del nuovo approccio sia nella fase di costruzione, sia nella fase di spiegabilità dei risultati finali in fase predittiva. Anche nel contesto dei dati tabulari e dell'algoritmo Random Forest l'applicazione di tecniche XAI consente una maggiore trasparenza e comprensione del modello proposto. In particolare nel contesto dei dati tabulari, la XAI si presta maggiormente ad identificare quali caratteristiche sono più importanti per le decisioni di un modello e come queste caratteristiche influenzano le previsioni. Questo è particolarmente utile nei dati tabulari dove le relazioni tra variabili possono essere intricate e non immediatamente evidenti. L'applicazione della XAI per il caso d'uso PUMP & Dump, dunque, è stata condotta attraverso due approcci distinti utilizzando di fatto due tecniche

diverse che si adattano ai singoli obiettivi:Features Contribution Analysis e Single Instance Explanation

Features Contribution Analysis

La Features Contribution Analysis (FCA) è un approccio nell'ambito dell'Explainable AI che si concentra sull'analisi del contributo di ciascuna caratteristica nel processo decisionale di un modello di machine learning. Questa analisi aiuta a comprendere come le singole variabili influenzino il risultato finale del modello. In pratica, si valuta quanto ogni caratteristica contribuisca alla previsione, permettendo agli utenti di identificare quali dati sono più rilevanti per le decisioni del modello. Questo tipo di analisi è particolarmente utile per migliorare la trasparenza dei modelli complessi e per guidare l'ottimizzazione e la selezione delle caratteristiche. Per calcolare la FCA è stato utilizzato SHAP (SHapely Additive exPlanations) [51] un metodo basato sulla teoria dei giochi per spiegare i risultati di modelli di machine learning. Nell'implementazione della libreria nel linguaggio Python [65] per modelli come Random Forest, si usa il TreeExplainer. L'analisi condotta sul dataset oggetto di sperimentazione ha prodotto i seguenti risultati:

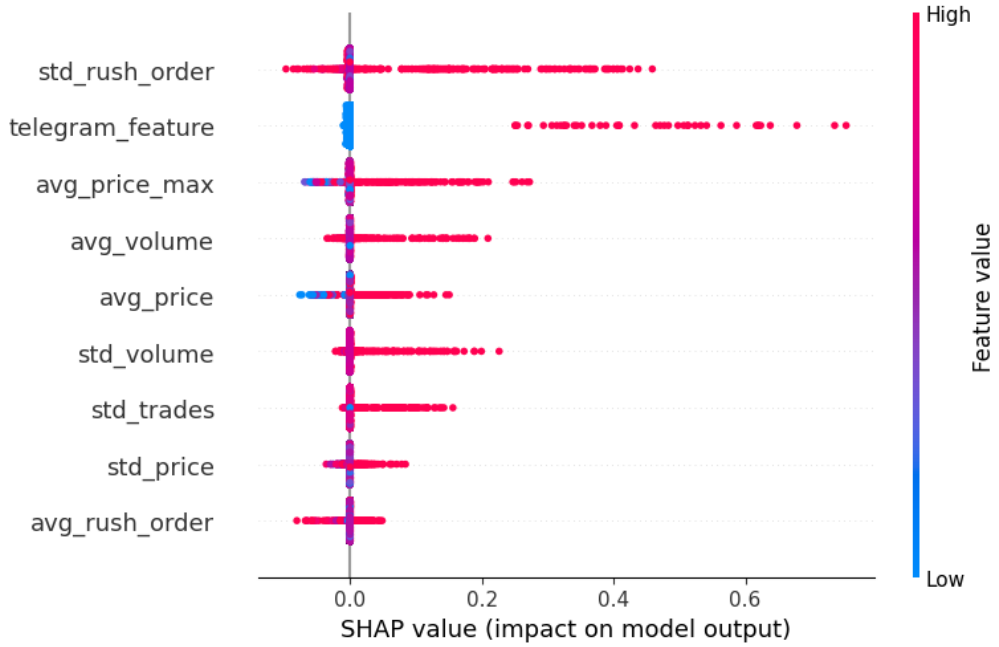


Figura 4.2: Features Contribution Analysis su RF (5 Folds) per Chunk-Size 5 Sec.

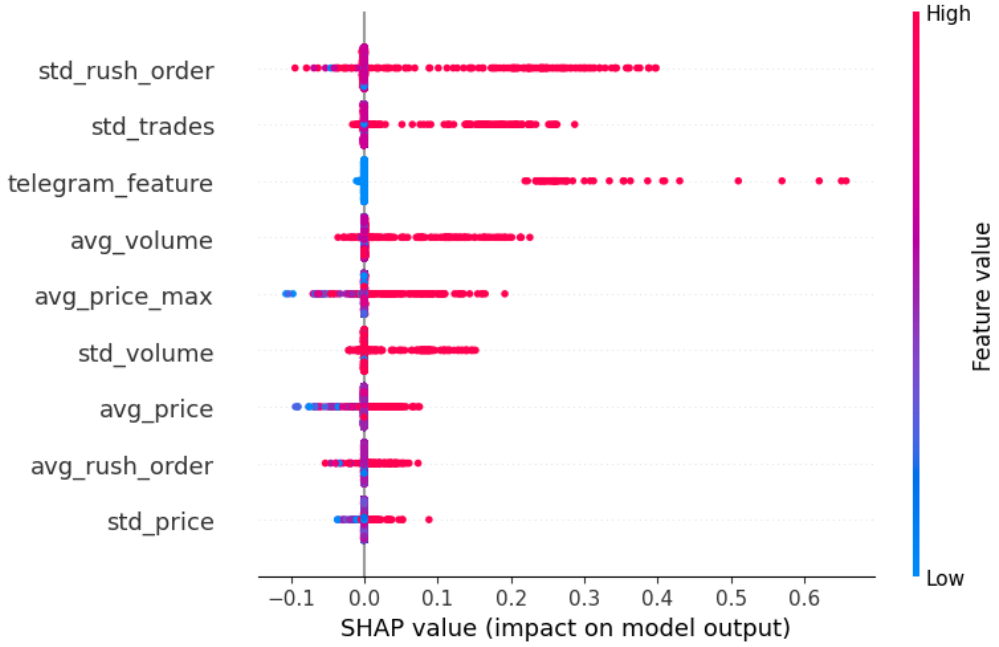


Figura 4.3: Features Contribution Analysis su RF (5 Folds) per Chunk-Size 15 Sec.

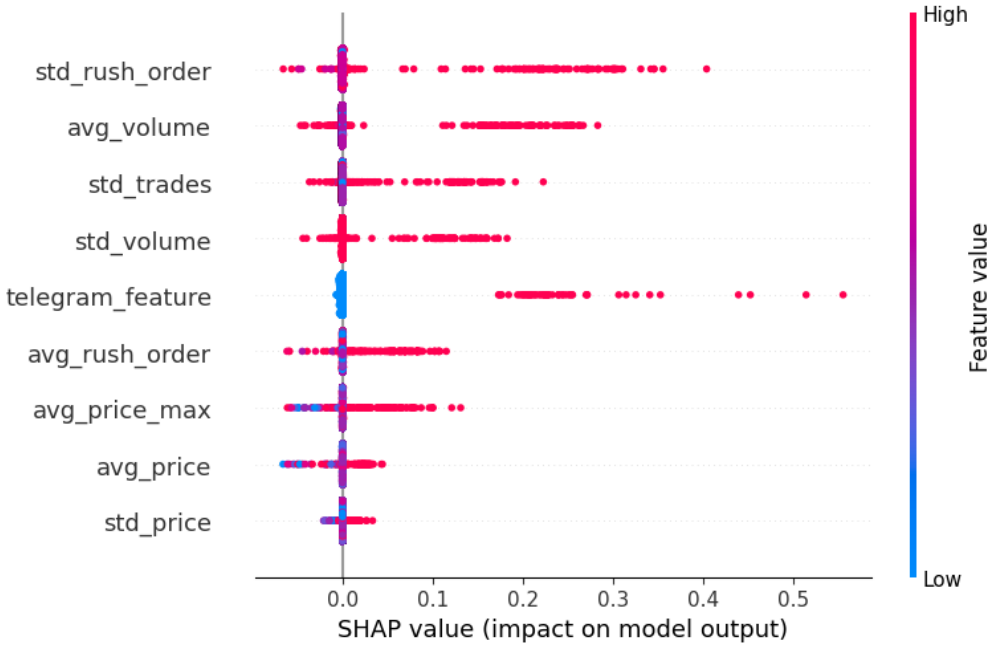


Figura 4.4: Features Contribution Analysis su RF (5 Folds) per Chunk-Size 25 Sec.

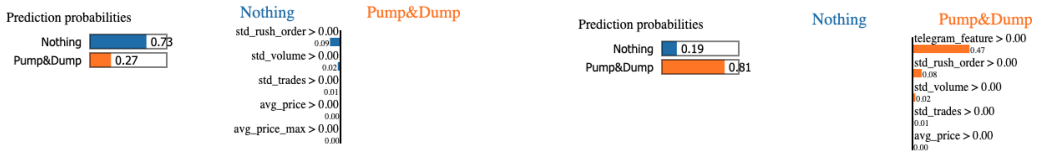
Le figure mostrano che la caratteristica sociale ha un impatto notevole su tutti i set di dati. Tuttavia, la sua importanza diminuisce all'aumentare della dimensione delle finestre temporali. Questo è prevedibile perché, per finestre di 15 e 25 secondi, sono necessari meno dati per catturare la stessa quantità di informazioni rispetto ai dati più dettagliati delle finestre di 5 secondi. Di conseguenza, ciò implica una riduzione delle informazioni sociali presenti nell'input per le dimensioni di finestra più grandi.

Single Instance Explanation

La Single Instance Explanation (SIE) nell'ambito dell'Explainable AI si riferisce all'analisi e alla spiegazione delle decisioni prese da un modello di machine learning per una singola istanza o predizione. Questo tipo di spiegazione aiuta a comprendere quali caratteristiche specifiche del dato di input hanno influenzato maggiormente la previsione del modello, permettendo così agli utenti di ottenere una visione dettagliata del comportamento del modello a livello microscopico. Come strumento per la SIE è stato

(a) Con Social Feature (b) Senza Social Feature

Figura 4.5: Single Instance Explanation Chunk-Size 5 Sec



(a) Con Social Feature (b) Senza Social Feature

Figura 4.6: Single Instance Explanation Chunk-Size 15 Sec

utilizzato LIME (Local Interpretable Model-agnostic Explanations)[44]. LIME è, appunto, uno strumento utilizzato per fornire spiegazioni a livello di singola istanza su come un modello di machine learning, arriva a una specifica previsione, specialmente in presenza di dati tabulari. LIME approssima il modello globale complesso con un modello locale più semplice e interpretabile intorno alla previsione di interesse. Questo permette di capire quali caratteristiche contribuiscono maggiormente alla previsione di una particolare istanza, offrendo così una spiegazione su misura per quella specifica osservazione.

Le analisi descritte nelle figure sono state condotte prendendo in considerazione, per ogni dataset, le istanze che il modello di La Morgia et al. [64] prevede come classe 0 (Nessun evento) e il modello proposto che

(a) Con Social Feature (b) Senza Social Feature

Figura 4.7: Single Instance Explanation Chunk-Size 25 Sec

incorpora le caratteristiche sociali prevede come classe 1 (Pump & Dump). L'analisi dimostra l'importante peso delle caratteristiche sociali per tutti i tipi di dataset. In molte istanze, specialmente quando sono assenti i valori delle altre caratteristiche, è in grado di ribaltare completamente la previsione locale (tutti uguali a 0).

4.2.4 Conclusioni

Lo studio dimostra 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. Infatti, il nuovo approccio, evidenzia l'importanza dei dati dei social media nel rilevare questi schemi nelle criptovalute, superando metodi precedenti e raggiungendo un livello avanzato nello stato dell'arte. 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, la Single Instance Explanation, promuove 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.

4.3 Legal Text Segmentation attraverso Breakpoint Detection

4.3.1 Introduzione

Proseguendo lo studio riferito all'applicazione dei tecniche di XAI su diverse tipologia di casi d'uso, osserviamo ora il contesto della segmentazione automatica di testi, attraverso approcci Machine Learning based. I documenti sono tipicamente formati da segmenti semanticamente coerenti, che variano in lunghezza e argomento. La segmentazione del testo permette di dividere questi documenti in unità più piccole e maneggevoli, note come segmenti ed buon risultato appare come cruciale per estrarre informazioni rilevanti dai testi. Tuttavia, la segmentazione di documenti presenta sfide particolari quando il dominio, come quello legale oggetto della sperimentazione, presenta una complessa terminologia, lunghezza considerevole e possibili distorsioni dovute a testi manoscritti e scansionati.

Nell'ambito della segmentazione del testo, esistono vari approcci neurali, inclusi quelli specifici per il settore legale. Tuttavia, il metodo proposto, introduce un approccio innovativo basato sul rilevamento dei punti di interruzione, o "breakpoint detection". Tale approccio si concentra sulla capacità di identificare e classificare le linee di testo che segnano l'inizio di un nuovo segmento all'interno di un documento.

Procedendo con l'analisi della letteratura, sono stati investigati diversi studi che hanno esplorato le tecniche di segmentazione del testo, analizzandole sia in un contesto più ampio sia specificamente all'interno del settore dei documenti legali. I metodi basati su regole sono stati utilizzati per anni a causa del loro approccio semplice nell'identificare schemi di scrittura all'interno dei testi. Nel loro articolo, Bali et al.[66] esaminano la segmentazione del testo a livello di frase per la lingua malese, tramite la pre-elaborazione del testo e la costruzione di regole appropriate. In modo simile, Saniati et al.[67] propongono un approccio di segmentazione del testo basato sulla distribuzione delle entità

nominate per rilevare i cambiamenti di sotto temi. Anche se i metodi basati su regole sono veloci e performanti, richiedono comunque una manutenzione impegnativa se vengono processati diversi tipi di documenti o se le parole chiave che identificano l'inizio di un segmento sono errate.

Parallelamente, sono stati analizzati anche studi basati su tecniche statistiche e di deep learning, approfondendo alcuni approcci notevoli che utilizzano reti BiLSTM in diverse modalità: un'architettura gerarchica che usa BiLSTMs sia per generare la rappresentazione delle frasi sia per la classificazione [68], o migliorando i BiLSTMs con CNN e meccanismi di attenzione [69]. Una ricerca rilevante esaminata, utilizza i Conditional Random Fields[70] per la segmentazione automatica delle decisioni delle corti ceche in parti composte da più paragrafi [71]. Nonostante la mancanza di documenti e l'uso di un modello probabilistico, il loro metodo raggiunge un punteggio F1 medio di 0.703 su 7 sezioni. Lukasik et al.[72] propongono, invece, tre diverse architetture per la segmentazione del testo basate su BERT[73], sfruttando l'importanza del contesto locale del documento. In modo simile, Glavaš et al.[74] propongono un modello consapevole della coerenza basato su due transformer gerarchici molto efficaci nel trasferimento linguistico zero-shot. Recentemente, Aumiller et al. [75] propongono un nuovo approccio per la segmentazione del testo considerando il cambio di argomento tra le frasi come candidato per un nuovo confine di segmento (STP). L'utilizzo di modelli transformer, inclusi BERT, ha portato significativi avanzamenti nella segmentazione del testo, catturando informazioni contestuali e migliorando l'accuratezza della segmentazione. L'approccio proposto si basa, appunto, su queste tecniche introducendo il rilevamento dei punti di interruzione usando un modello basato su transformer pre-addestrato su testi legali italiani. Passando quindi da un approccio a regole, dove, per caratteristiche costitutive, risulta essere un approccio interpretabile e quindi si comprende, seppur in alcuni casi applicando dei metodi di semplificazione se le regole sono numerose e fortemente sovrapposte, ad un approccio deep learning based, di fatto una scatola nera, la corretta applicazione di metodi e strumenti di Explainable AI

risulta preziosa ai fini di validazione e utilizzo del metodo proposto[76].

4.3.2 Dataset e metodo proposto

Il dataset utilizzato in questo studio comprende 730 documenti della Corte di Cassazione, divisi in 520 per l'addestramento e 210 per il test, con una media di 9563 parole per documento. I documenti, ottenuti da diverse fonti per assicurare varietà e rappresentatività, (*ItalgiureWeb*, *Ricerca Giuridica* e *La Legge per tutti*), sono stati etichettati manualmente con l'aiuto di un motore di regole creato ad hoc, che utilizza regole basate su approcci lessicali e la libreria regex per identificare schemi nei testi legali. Ogni documento è strutturato in sezioni che includono il testo grezzo, le righe con allineamenti e etichette corrispondenti e le sezioni con gli indici di inizio e fine di ciascun segmento. Le sezioni presenti si dividono in:

- intestazione: dettagli della corte di riferimento per la sentenza (non considerata nella formazione/test del modello);
- attori: elenco dei querelanti nel processo;
- imputati: elenco degli imputati nel processo;
- fatti: riassunto delle circostanze di fatto attorno al caso;
- procedimenti: il reale svolgimento del processo;
- motivazioni: ragione dell'esito della sentenza;
- aggiuntivi: informazioni supplementari ritenute rilevanti per la sentenza;
- esito: può includere verdetti, ordini, ingiunzioni o altre determinazioni legali effettuate dalla corte;
- firma: firma del giudice o del presidente del consiglio.

Per la natura del problema esaminato, lo sbilanciamento del dataset appare evidente: solo il 5,7% delle righe totali sono breakpoint, mentre il resto non

lo è. Questo squilibrio, su un dataset di oltre 98.000 righe, potrebbe portare il modello a sovrastimare la classe dominante e a non rilevare accuratamente i breakpoint ma, nonostante lo squilibrio, il modello di segmentazione del testo proposto ha mostrato prestazioni notevoli, ottenendo buoni risultati anche con pochi breakpoint. Sono stai testati, inoltre, diversi approcci per ribilanciare il dataset tramite re-sampling, ma non hanno migliorato le prestazioni, evidenziando, di fatto, le capacità intrinseche del modello di gestire lo squilibrio grazie alla tipologia di architettura utilizzata. In particolare, il modello proposto per la text segmentation in ambito legale, sfrutta un modello basato su transformer specificamente personalizzato sul dataset legale appena descritto. Chalkidis et al. [77] hanno sviluppato il primo modello innovativo basato su transformer per il dominio legale inglese, addestrando BERT [73] sui corpora specifici del settore. Analogamente, ITALIAN-LEGAL-BERT [78] è il risultato di un ulteriore addestramento del modello ITALIAN XXL BERT su corpora di diritto civile italiano. Questo addestramento supplementare è consistito in 4 epoche su 3.7 GB di testo dall'Archivio Giurisprudenziale Nazionale, un repository pubblico contenente milioni di documenti legali italiani. ITALIAN-LEGAL-BERT è stato scelto per sperimentare il nuovo approccio proposto perché ritenuto, in prima istanza, specificamente vantaggioso e pertinente al dominio legale. È stata, quindi, considerata la necessità di un modello addestrato non solo su un ampio corpus di testo italiano pre-elaborato, ma anche progettato per cogliere le complessità e le sfumature del dominio legale. La premessa di partenza è che nei testi legali è possibile riconoscere schemi o modelli ricorrenti specifici su cui eseguire la segmentazione del testo. Da questa assunzione è possibile sviluppare un sistema che identifichi prima paragrafi o sezioni caratteristici, delimitati da specifici titoli o formule introduttive, e poi segmenti il testo. Data questa premessa, possiamo definire un **breakpoint** in un testo legale come una riga nel testo che corrisponde all'inizio di un nuovo segmento all'interno del documento. Poiché le righe dei breakpoint sono i titoli o le formule introduttive delle sezioni, queste sono incluse come prima riga

del segmento. La riga che precede il breakpoint è invece l'ultima riga del segmento precedente e viene inclusa come ultima riga della sezione precedente.

Si consideri $D = \{T_1, T_2, \ldots, T_K\}$ il dataset composto da K documenti testuali e ogni documento $T_i = \{s_1^i, s_1^i, \ldots, s_N^i\}$, una sequenza di N sezioni. Inoltre, ogni sezione $s_j = \{l_1^{i,j}, l_1^{i,j}, \ldots, l_M^{i,j}\}$ è composta da M linee (o frasi). Ad ogni linea è associata un'etichetta di classificazione $y \in (0, 8)$ dove 0 indica che la linea non è un breakpoint, mentre le restanti etichette indicano per quale tipo di sezione rappresenta la linea di breakpoint (di inizio). Il metodo proposto, in definitiva, si concentra sull'identificare le frasi che segnano i confini tra diverse sezioni all'interno di un documento legale, definite come breakpoint. Quando viene rilevato un breakpoint, l'etichetta associata fornisce informazioni sulla classe della sezione. Rilevando efficacemente questi breakpoint, possiamo segmentare accuratamente il documento in unità significative per migliorare il recupero e l'analisi delle informazioni.

4.3.3 Risultati sperimentali

In questa sezione, sono descritti risultati dell'approccio proposto per la segmentazione di documenti legali basato sul modello ITALIAN-LEGAL-BERT. Le metriche di valutazione utilizzate includono il Balanced F1-score, l'Accuracy e il Categorical WindowDiff, che forniscono misure complete della qualità della segmentazione tenendo conto dello squilibrio delle etichette. Inoltre, sono riportati e descritti due diversi confronti: il primo con ITALIAN XXL BERT per comprendere le differenze e sottolineare l'importanza dell'addestramento specifico per il contesto. Il secondo confronto è fatto con un motore basato su regole, per mostrare sia i vantaggi, sia i limiti di questo approccio, che si basa su template predefiniti e regole linguistiche ed euristiche per identificare le linee di interruzione. Nei primi anni di sviluppo di metodi e tecniche per la Segmentazione del Testo, la metrica di valutazione P_k,*Statistical Models for Text Segmentation* [79], è stato considerata uno standard per la valutazione degli algoritmi di segmentazione

del testo. Tuttavia, un esame teorico ha rivelato diverse preoccupazioni: P_k enfatizza più la penalizzazione dei falsi negativi rispetto ai falsi positivi, penalizza eccessivamente i quasi-abbattimenti e è influenzato dalle variazioni nella distribuzione delle dimensioni dei segmenti. Di conseguenza, è stata proposta la metrica WindowDiff [80], che utilizza una finestra mobile di dimensioni fisse attraverso il testo e penalizza l'algoritmo quando il numero di confini all'interno della finestra non corrisponde al vero numero di confini per quella finestra di testo. Tuttavia, WindowDiff considera solo la sequenza di segmentazione senza tener conto della classe del segmento identificato. In questo contesto, viene proposta la metrica Categorical WindowDiff, calcolando i valori WindowDiff individualmente per ciascun confine (tipo di segmento) e successivamente aggregati senza bisogno di operazioni supplementari. Nella fase di text pre-processing, al fine di preparare opportunamente le frasi in input e renderle adatte alla fase di apprendimento, oltre ad eliminare sorgenti di rumore come numeri di pagina, caratteri specifici come i caratteri di nuova riga e le barre, è stato aggiunta un'informazione supplementare sull'allineamento della riga della frase, considerando che l'aggiunta della funzionalità relativa all'allineamento del testo all'interno di un documento fosse cruciale per migliorare la comprensione del modello e delle informazioni spaziali presenti nel testo. La Figura 4.8 conferma le ipotesi fatte, mostrando la distribuzione degli allineamenti sulle linee dei breakpoint, che è fortemente inclinata verso l'allineamento "centrale". I due grafici a barre mostrano la distribuzione dell'allineamento del testo sulle righe breakpoint durante il training (a sinistra) e il test (a destra).

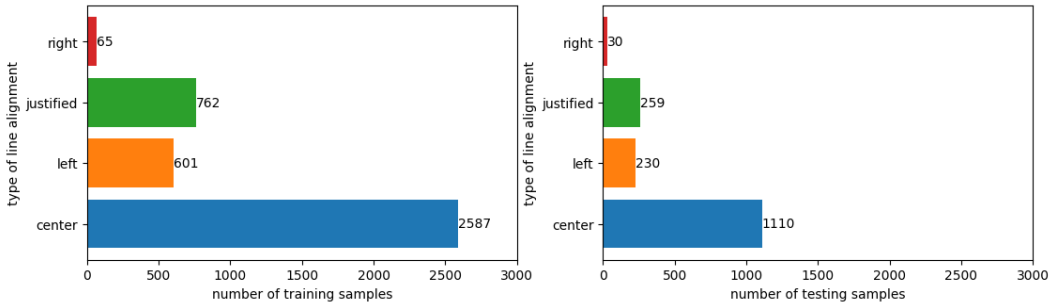
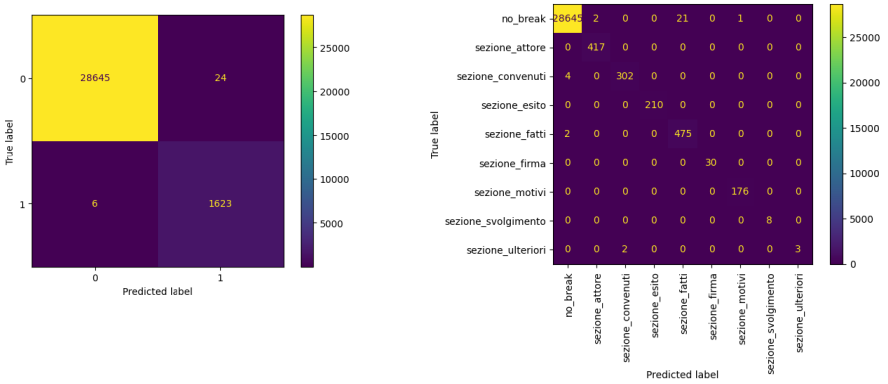


Figura 4.8: Legal Text Segmentation: distribuzione dell'allineamento del testo.

È stato effettuato un processo di "fine-tuning" del modello durato per ulteriori 10 epoche di addestramento, utilizzando un batch di 128 righe di testo su una singola NVIDIA Tesla T4 da 16 GB. È stato utilizzato l'"early stopping" con una "patience" di 3 sulla validazione della metrica di loss. Inoltre, è stato impiegato l'ottimizzatore AdamW con un tasso di decadimento di 0,8 ed è stato applicato un tasso di apprendimento di 4e-05 utilizzando uno scheduler di decadimento polinomiale.

Per la sperimentazione del modello proposto, sono stai definiti due diversi scenari. Nel primo, è stata valutata l'accuratezza della previsione dei breakpoint sulle righe del set di test, considerando quindi una classificazione binaria dove il modello predice solo se una linea è o meno un breakpoint. Le matrici di confusione,Figura 4.9, mostrano che il modello raggiunge prestazioni notevoli, con un punteggio F1 di 0.98 nella classificazione binaria dei breakpoint. È importante sottolineare anche che i test effettuati senza includere la feature dell'allineamento del testo nel pre-processing non hanno portato a risultati altrettanto positivi, con un punteggio F1 di circa 0.86.



(a) Confusion Matrix della breakpoint prediction.

(b) Confusion Matrix della breakpoint class prediction.

Figura 4.9: Legal Text Segmentation: Confusion Matrix

Successivamente, per il secondo scenario di sperimentazione è stato testato il modello sul task completo di segmentazione del testo, che considera sia la previsione delle righe breakpoint sia la previsione della classe (o sezione) a cui i breakpoint appartengono. I risultati mostrati in tabella Tabella 4.3 comprendono entrambi gli scenari e presentano gli F1-score sui due tipi di classificazione, binaria e multiclasse, e le metriche di WindowDiff categorico per ITALIAN XXL BERT, ITALIAN-LEGAL-BERT e il motore basato su regole.

Model	F1 Binary	F1 Multi - Macro avg	Categorical WindowDiff
ITALIAN XXL BERT	0.90	0.78	6.833
ITALIAN-LEGAL-BERT	0.98	0.96	0.324
Rule-Engine	0.99	0.99	0.002

Tabella 4.3: Legal Text Segmentation: confronto risultati su approcci diversi.

Durante lo sviluppo del modello, sono stati condotti inoltre alcuni studi per testare la robustezza dello stesso in presenza dati di test alterati. La resilienza e le prestazioni del modello sono state esaminate in relazione a variazioni nei dati di input, con l'obiettivo di valutare la sua capacità di generalizzazione e il suo comportamento in situazioni avverse o caratterizzate da rumore. Le perturbazioni introdotte nei dati di test sono state progettate

per simulare scenari realistici in cui la qualità dei dati potrebbe essere compromessa o influenzata da fattori disturbanti. Le perturbazioni includono principalmente alterazioni testuali che potrebbero derivare da un sistema OCR che identifica in modo errato i caratteri per lettere o numeri simili e nella sezione successiva che descrive il metodoREMOAC saranno approfonditi, appunto, i concetti di perturbazione e anomalie OCR. Per la sperimentazione sono stati applicati diversi tassi di perturbazione sui caratteri: 10%, 30%, 50% e 80%. La tabella mostra le metriche di accuratezza sui dati perturbati con un diverso tasso di perturbazione.

	ITALIAN-LEGAL-BERT		**RULE-ENGINE**	
Perturbation %	**F1 Binary**	**F1 Multi - Macro avg**	**F1 Binary**	**F1 Multi - Macro avg**
10%	0.77	0.70	0.26	0.32
30%	0.62	0.56	0.24	0.26
50%	0.55	0.52	0.24	0.27
80%	0.53	0.52	0.24	0.27

Tabella 4.4: Legal Text Segmentation: valutazioni performance del modello in presenza di perturbazioni testuali.

Come mostrato in Tabella 4.4, confrontando il modello proposto con il motore basato su regole, emerge un netto divario di prestazioni, specialmente nelle classificazioni multiclasse con l'aumentare del tasso di perturbazione. Il modello proposto mostra un'eccezionale capacità di adattamento a vari input testuali, riconoscendo e internalizzando schemi complessi nei dati, mantenendo buone prestazioni di generalizzazione anche con input alterati. Invece, il motore basato su regole, che usa schemi linguistici prestabiliti, non si adatta altrettanto bene alle dinamiche dei dati perturbati. La rigidità di questo sistema deriva dalla sua incapacità di apprendimento, contrariamente all'adattabilità mostrata dal nostro modello di deep learning.

4.3.4 Applicazione di tecniche di XAI

L'integrazione di modelli e tecniche XAI è stata attentamente progettata e personalizzata per supportare tre aspetti cruciali. In primis, la validazione

del processo di etichettatura, **Bias Labeling**, assicurando che le etichette assegnate ai dati fossero accurate e affidabili. In secondo luogo, la XAI si è concentrata sulla validazione della robustezza del modello, **Model Robustness**, per poi predisporre, in maniera piu generica, la spiegabilità per l'utente finale **Model Explainability**, assicurandosi che i risultati del modello siano comprensibili e trasparenti, rendendo chiaro il processo decisionale del modello agli utenti stessi.

Bias Labeling

Durante l'analisi dei risultati intermedi del modello, l'impiego della XAI ha evidenziato un significativo numero di errori di etichettatura fatti dagli esperti di dominio, che influenzavano negativamente le prestazioni del modello. Questa scoperta ha portato alla correzione dei dati sia nel set di addestramento che in quello di test, migliorando così l'accuratezza complessiva del modello. Questa azione ha evidenziato l'importanza di un'accurata verifica e revisione dei dati utilizzati in fase di addestramento e test per garantire l'efficacia e l'affidabilità del modello. Come si evince dalla Figura 4.10 il modello aveva predetto con particolare percentuale elevata che la frase fosse un breakpoint e appartenesse alla categoria "Label 4". La tabella di verità presente, invece, riportava come "label 0" (no breakpoint), la frase esaminata. Da una prima analisi è emerso che la problematica riguardava esclusivamente errori di predizioni nel caso in cui la tabella di verità riportava la frase come non breakpoint mentre il modello prediceva una label di tipo breakpoint. Sono stati quindi raccolti tutte le discordanze, calcolati gli shape values e costruiti una serie di grafici da sottoporre agli esperti di dominio per procedere ad una validazione del dataset. Lo stesso procedimento è stato iterato anche per il dataset di training.

Model Explainability

Per quanto concerne la spiegabilità del modello e la sua relativa analisi, il contributo principale di questo studio è stato l'identificazione di due

Figura 4.10: Legal Text Segmentation: esempio di bias labeling intercettato

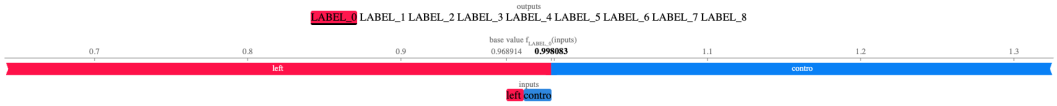


Figura 4.11: Legal Text Segmentation: esempio di interpretazione del modello

limiti principali nell'approccio utilizzato, emersi dall'analisi degli errori del modello. Il primo riguarda i casi in cui una frase non rappresenta un breakpoint, ma viene comunque prevista come tale. In queste situazioni, si osserva che il contesto testuale ha una maggiore influenza sulla caratteristica posizionale dell'allineamento del testo. Il secondo limite emerge quando il contesto testuale non è sufficiente a guidare la previsione. In questi casi, la caratteristica posizionale diventa determinante, sottolineando l'importanza delle informazioni spaziali nel testo. Infatti dalla Figura 4.10 si evince come la feature posizionale "left" concorri maggiormente a spostare la previsione del modello verso la "label 0" (no breakpoint), nonostante il modello abbia imparato che la parola "contro" ad inizio della frase rappresenti con molta probabilità una caratteristica discriminatoria per far si che la frase sia un breakpoint. Queste osservazioni evidenziano la complessità e la necessità di bilanciare diversi fattori nell'analisi del testo.

Model Robustness

Come descritto nella fase precedente, il presente approccio, in particolare rispetto ad un approccio a regole, vede nella robustezza uno dei suoi principali vantaggi. Per tali ragioni studi e analisi di tecniche di XAI si sono concentrate sull'emersione di tale proprietà e per capire la coerenza con i risultati ottenuti. Per procedere con tale analisi è stato necessario riproporre l'applicazione dei moduli XAI ad un dataset perturbato, ovvero , un dataset dove in maniera

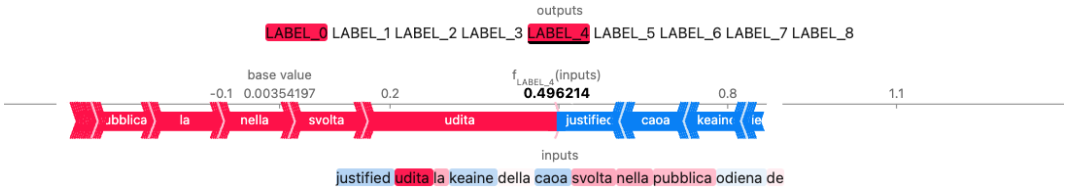


Figura 4.12: Legal Text Segmentation: esempio di Model Robustness

sintetica sono stati simulati alcuni tipi errori che possono commettere sistemi di OCR. Dall'analisi del set di dati perturbato è emerso che, nel caso di testi composti da una singola parola, la caratteristica posizionale è predominante e influisce significativamente sulla previsione. Al contrario, in testi formati da più parole, anche se le parole perturbate tendono a incrementare leggermente la probabilità della classe non-breakpoint, i token di parole non perturbate spesso prevalgono, contribuendo alla corretta classificazione. Questo fenomeno evidenzia come il contesto più ampio e la struttura del testo siano cruciali per l'interpretazione corretta.

4.3.5 Considerazioni

L'approccio proposto si è rilevato efficace per la segmentazione del testo nei documenti legali, dimostrando prestazioni eccezionali e contribuendo in modo innovativo al campo 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 classificazione testuale proposto, i risultati raggiunti si sono rilevati di particolare rilevanza. Infatti, la validazione del processo di etichettatura, Bias Labelling, 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, assicura che il modello sia resistente a varie condizioni e input imprevisti. Infine, ma non meno importante, l'applicazione della XAI per la Model Explainability per l'utente finale garantirà che i risultati siano comprensibili e trasparenti, illustrando chiaramente il processo decisionale del modello. Questi aspetti si sono rilevati cruciali per valutare l'efficacia e la fiducia nel modello.

4.4 REMOAC: A Retroactive Explainable Method for OCR Anomalies Correction

4.4.1 Introduzione

Proseguendo con la descrizione di scenari di applicazione di tecniche di XAI, come ultimo caso d'uso, troviamo un utilizzo diversificato ed innovativo delle tecnologie di XAI. Il contesto di riferimento è rappresentato dalla text classification correlato ai sistemi di Optical Character Recognition (OCR) per la digitalizzazione di documenti scansionati. Nei sistemi di Information Management, l'importanza di disporre di un eccellente strumento di OCR nell'era digitale risulta fondamentale, in particolare perché, nonostante l'ampia diffusione della digitalizzazione, molti settori continuano a fare affidamento su documenti cartacei. In diversi domini, come quelli legali, medici, educativi e governativi, i documenti vengono ancora spesso creati o archiviati in formato cartaceo. La digitalizzazione di questi documenti attraverso la scansione è un passo essenziale, ma non sufficiente per renderli pienamente funzionali in un ambiente digitale.

Per riuscire quindi ad utilizzare e valorizzare tale tipologia di documenti si rende necessario l'utilizzo di sistemi di OCR, una tecnologia che permette di convertire diversi tipi di documenti, come scansioni di testi stampati, manoscritti o immagini di testo, in dati testuali modificabili e ricercabili. Questo non solo facilita l'archiviazione e la gestione dei documenti, ma apre anche la strada all'analisi dei dati, all'accessibilità e alla condivisione efficiente delle informazioni. Un'elaborazione di alta qualità riduce errori e imprecisioni che possono verificarsi durante la conversione del testo, assicurando che i dati convertiti siano fedeli all'originale. Questo aspetto risulta cruciale, ad esempio, nei casi in cui la precisione dei dati può avere implicazioni legali o in ambiti dove gli errori possono influenzare significativamente l'interpretazione dei risultati, come nella ricerca medica.

I sistemi OCR però soffrono di alcune problematiche legate non solo

alla capacità e le performance dei sistemi in se ma dipendono anche molto dalla qualità dei documenti da digitalizzare. Infatti gli errori prodotti da uno strumento di OCR possono essere vari e influenzare significativamente la qualità del testo digitalizzato. Schematizzando le principali tipologie di anomalie (errori), troviamo:

- Errori di riconoscimento dei caratteri: questo è l'errore più comune e avviene quando l'OCR interpreta in modo errato i caratteri individuali. Ad esempio, può scambiare una "l" (elle minuscola) per un "1" (uno) o una "O" (o maiuscola) per uno "0" (zero). Questi errori sono frequenti con font poco chiari o in presenza di testo danneggiato.

- Errori di formattazione: l'OCR può avere difficoltà nel mantenere la formattazione originale del documento. Ciò include problemi nel riconoscere correttamente i margini, gli allineamenti, gli spazi tra paragrafi, l'indentazione, elenchi puntati o numerati, e la formattazione del testo come grassetto o corsivo.

- Errori di riconoscimento di layout: in documenti con layout complessi, come quelli con colonne multiple, immagini o grafici, l'OCR può non riuscire a interpretare correttamente l'ordine del testo o includere elementi non testuali nel testo estratto.

- Errori di parola intera: a volte, l'OCR può sbagliare nell'interpretazione di intere parole, specialmente se sono scritte con una calligrafia insolita, se sono parzialmente oscurate o danneggiate.

- Errori di linguaggio e contesto: gli strumenti OCR possono avere difficoltà con le lingue che non sono state adeguatamente incorporate nel loro algoritmo o con termini tecnici e di nicchia. Inoltre, possono mancare di contesto, portando a errori nell'interpretazione di parole che si scrivono allo stesso modo ma hanno significati diversi a seconda del contesto (omografi).

- Errori dovuti alla qualità dell'immagine: scansioni di bassa qualità, immagini sfocate, o con contrasto insufficiente possono portare a un aumento degli errori di riconoscimento da parte dell'OCR.

- Errori nei documenti antichi o danneggiati: documenti con inchiostro sbiadito, macchie, strappi o pieghe possono presentare sfide significative per l'OCR, risultando in un alto tasso di errori.

Ridurre questi errori richiede, oltre l'utilizzo di software OCR avanzati e scansioni di alta qualità dei documenti anche strumenti di post elaborazioni che migliorino la qualità del testo prodotto. Tra questi strumenti di post elaborazione troviamo quelli che si occupano di correzione ortografica, o spell checking. Dopo che un documento è stato scansionato e il testo è stato estratto tramite OCR, lo strumento di spell checking analizza il testo per identificare parole che sembrano essere scritte in modo errato. Quando viene rilevato un potenziale errore, lo strumento di spell checking propone correzioni, spesso basate su algoritmi che considerano la somiglianza tra le parole e le possibili varianti ortografiche basandosi ad esempio su una distanza di edit minima (come il numero di lettere che devono essere cambiate, aggiunte o rimosse per ottenere una parola corretta). Alcuni strumenti di spell checking avanzati, invece, utilizzano l'intelligenza artificiale e l'apprendimento automatico per capire meglio il contesto in cui una parola viene utilizzata, permettendo loro di proporre suggerimenti più accurati, specialmente in testi con terminologia specifica di un determinato dominio. Qualsiasi sia l'approccio da utilizzare, gli strumenti di spell checking possono essere integrati con dizionari specifici (ad esempio, terminologia legale, medica, scientifica) per migliorare la precisione. Costruire uno strumento di spell checking, perimetrando dominio ed applicazione, rappresenta la base dell'intuizione proposta da REMOAC e la relativa applicazione delle tecniche di XAI nell'ambito della correzione delle anomalie prodotte dai sistemi di OCR. Tale metodo si concentra, infatti, sulla correzione delle anomalie OCR nel contesto specifico della classificazione del testo, ovvero, si pone l'obiettivo di

correggere parte del testo per migliorare le performance di un text classfier AI based. REMOAC, utilizza modelli e tecniche di spiegabilità per individuare specifici problemi all'interno di una frase e una volta identificate queste anomalie, interviene applicando correzioni mirate. L'aspetto distintivo di questo approccio risiede nella sua capacità di eseguire una correzione selettiva dei token. Invece di correggere indiscriminatamente tutto il testo, REMOAC si concentra esclusivamente su quei token che presentano errori e che, se non corretti, potrebbero compromettere l'efficacia del classificatore. Questa metodologia mirata, non solo ottimizza il processo di correzione, ma potenzia significativamente le prestazioni di un classificatore di testo, garantendo l'ottenimento di risultati accurati anche in contesti applicativi reali e complessi.

4.4.2 Il metodo REMOAC

Come anticipato nel paragrafo introduttivo REMOAC [81], vuole correggere puntualmente e perimetralmente le anomalie prodotte dai sistemi di OCR per migliorare le performance di un task specifico che è quello della Text Classification. Resta quindi preliminare la presenza di un modello di Text Classification, machine learning based, sul quale applicare il metodo REMOAC per aumentarne le performance. Per perseguire tale obiettivo, REMOAC si compone di due moduli differenti che si occupano delle due fasi distinte relative alla gestione delle anomalie : OCR Anomaly Detection e OCR Anomaly Correction.

OCR Anomaly Detection

Lo stato dell'arte presenta diversi lavori riguardo l'utilizzo di un approccio di anomaly detection per individuare e correggere anomalie nei testi [82, 83]. Stabilire se una data frase, in un testo, contiene anomalie o errori e determinarne la posizione all'interno della frase è il primo passo da effettuare nel processo di correzione delle anomalie OCR. Tale tipologia di task, non sempre si presenta come un'attività semplice da sviluppare, dal momento che la base di questi errori è rappresentato da parole ripetute e mancanti che

complica, di fatto, il task stesso[84]. A volte, ciò che potrebbe sembrare un'anomalia in un testo OCR non è effettivamente un errore, ma può derivare da diverse cause come la diversità nelle convenzioni di scrittura e formattazione adottate.[85].

Dopo aver verificato la presenza di anomalie in una frase, si procede alla classificazione di queste anomalie secondo alcune tipologie di tassonomie e attraverso diversi framework che utilizzano approcci diversificati [86].

Durante gli anni i task di anomaly detection per dati testuali, ma in generale tutti i task Natural Language Processing based, hanno subito una rivoluzione con lo sviluppo di BERT (Bidirectional Encoder Representations from Transformers)[73], grazie al quale sono stati introdotti gli embedding contestuali, costruiti sulla base del contesto in cui appare una parola[87]. A differenza del tradizionale pre-addestramento da sinistra a destra, BERT è inteso per essere utilizzato per un pre-addestramento bidirezionale e di conseguenza, può essere utilizzato per numerosi compiti NLP. Superando i predecessori come ELMo [88] or Flair [89], ha avanzato lo stato dell'arte del pre-addestramento ed è diventato la soluzione all'avanguardia de facto.

Considerando, quindi, gli errori promossi da un sistema di OCR come anomalie testuali, il primo passo dell'approccio proposto è classificare una frase per rilevare se è stata alterata o meno. Consideriamo un dataset $D = l_1, l_2, \ldots, l_K$ costituito da K frasi. Ogni frase è associata a un'etichetta di classificazione y che può essere 0 o 1, rappresentando rispettivamente frasi non alterate e alterate. Tale dataset deve essere dato in input ad un classificatore testuale basato su Transformer e su modello BERT. In particolare, l'architettura di BERT per la classificazione di frasi è basata su un modello di rete neurale Transformer che apprende rappresentazioni di testo bidirezionali, analizzando il contesto di ogni parola in una frase sia dalla parte sinistra che destra. Nella classificazione di frasi, BERT considera l'intera sequenza di input (ad esempio, una frase o un paragrafo) per prevedere l'etichetta di output. Viene utilizzato un token speciale [CLS] all'inizio della sequenza di input, la cui rappresentazione finale viene utilizzata per la classificazione. Le

rappresentazioni di parole generate da BERT vengono poi passate attraverso uno o più strati lineari con una funzione di attivazione softmax per ottenere le probabilità delle classi. Nel modello proposto, tale classificazione si traduce in una classificazione binaria, che ha l'obiettivo di determinare se una frase contiene o meno anomalie testuali.

OCR Anomaly Correction

Come anticipato precedentemente, la strategia proposta da REMOAC, mira a migliorare le prestazioni di un classificatore testuale, machine learning based, rilevando e correggendo eventuali errori introdotti da un modulo OCR durante la scansione delle pagine, utilizzando tecniche di explainable AI. L'idea di base è che, se una frase viene "alterata" dagli errori OCR, rendendo difficile la corretta classificazione, correggendo il token errato, la previsione potrebbe diventare corretta. L'approccio si divide in due fasi: una prima fase "preliminare", da effettuare all'inizio o solo in caso di modifica (riaddestramento) del modello di Text Classification, dove si costruisce il dizionario dei token più rappresentativi del classificatore testuale da migliorare, ed una fase "attiva" che prevede prima una fase di detection per rilevare le anomalie ed in caso positivo applicare un flusso correttivo, il tutto sempre basato su risultati proveniente dall'applicazione di tecniche di Explainable Ai. Vediamo di seguito in dettaglio le due fasi.

Fase Preliminare: Nella fase preliminare (Preliminary Phase), in figura 4.13, si analizza il classificatore testuale (Main classifier Model), precedentemente addestrato, oggetto del miglioramento, applicando SHAP (SHapley Additive exPlanations) [51], un framework per interpretare le previsioni e identificare le misurazioni relativi alla feature importance del modello. Lo scopo dell'applicazione di SHAP è di identificare le feature (nel caso del testo le parole o token) che contribuiscono in maniera più elevata alle predizioni del classificatore. In particolare, per tutte le etichette di classificazione previste che sono predette dal classificatore oggetto del miglioramento, si estraggono le prime N Parole (words) (NW), che in media, hanno contribuito maggiormente

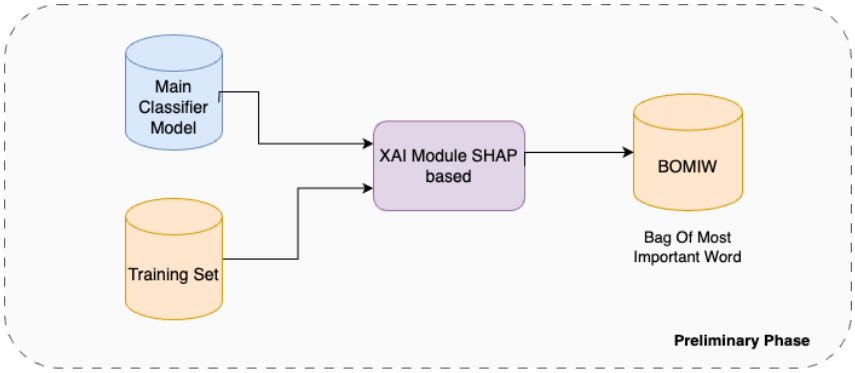


Figura 4.13: REMOAC: Fase preliminare

ad orientare, correttamente, il classificatore sulla specifica predizione. Tale estrazione si basa sulla valutazione complessiva di tutti gli shape values per ogni item (frase) previsto per la fase di training del classificatore testuale. Così facendo, vengono estratte le prime N parole più rappresentative per ogni etichetta analizzata. Le parole uniche estratte andranno a costituire il nostro BOMIW (Bag of Most Important Words), un dizionario ridotto di parole significative per un dominio e task specifici, con l'obiettivo, appunto, di perimetrare e ottimizzare le possibili sostituzioni di anomalie presenti nei testi.

Fase Attiva:

Nella fase attiva si applica il metodo REMOAC ad una singola frase per correggere le anomalie e procedere poi con la classificazione (Main Classifier Model). Come descritto in figura 4.14, la fase attiva è divisa in 3 step distinti:

- **Step 1 - Anomaly Detection e Most Important Error Token**
  Nel primo step la frase da dare in pasto al classificatore testuale, viene analizzata dal modulo di OCR Anomaly Detection che classifica la frase stessa come frase contenente o meno anomalie (errori) al suo interno. Se il modulo di anomaly detection ritiene che nella frase non ci siano errori, il workflow REMOAC si interrompe e restituisce la frase nella sua versione originale. Diversamente se la frase è classificata come "anomala" attraverso l'applicazione di un modulo XAI basato su SHAP, che in ingresso ha sia il OCR Anomaly Detection Model che la frase esaminata,

si estrae il token (la parola), il MIET (Most Important Error Token), che in percentuale maggiore ha contribuito proprio a classificare come 'anomala" la frase esaminata dal OCR Anomaly detection classifier.

- **Step 2 - Token Similarity**

  Nel secondo step, il MIET viene confrontato con tutte le istanze della BOMIW attraverso la misurazione della distanza fra stringhe basata su due misure differenti Longest Contiguous matching Subsequence (LCS) [90] e Levenshtein Distance [91]. Una volta selezionata la parola con la probabilità più alta di similarità, se tale probabilità supera la soglia impostata, viene identificata come una parola "cleaned".

- **Step 3 - Cleaner**

  Nello terzo step la parola MIET presente nella frase viene sostituita in tutte le sue occorrenze con la parola "cleaned" presente nella BOMIW. Viene cosi restituita da REMOAC la frase corretta puntalmente solo per quanto concerne il MIET, ovvero la parola che secondo il classificatore di OCR Anomalies era quella con più probabilità a rappresentare un errore.

4.4.3 Sperimentazione e risultati

Per la sperimentazione del metodo REMOAC è stato selezionato il caso d'uso precedentemente illustrato inerente la Legal Text Segmentation (LTS) in ambito legale, un modello che identifica e classifica sequenzialmente le frasi che rappresentano breakpoint semantici di documenti legali. Essendo un caso d'uso di text classification e riferendosi ad un dominio specifico, di fatto quello legale, la LTS rappresentava un ideale caso di test per REMOAC. Inoltre, come descritto nel paragrafo che descriveva la robustezza del modello nel capitolo precedente, seppur rispetto ad un approccio a regole, risultava più robusto rispetto alle anomalie presenti nel testo, le performance degradavano al degradare della qualità del testo prodotto dalla OCR.

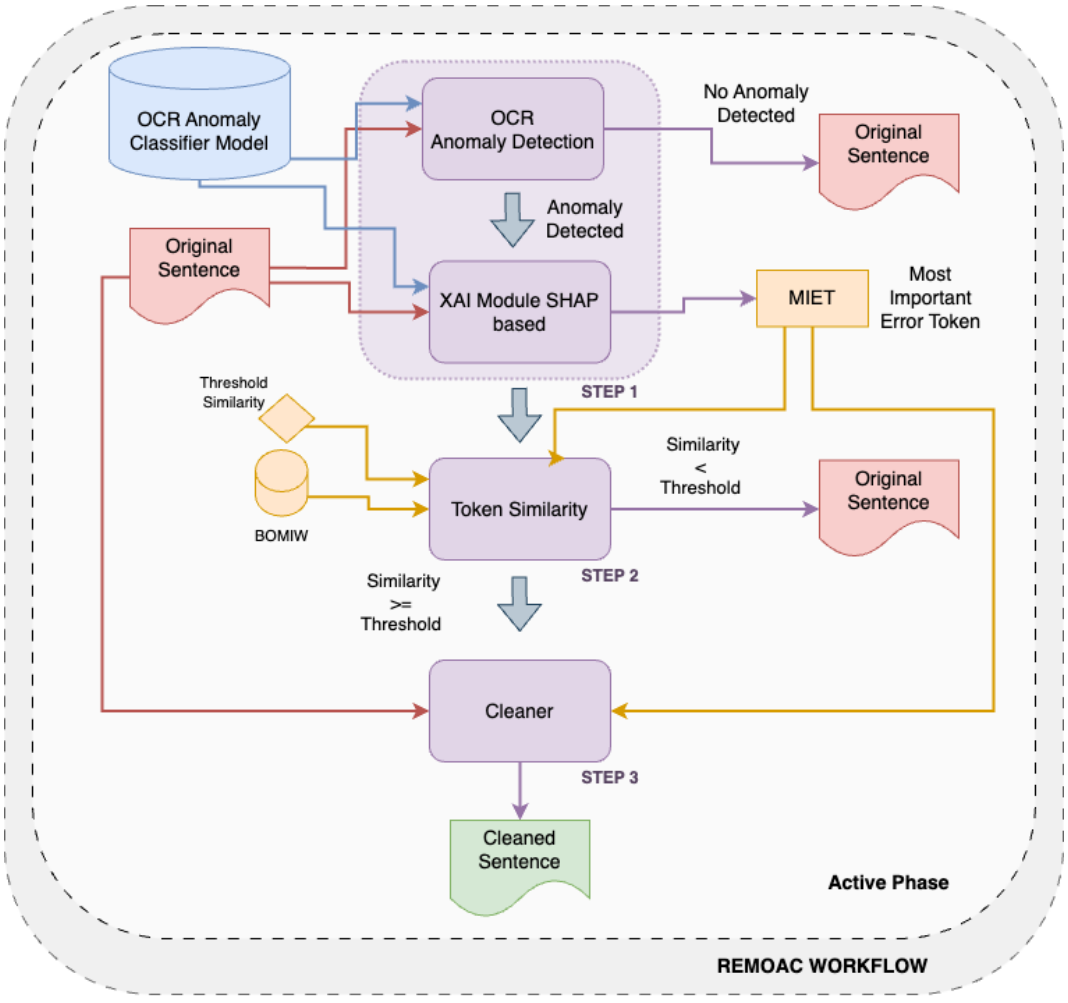


Figura 4.14: REMOAC: Fase Attiva

Main Text Classifier

Il modello, precedentemente illustrato di LTS è basato su un'architettura transformer e su ITALIAN-LEGAL-BERT, ed è riconducibile ad un problema di multi label text classification, dove ad ogni linea è associata un etichetta che lo caratterizza come appartenne ad una classe specifica, dalla "label 0" che lo identifica come "no breakpoint" ad una possibile classe di "break point" sulla base del tipo di sezione che introduce. Nonostante dallo studio precedente, durante la fase sperimentale, era emersa una sua particolare caratteristica di robustezza, che lo confrontava positivamente rispetto ad un approccio a regole, le anomalie prodotte da sistemi OCR impattano, comunque, con una certa percentuale, sulle performance del modello. Questi aspetti lo rendono dunque un caso d'uso ideale per validare l'approccio REMOAC.

Dataset

Il dataset per la sperimentazione di REMOAC rappresenta un ulteriore elaborazione del dataset previsto per la LTS. Si parte sempre dal set di documenti della Corte di Cassazione in particolare 730 (520 per l'addestramento e 210 per il test) acquisiti da diverse fonti (*ItalgiureWeb*, *Ricerca Giuridica* e *La Legge per tutti*) per massimizzare la diversità dei dati e ottenere campioni rappresentativi. Come per la Legal Text Segmentation, i testi estratti sono sempre organizzati in singole righe per ogni documento, in modo che l'input del modello fosse una singola frase. Per l'addestramento e la valutazione di REMOAC è stato riprocessato il dataset della LTS creandone una sua copia speculare contenente però in modalità sintetica anomalie. In particolare, ogni singola frase del dataset duplicato è stata sottoposta ad un processo di perturbazione, rispecchiando le anomalie che comunemente si manifestano durante il parsing dei testi da parte di un sistema OCR. Questo processo è utile per simulare gli scenari reali in cui i professionisti legali del testo lavorano spesso con documenti testuali digitalizzati, molti dei quali vengono elaborati con l'OCR. Il processo di perturbazione ha introdotto variazioni quali anomalie nel riconoscimento dei caratteri, discrepanze nella formattazione e

occasionali interpretazioni errate che riproducono fedelmente le sfide incontrate nelle applicazioni OCR. Il grado di perturbazione applicato alle frasi è minimo; pertanto, solo il 10% dei caratteri di ogni frase è stato perturbato [92]. Questa scelta è motivata dal fatto che addestrare un modello a rilevare un grado di perturbazione minimo lo rende in grado di rilevare tassi di perturbazione più elevati. Il dataset in definitiva comprende quindi il doppio delle righe del dataset di LTS perfettamente bilanciato su due tipi distinti di etichette: "1" per le righe perturbate e "0" per quelle non perturbate, conservando però l'etichetta del task principale, ovvero se rappresenta o meno un breakpoint. Quindi per ogni frase estratta avremo sia una tabella di verità riguardante il problema di text segmentation (se sia o meno un breakpoint) sia se la frase sia stata perturbata o meno e che quindi contenga delle anomalie.

Algoritmi per lo spell checking in lingua italiana

È stato inoltre condotto uno studio dello stato dell'arte sulla disponibilità di librerie open source per lo spell checking nella lingua italiana, allo scopo di confrontare l'approccio REMOAC con altre tipologie di sistemi per l'individuazione e la correzione di anomalie. Tale studio ha comportato la selezione delle seguenti librerie disponibili.

- **Enchant**

  La libreria Enchant [93] di Dom Lachowicz è una libreria software per la correzione ortografica. È progettata per fornire un'interfaccia semplice e unificata a diverse librerie di correzione ortografica esistenti, come Aspell, MySpell, Hunspell, e altri. Enchant è ampiamente utilizzata per la sua compatibilità con vari formati di dizionari e il suo approccio flessibile alla gestione degli errori ortografici. Questa caratteristica la rende una scelta popolare per gli sviluppatori che necessitano di implementare funzionalità di controllo ortografico in diverse applicazioni. Per la sperimentazione è stato utilizzata un'implementazione nel linguaggio python di Enchant: PyEnchant [94]

- **Autocorrect**

  Autocorrect [95] è una libreria python per lo spell checking multi language e supporta la lingua italiana. L'implementazione è basata sul più noto correttore di Peter Norvig [96], esempio noto di implementazione di correttore ortografico. Infatti questo correttore è noto per la sua brevità e efficienza, dimostrando come sia possibile implementare un correttore ortografico funzionale con meno di 25 linee di codice. L'algoritmo si basa sull'idea che la parola corretta è spesso a una distanza di edit minima dalla parola errata, utilizzando operazioni come l'inserimento, la cancellazione, la sostituzione o il trasposto di lettere. Questo approccio è diventato un classico esempio nell'ambito dell'elaborazione del linguaggio naturale e dell'apprendimento automatico.

- **Spell Checker - Spark NLP**

  Spell Checker [97] è un modello sequence-to-sequence di John Snow Labs (Context Spell Checker) che rileva e corregge gli errori ortografici nel testo in ingresso in lingua italiana. È basato sull'automa di Levenshtein per generare le correzioni candidate e su un modello linguistico neurale per classificare le correzioni. Il modello è addestrato con PySpark 2.4.x e SparkNLP 3.4.1.

Applicazione di REMOAC

Considerati i dataset precedentemente descritti, REMOAC è stato testato in due diversi scenari variando gli iperparametri che permette di configurare ovvero:

- **N Words** : nella costruzione della **BOMIW** (Bag of Most Important Words) rappresenta il numero di parole ritenute importanti per la previsione di ogni etichetta del classificatore principale.

- **Similarity threshold**: rappresenta la soglia di similarità minima da considerare nel confronto tra il MIET (Most Important Error Token) estratto da una frase considerata come avente anomalie e le parole presenti nella BOMIW.

- **Algoritmo di string distance**:LCS [90] e distanza di Levenshtein [91].

Nel primo scenario di sperimentazione è stato valutato l'applicazione di REMOAC sul caso singolo per verificare, in prima istanza, i benefici apportati dal metodo. Come parametri selezionati, i migliori risultati si sono avuti considerando rispettivamente:

- N Words : 10.

- Similarity threshold: 0.80

- Algoritmo di string distance:distanza di Levenshtein [91] (scelta ininfluente).

Come si evince dalla tabella Tabella 4.5 che mostra il confronto dei risultati del classificatore principale con e senza REMOAC, si è voluto valutare anche l'impatto dell'eventuale errore commesso, nello step 1 della fase attiva, dall'Anomaly OCR Detection classifier, considerando per quel ciclo di test tutte le sentenze come perturbate e quindi applicando tutti gli step del metodo REMOAC.

Model	Balanced Accuracy	F1 Binary	F1 Multi - Macro avg
Legal Text Classifier without REMOAC	0.82	0.77	0.70
LTC - REMOAC w/ OCR Anomaly Detection prediction	0.94	0.93	0.91
LTC - REMOAC w/o OCR Anomaly Detection prediction	0.94	0.94	0.91

Tabella 4.5: REMOAC: Scenario 1 - risultati

Valutata positivamente la validità dell'approccio REMOAC è stato considerato un secondo scenario di sperimentazione che metteva a confronto REMAOC stesso con diversi approcci di spell checking, presenti allo stato dell'arte, per la lingua italiana. In questo caso sono stato considerati e creati

diversi dataset aumentando la percentuale di perturbazione e generazione anomalie.

Come parametri selezionati, i migliori risultati per REMOAC si sono avuti considerando rispettivamente:

- N Words : 10.
- Similarity threshold: per i dataset con percentuale di perturbazione 10% e 30% è stata usato una soglia di 0.80. Per i dataset con perturbazione pari al 50% e 80% una soglia dello 0.1.
- Algoritmo di string distance:distanza di Levenshtein [91] (scelta ininfluente).

I risultati della sperimentazione del secondo scenario sono presenti nelle tabelle successive:

10 % PERTUBATION				
Model	**Accuracy**	**Balanced Accuracy**	**F1 Binary**	**F1 Multi - Macro avg**
REMOAC	**0.993498**	**0.944744**	**0.93639**	**0.912441**
Autocorrect	0.990824	0.918725	0.90758	0.880623
Enchant	0.989867	0.910692	0.897152	0.802907
SPARK NLP	0.980296	0.818206	0.776488	0.702643

Tabella 4.6: REMOAC: Scenario 2 con dataset pertubato al 10%

30 % PERTUBATION				
Model	**Accuracy**	**Balanced Accuracy**	**F1 Binary**	**F1 Multi - Macro avg**
REMOAC	**0.984553**	**0.861565**	**0.834395**	**0.857867**
Autocorrect	0.97046	0.728187	0.624423	0.620843
Enchant	0.975048	0.770561	0.7	0.657486

SPARK NLP	0.964123	0.666939	0.50023	0.589146

Tabella 4.7: REMOAC: Scenario 2 con dataset pertubato al 30%

50 % PERTUBATION				
Model	**Accuracy**	**Balanced Accuracy**	**F1 Binary**	**F1 Multi - Macro avg**
REMOAC	**0.970592**	**0.729993**	**0.627353**	**0.592732**
Autocorrect	0.964783	0.675683	0.517849	0.5875
Enchant	0.970427	0.729038	0.625105	0.591754
SPARK NLP	0.960459	0.633158	0.420136	0.526883

Tabella 4.8: REMOAC: Scenario 2 con dataset pertubato al 50%

80 % PERTUBATION				
Model	**Accuracy**	**Balanced Accuracy**	**F1 Binary**	**F1 Multi - Macro avg**
REMOAC	**0.9765**	**0.802305**	**0.735316**	**0.758161**
Autocorrect	0.964354	0.672272	0.509982	0.605107
Enchant	0.971549	0.738605	0.643507	0.623576
SPARK NLP	0.960261	0.631317	0.415534	0.528105

Tabella 4.9: REMOAC: Scenario 2 con dataset pertubato al 80%

4.4.4 Considerazioni

In questo scenario di applicazione delle tecniche di XAI, è stata affrontata la sfida critica del rilevamento e della correzione delle anomalie di OCR nel dominio della classificazione dei testi. Utilizzando metodi di XAI (Explainable Artificial Intelligence), è stato costruito REMOAC, un metodo retroattivo

spiegabile per la correzione di anomalie OCR, pensato per affrontare le sfide complesse poste dalle anomalie OCR in questo dominio specializzato, concentrando la correzione solo sui token che potrebbero influenzare il modello di classificazione. E' stato dimostrato che il metodo può migliorare le prestazioni del task di classificazione in presenza di anomalie provenienti da sistemi di OCR. L'intuizione proposta si basa sull'utilizzo dei valori SHAP del classificatore di testo in una modalità attiva, che va oltre la comprensione del funzionamento del classificatore. L'idea è stata quella di costruire un dizionario limitato di termini, ritenuti più influenti nelle decisioni del classificatore testuale, cosi da permimetrare e indirizzare il correttore ortografico solo su possibili scelte proprie del dominio e ritenute importanti ai fini del task di classificazione. Infatti l'obiettivo di base non è correggere l'intero testo, ma limitare tale correzioni ai token che l'OCR anomaly detection ritiene quello con probabilità maggiormente anomalo e promuovere un ventaglio di possibili soluzioni che non comprenda l'intero dizionario della lingua italiana o un dizionario di dominio esteso, ma un set parole (bag of words) caratterizzanti sia il dominio sia il classificatore testuale precedentemente addestrato. In alternativa alle precedenti applicazioni che utilizzavano la XAI orientandola a diversi stakeholder e a diversi obiettivi uno dei quelli anche al miglioramento stesso del task preso in esame, con REMOAC si è voluto dimostrare che la conoscenza prodotta dalla XAI può essere utilizzata direttamente ed in moto automatizzato per migliorare le performance dello stesso modello di Machine Learning che si è voluto esaminare per capirne il funzionamento. Oltre ad estendere l'approccio retroattivo della XAI su altri casi d'uso inerenti la text classification, una direzione promettente per ricerche future è rappresentato dalla sperimentazione di REMOAC sia su altri contesti linguistici e in differenti task di NLP come la Named Entity Recognition, Topic Detection e Keyword Extraction. Parallelamente, non meno importante, sarà estendere la sperimentazione di REMOAC per includere lingue oltre l'italiano considerando altri domini per dimostrare il potenziale agnostico del metodo alla lingua.

CAPITOLO 5

CONCLUSIONI

Nel presente lavoro, è stata affrontata la capacità delle tecniche di XAI di soddisfare diversi requisiti per un'AI affidabile nel contesto dei DSS. Lo studio e analisi dello stato dell'arte sia per quanto concerne la letteratura che gli strumenti disponibili, ha evidenziato una grande eterogeneità sia nella disponibilità di tecniche e strumenti, sia nella costruzione di un linguaggio comune che potesse adeguatamente favorire e trasferire l'importanza dell'introduzione della XAI nei processi aziendali. Tale complessità ed eterogeneità può risultare un forte ostacolo per l'adozione di questi strumenti, in particolare nei contesti enterprise, che richiedono, quasi sempre, approcci metodologici fortemente chiari, contestuali e composti da procedure ordinati. È nata cosi l intuizione di razionalizzare e sintetizzare 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 voleva porre l'attenzione sulla complessità e sull'importanza dei principi di un AI responsabile, trasferendo la necessita 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. Come riportato in [98] lo human-centered design mira a "consentire agli utenti di raggiungere obiettivi in modo efficace, efficiente e soddisfacente, tenendo conto del contesto di utilizzo". Nel contesto dei DSS, un design centrato sull'uomo della loro interfaccia utente aumenta sicuramente la capacità di tutti gli attori del processo di comprendere e agire correttamente basandosi sulle proposizione offerte dai sistemi, migliorando di fatto le decisioni e quindi le performance aziendali.

La sperimentazione e l'applicazione degli strumenti nei diversi casi d'uso individuati ha evidenziato 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.

5.1 Sviluppi futuri

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 metodologia 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. Questi strumenti, che favoriranno e guideranno ancor di più l'applicazione del

framework, dovranno poi essere contemplati ed organizzati in un quadro più vasto e complesso, una “reference architecture” che fornirà una linea guida, un modello per la progettazione e l’implementazione dei sistemi di XAI. Il modello architetturale dovrà essere pensato e progettato per essere personalizzato e adattato alle esigenze specifiche di un progetto o di un’organizzazione, rappresentando, di fatto, un punto di partenza comune comprensivo di best practices per la progettazione di sistemi. In conclusione, come evidenziato nelle considerazioni di REMOAC descritte nel paragrafo 4.4.4, 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.

BIBLIOGRAFIA

[1] Daniel Gruen Q. Vera Liao e Sarah Miller. «Questioning the AI: Informing Design Practices for Explainable AI User Experiences». In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020).

[2] Gagan Bansal et al. «Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance.» In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021).

[3] Upol Ehsan et al. «Expanding Explainability: Towards Social Transparency in AI Systems.» In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021).

[4] Niall Docherty e Asia J. Biega. «(Re)Politicizing Digital Well-Being: Beyond User Engagements.» In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022).

[5] Shivani Kapania et al. «Because AI is 100% Right and Safe: User Attitudes and Sources of AI Authority in India.» In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022).

[6] Ge Wang et al. «Informing Age-Appropriate AI: Examining Principles and Practices of AI for Children.» In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022).

[7] Hao-Fei Cheng et al. «Soliciting Stakeholders' Fairness Notions in Child Maltreatment Predictive Systems.» In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021).

[8] Min Kyung Lee e Katherine Rich. «Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust.» In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021).

[9] Jakub Mlynar et al. «AI beyond Deus Ex Machina – Reimagining Intelligence in Future Cities with Urban Experts.» In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022).

[10] Wencan Zhang, Mariella Dimiccoli e Brian Y Lim. «Debiased-CAM to Mitigate Image Perturbations with Faithful Visual Explanations of Machine Learning.» In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022).

[11] Hyanghee Park et al. «Human-AI Interaction in Human Resource Management: Understanding Why Employees Resist Algorithmic Evaluation at Workplaces and How to Mitigate Burdens.» In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022).

[12] Nur Yildirim et al. «Because AI is 100% Right and Safe: User Attitudes and Sources of AI Authority in India.» In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022).

[13] Hariharan Subramonyam et al. «Solving Separation-of-Concerns Problems in Collaborative Design of Human-AI Systems through Leaky Abstractions.» In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022).

[14] Kaely Hall et al. «Supporting the Contact Tracing Process with WiFi Location Data: Opportunities and Challenges.» In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022).

[15] Ramya Akula e Ivan Garibay. «Ethical AI for Social Good.» In: *International Conference on Human-Computer Interaction (HCII 2021)* (2021).

[16] Steve J. Bickley e Benno Torgler. «Cognitive architectures for artificial intelligence ethics.» In: *AI & SOCIETY (June 2022) 1-19.* (2022).

[17] Brandon M. Booth et al. «Bias and Fairness in Multimodal Machine Learning: A Case Study of Automated Video Interviews.» In: *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21). Association for Computing Machinery, New York, NY, USA, 268–277.* (2021).

[18] Brandon M. Booth et al. «Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data.» In: *ACM Hum.-Comput. Interact. 5, CSCW2, Article 438 (October 2021), 27 pages.* (2021).

[19] Andreas Cebulla et al. «Applying ethics to AI in the workplace: the design of a scorecard for Australian workplace health and safety.» In: *AI & society (May 2022)* (2022).

[20] Santtu Lehtinen Robert Gianni e Mika Niemine. «Governance of Responsible AI: From Ethical Guidelines to Cooperative Policies.» In: *Frontiers in Com- puter Science 4, Article 873437 (May 2022), 17 pages* (2022).

[21] Birte Keller Kimon Kieslich e Christopher Starke. «Artifcial intelligence ethics by design. Evaluating public perception on the importance of ethical design principles of artifcial intelligence.» In: *Big Data & Society 9, 1 (Jan. 2022)* (2022).

[22] Nuria Oliver Bruno Lepri e Alex Pentland. «Ethical machines: The human-centric use of artifcial intelligence.» In: *Science 24, 3, Article 102249 (March 2021) 17 pages.)* (2022).

[23] Pradeep Paraman e Sanmugam Anamalah. «thical artifcial intelligence framework for a good AI society: principles, opportunities and perils.» In: *AI & Society (June 2022), 1-17* (2022).

[24] Jan Auernhammer. «Human-centered AI: The role of Human-centered Design Research Human-centered AI: The role of Human-centered Design Research in the development of AI». In: *Boess, S., Cheung, M. and Cain, R. (eds.), Syner gy - DRS International Conf erence 2020 , 11-14 August* (2020).

[25] Mustafa Demir Nancy Cooke e Lixiao Huang. «A framework for human autonomy team research». In: *International Conference on Human-Computer Interaction (HCII '20). Springer, Cham, 134-146* (2020).

[26] Andreas Holzinger et al. «Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artifcial intelligence.» In: *Inf. Fusion 79, C (Mar 2022), 263–278* (2022).

[27] M. Turek. *Explainable artificial intelligence (xai).* https://www.darpa.mil/program/explainable-artificial-intelligence. 2018.

[28] High-Level Expert Group on AI. *Explainable artificial intelligence (xai).* https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. 2019.

[29] Jan A. Kors Aniek F. Markus e Peter R. Rijnbeek. «The role of explainability in creating trustwor- thy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies.» In: *Journal of Biomedical Informatics, 113:10365* (2021).

[30] Zeynep Akata et al. «A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence.» In: *Computer, 53(8):18–2* (2020).

[31] Supriyo Chakraborty et al. «Interpretability of deep learning models: A survey of results». In: *2017 IEEE SmartWorld, Ubiquitous Intelligence and Computing, Advanced and Trusted Computed, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. 2017, pp. 1–6. DOI: 10.1109/UIC-ATC.2017.8397411.

[32] Angelos Chatzimparmpas et al. «A survey of surveys on the use of visualization for interpreting machine learning models». In: *Information Visualization* 19.3 (2020), pp. 207–233. DOI: 10.1177/1473871620904671. eprint: https://doi.org/10.1177/1473871620904671. URL: https://doi.org/10.1177/1473871620904671.

[33] Alejandro Barredo Arrieta et al. «Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI». In: *Information Fusion* 58 (2020), pp. 82–115. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2019.12.012. URL: https://www.sciencedirect.com/science/article/pii/S1566253519308103.

[34] F. Ross. *AI Ethics for Enterprise AI (2019.* https://economics.harvard.edu/files/economics/files/rossi francesca_42219_aiethicsforenterpriseai_ec3118hbs.pdf. 2019.

[35] Alberto Fernandez et al. «Evolutionary Fuzzy Systems for Explainable Artificial Intelligence: Why, When, What for, and Where to?» In: *IEEE Computational Intelligence Magazine* 14.1 (2019), pp. 69–81. DOI: 10.1109/MCI.2018.2881645.

[36] M. Gleicher. «A framework for considering comprehensibility in modeling.» In: *Big data 4 (2) (2016) 75–88* (2016).

[37] M. W. Cravenr. «Extracting comprehensible models from trained neural networks». In: *Tech. rep., University of Wisconsin-Madison Department of Computer Sciences (1996* (1996).

[38] Riccardo Guidotti et al. *A Survey Of Methods For Explaining Black Box Models*. 2018. arXiv: 1802.01933 [cs.CY].

[39] Zachary C. Lipton. *The Mythos of Model Interpretability*. 2017. arXiv: 1606.03490 [cs.LG].

[40] Krishna Gade et al. «Explainable AI in Industry: Practical Challenges and Lessons Learned». In: *Companion Proceedings of the Web Conference 2020*. WWW '20. Taipei, Taiwan: Association for Computing Machinery, 2020, pp. 303–304. ISBN: 9781450370240. DOI: 10.1145/3366424.3383110. URL: https://doi.org/10.1145/3366424.3383110.

[41] Waddah Saeed e Christian Omlin. «Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities». In: *Knowledge-Based Systems* 263 (2023), p. 110273. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2023.110273. URL: https://www.sciencedirect.com/science/article/pii/S0950705123000230.

[42] Marko Robnik-Šikonja e Marko Bohanec. «Perturbation-Based Explanations of Prediction Models». In: *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*. A cura di Jianlong Zhou e Fang Chen. Cham: Springer International Publishing, 2018, pp. 159–175. ISBN: 978-3-319-90403-0. DOI: 10.1007/978-3-319-90403-0_9. URL: https://doi.org/10.1007/978-3-319-90403-0_9.

[43] Yin Lou, Rich Caruana e Johannes Gehrke. «Intelligible Models for Classification and Regression». In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '12. Beijing, China: Association for Computing Machinery,

2012, pp. 150–158. ISBN: 9781450314626. DOI: 10.1145/2339530.2339556. URL: https://doi.org/10.1145/2339530.2339556.

[44] Carlos Guestrin Marco Tulio Ribeiro Sameer Singh. *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery e Data Mining, San Francisco, CA, USA, August 13-17, 2016, 2016.

[45] Saumitra Mishra, Bob L. Sturm e Simon Dixon. «Local Interpretable Model-Agnostic Explanations for Music Content Analysis». In: *International Society for Music Information Retrieval Conference*. 2017. URL: https://api.semanticscholar.org/CorpusID:795766.

[46] Marco Tulio Ribeiro, Sameer Singh e Carlos Guestrin. *Nothing Else Matters: Model-Agnostic Explanations By Identifying Prediction Invariance*. 2016. arXiv: 1611.05817 [stat.ML].

[47] Rikard Konig, Ulf Johansson e Lars Niklasson. «G-REX: A Versatile Framework for Evolutionary Data Mining». In: *2008 IEEE International Conference on Data Mining Workshops*. 2008, pp. 971–974. DOI: 10.1109/ICDMW.2008.117.

[48] U. Johansson, Rikard König e Lars Niklasson. «The Truth is In There - Rule Extraction from Opaque Models Using Genetic Programming». In: *The Florida AI Research Society*. 2004. URL: https://api.semanticscholar.org/CorpusID:263862820.

[49] U. Johansson, Lars Niklasson e Rikard König. «Accuracy vs. comprehensibility in data mining models». In: 2004. URL: https://api.semanticscholar.org/CorpusID:5889873.

[50] Guolong Su et al. *Interpretable Two-level Boolean Rule Learning for Classification*. 2016. arXiv: 1606.05798 [stat.ML].

[51] Scott M Lundberg e Su-In Lee. «A Unified Approach to Interpreting Model Predictions». In: *Advances in Neural Information Processing Systems 30*. A cura di I. Guyon et al. 2017.

[52] Paulo Cortez e Mark J. Embrechts. «Opening black box Data Mining models using Sensitivity Analysis». In: *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. 2011, pp. 341–348. DOI: 10.1109/CIDM.2011.5949423.

[53] Alex Goldstein et al. *Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation*. 2014. arXiv: 1309.6392 [stat.AP].

[54] Helena Löfström, Karl Hammar e Ulf Johansson. «A Meta Survey of Quality Evaluation Criteria in Explanation Methods». In: *Intelligent Information Systems*. A cura di Jochen De Weerdt e Artem Polyvyanyy. Cham: Springer International Publishing, 2022, pp. 55–63. ISBN: 978-3-031-07481-3.

[55] Domenico Alfano, Roberto Abbruzzese e Domenico Parente. «Pump and Dump Cryptocurrency Detection Using Social Media». In: *Proceedings of the 12th International Conference on Data Science, Technology and Applications, DATA 2023, Rome, Italy, July 11-13, 2023*. A cura di Oleg Gusikhin, Slimane Hammoudi e Alfredo Cuzzocrea. SCITEPRESS, 2023, pp. 235–240. ISBN: 978-989-758-664-4. DOI: 10.5220/0012059300003541. URL: https://doi.org/10.5220/0012059300003541.

[56] Raidgar98. *Visualisation of pump and dump scam based on SLO-BNB cryptocurrency example*. https://upload.wikimedia.org/wikipedia/commons/6/6d/Visualisation_of_pump_and_dum

[57] A. S. Kyle e S. Viswanathan. *How to define illegal price manipulation*. American Economic Review, 2008.

[58] Josh Kamps e Bennett Kleinberg. *To the moon: defining and detecting cryptocurrency pump-and-dumps*. Crime Science, 2018.

[59] S. Alireza Hashemi Golpayegani Neda Soltani Elham Hormizi. *Anomaly Detection in Q&A Based Social Networks*. Advances in Intelligent Systems e Computing, 2018.

[60] Ruixiang Wang Ryan G. Chacon Thibaut G. Morillon. *Will the reddit rebellion take you to the moon? Evidence from WallStreetBets*. Financial Markets e Portfolio Management, 2022.

[61] Mehmet Serdar Guzel Firat Akba Ihsan Tolga Medeni e Iman Askerzade. *Manipulator Detection in Cryptocurrency Markets Based on Forecasting Anomalies*. IEEE International Conference on Artificial Intelligence in Engineering e Technology, 2022.

[62] Fred Morstatter Mehrnoosh Mirtaheri Sami Abu-El-Haija, Greg Ver Steeg e Aram Galstyan. *Identifying and Analyzing Cryptocurrency Manipulations in Social Media*. IEEE Transactions on Computational Social Systems, 2021.

[63] F. Allen e D. Gale. *Stock-price manipulation*. The Review of Financial Studies, 1992.

[64] F. Sassi La Morgia A. Mei. *Pump and Dumps in the Bitcoin Era: Real Time Detection of Cryptocurrency Market Manipulations*. IEEE, 2020.

[65] Scott Lundberg. *Shap Python Library*. https://shap-lrjball.readthedocs.io/en/latest/index.html.

[66] Bali Ranaivo-Malancon. «Building a Rule-Based Malay Text Segmentation Tool». In: *2011 International Conference on Asian Language Processing*. 2011.

[67] Saniati e Ayu Purwarianti. «Rule based approach for text segmentation on Indonesian news article using named entity distribution». In: *2014 International Conference on Data and Software Engineering (ICODSE)*. 2014.

[68] Omri Koshorek et al. «Text Segmentation as a Supervised Learning Task». In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2018.

[69] Pinkesh Badjatiya et al. «Attention-Based Neural Text Segmentation». In: 2018.

[70] John D. Lafferty, Andrew McCallum e Fernando Pereira. «Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data». In: *International Conference on Machine Learning*. 2001.

[71] Jakub Harasta et al. «Automatic Segmentation of Czech Court Decisions into Multi-Paragraph Parts». In: 2019.

[72] Michal Lukasik et al. «Text Segmentation by Cross Segment Attention». In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

[73] Jacob Devlin et al. «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2019.

[74] Goran Glavaš e Swapna Somasundaran. «Two-Level Transformer and Auxiliary Coherence Modeling for Improved Text Segmentation». In: *Proceedings of the AAAI Conference on Artificial Intelligence* (2020).

[75] Dennis Aumiller et al. «Structural Text Segmentation of Legal Documents». In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. 2021.

[76] Andrea Lombardi, Domenico Alfano e Roberto Abbruzzese. «Legal Text Segmentation Through Breakpoint Detection». In: dic. 2023. ISBN: 9781643684727. DOI: 10.3233/FAIA230968.

[77] Ilias Chalkidis et al. «LEGAL-BERT: The Muppets straight out of Law School». In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2020.

[78] Daniele Licari e Giovanni Comandè. «ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law». In: *Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management*.

[79] Doug Beeferman, Adam L. Berger e John D. Lafferty. «Statistical Models for Text Segmentation». In: *Machine Learning* (1999).

[80] Lev Pevzner e Marti A. Hearst. «A Critique and Improvement of an Evaluation Metric for Text Segmentation». In: *Computational Linguistics* (2002).

[81] Roberto Abbruzzese, Domenico Alfano e Andrea Lombardi. «REMOAC: A Retroactive Explainable Method for OCR Anomalies Correction in Legal Domain». In: dic. 2023. ISBN: 9781643684727. DOI: 10.3233/FAIA230985.

[82] Eleazar Eskin. «Detecting Errors within a Corpus using Anomaly Detection». In: *1st Meeting of the North American Chapter of the Association for Computational Linguistics*. 2000.

[83] Pratip Samanta e Bidyut B. Chaudhuri. «A simple real-word error detection and correction using local word bigram and trigram». In: *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*. The Association for Computational Linguistics e Chinese Language Processing (ACLCLP), 2013.

[84] Martin Chodorow et al. «Problems in Evaluating Grammatical Error Detection Systems». In: *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee, 2012.

[85] Courtney Napoles et al. «Ground Truth for Grammatical Error Correction Metrics». In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 2015.

[86] Daniel Dahlmeier, Hwee Tou Ng e Siew Mei Wu. «Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English». In: *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, 2013.

[87] Samuel Bell, Helen Yannakoudakis e Marek Rei. «Context is Key: Grammatical Error Detection with Contextual Word Representations». In: *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, 2019.

[88] Matthew E. Peters et al. «Semi-supervised sequence tagging with bidirectional language models». In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2017.

[89] Alan Akbik, Duncan Blythe e Roland Vollgraf. «Contextual String Embeddings for Sequence Labeling». In: *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, 2018.

[90] David Maier. «The Complexity of Some Problems on Subsequences and Supersequences». In: *J. ACM* (1978).

[91] Frederic P. Miller, Agnes F. Vandome e John McBrewster. *Levenshtein Distance: Information Theory, Computer Science, String (Computer Science), String Metric, Damerau?Levenshtein Distance, Spell Checker, Hamming Distance*. Alpha Press, 2009.

[92] Edward Ma. *NLP Augmentation*. https://github.com/makcedward/nlpaug. 2019.

[93] Dom Lachowicz. *Enchant*. https://pyenchant.github.io/pyenchant/.

[94] Ryan Kelly. *PyEnchant*. https://abiword.github.io/enchant/.

[95] Filip Sondej. *Autocorrect*. https://https://github.com/filyp/autocorrect.

[96] Peter Norvig. *Spelling Corrector*. https://norvig.com/spell-correct.html.

[97] John Snow Labs. *Context Spell Checker*. https://sparknlp.org/2021/03/08/spellcheck_dl_it.html.

[98] International Organization for Standardization Standard. *Ergonomics of human-system interaction Human-centred design for interactive systems*. 2019.

ELENCO DELLE FIGURE

ELENCO DELLE TABELLE