



UNIVERSITÀ POLITECNICA DELLE MARCHE

FACOLTÀ DI INGEGNERIA

Dipartimento di Ingegneria Civile, Edile e Architettura

Corso di Laurea magistrale in Ingegneria Edile

Sviluppo di un sistema automatico per l'assegnazione delle priorità di intervento nella manutenzione di patrimoni immobiliari mediante tecniche di
ML

Development of an automated system for prioritizing interventions in real estate maintenance using ML techniques

Relatore:

Prof. Ing. Arch. **Marco D'Orazio**

Correlatori:

Prof. Ing. **Elisa Di Giuseppe**

Prof. Ing. **Gabriele Bernardini**

Tesi di laurea di:

Bruni Adamo

Anno accademico : 2020/21

INDICE

INTRODUZIONE	4
Motivazioni.....	4
Obiettivi.....	5
Articolazione della tesi	6
1 STATO DELL'ARTE	8
1.1 La manutenzione negli edifici	8
1.1.1 Riferimenti normativi.....	10
1.2 La prioritizzazione degli interventi manutentivi	14
1.2.1 Processo standard di manutenzione.....	14
1.2.2 Rassegna della letteratura.....	17
1.3 Text mining per la prioritizzazione di interventi manutentivi	24
1.3.1 Tecniche di ML utilizzate	25
1.3.2 Rassegna della letteratura.....	36
1.4 Conclusione sullo stato dell'arte	38
2 FASI, MATERIALI E METODI	39
2.1 Le fasi operative	39
2.2 Materiali	42
2.2.1 Caso studio: le richieste manutentive all'interno dell'Università Politecnica delle Marche	42
2.2.2 Analisi del set di dati raccolti	45
2.3 Metodi.....	47
2.3.1 HMA annotation.....	47
2.3.2 Creazione DTM.....	52
2.3.3 Applicazione Classificatori	61

2.3.4	Analisi dei risultati	63
3	RISULTATI	66
3.1	HMA annotation.....	66
3.1.1	Verifica di correlazione tra annotazioni.....	66
3.1.2	Verifica di concordanza tra annotatori selezionati.....	76
3.1.3	Verifica di correlazione tra annotatori selezionati	77
3.2	Applicazione classificatori	78
3.2.1	Presentazione dei risultati.....	78
3.2.2	Elaborazione del testo con selezione di parole (DTM_sel)	79
3.2.3	Elaborazione del testo incluso lo stemming (DTM_ste).....	83
3.2.4	Elaborazione del testo escluso lo stemming (DTM_noste)	87
3.3	Discussione dei risultati	93
3.3.1	Valutazione di influenza del filtraggio dei termini	93
3.3.2	Valutazione della capacità della predizione di classi specifiche	96
4	CONCLUSIONI.....	98
5	BIBLIOGRAFIA E SITOGRAFIA	101

INTRODUZIONE

Motivazioni

I costi operativi e di manutenzione di un edificio rappresentano circa il 75% del totale dei costi del suo intero ciclo di vita. Particolarmente importante è l'aliquota destinata alla manutenzione ordinaria e straordinaria. Motivo per il quale sono state sviluppate strategie di manutenzione preventiva e programmata utili a ridurre sia gli impatti economici che le interruzioni del servizio, in uso soprattutto da parte di organizzazioni detentrici di grandi patrimoni edilizi ed erogatrici di servizi di rilievo o critici (università, ospedali, ecc....). Nonostante la manutenzione programmata sia una prassi per molte di queste organizzazioni, l'insorgenza di eventi che richiedono attività manutentive a richiesta è ancora particolarmente elevata, specialmente laddove il patrimonio edilizio dell'organizzazione non è di recente realizzazione. In luoghi come università, ospedali, possono generarsi numerosi eventi contemporanei che richiedono attività manutentive, con la necessità di struttura tecnica dimensionalmente adeguata. Non essendo la gestione del patrimonio edilizio, il "core business" di queste organizzazioni, questa attività viene normalmente demandata a contactors esterni, per mezzo di contratti di facility management comprendenti uno o più servizi.

Un aspetto critico nel processo di gestione dell'attività manutentiva riguarda il fatto che dal momento nel quale viene generata la richiesta di intervento al momento nel quale si dà soluzione al problema, intercorre un intervallo temporale anche importante, nel quale possono essere interrotti anche servizi critici. E' fondamentale quindi la corretta assegnazione di priorità all'intervento richiesto, dato che possono verificarsi anche decine di eventi di questo tipo in contemporanea in grandi strutture. Generalmente il processo

di assegnazione della priorità viene effettuato manualmente, tramite tecnici che, interpretando il contenuto delle segnalazioni, le classificano, assegnano il grado di urgenza anche tramite verifiche e quindi assegnano l'intervento, determinando le risorse sia umane che economiche da impiegare per la risoluzione delle mancanze segnalate.

Per facilitare il processo di gestione di tali richieste, negli ultimi anni sono stati molti i sistemi computerizzati di gestione della manutenzione, (CMMS) sviluppati, così da supportare le attività di manutenzione, fornendo al personale un considerevole aiuto. Tuttavia, con l'utilizzo di questi strumenti le richieste manutentive sono spesso registrate in modo incoerente o alcuni dettagli vengono semplicemente omessi dai diversi utenti. Quindi anche se un set di dati, fornito da strumenti come il CMMS, contiene informazioni testuali importanti per la valutazione delle prestazioni di manutenzione e funzionamento di una struttura, risulta difficile recuperarle in modo rapido ed efficiente.

Tuttavia, il gruppo di Facility Management, anche se supportato da sistemi computerizzati, molte volte incontra il caso in cui l'elevato numero di richieste manutentive provoca una notevole difficoltà nella corretta individuazione della gravità di ognuna di esse. Tutto questo determina un rallentamento delle attività di manutenzione, svolte in seguito alle richieste, con conseguente perdita di efficacia del totale sistema di gestione dell'edificio.

Obiettivi

In risposta a questa esigenza la tesi in questione si pone l'obiettivo di sviluppare un sistema automatizzato, basato su tecniche di Machine

Learning (ML), applicata al linguaggio naturale per l'assegnazione delle priorità di intervento nella manutenzione di patrimoni immobiliari.

Utilizzando il dataset delle richieste di intervento generate nell'arco di 34 mesi nei 23 edifici dell'Università Politecnica delle Marche (12.000 interventi ca.) si è cercato di verificare se fosse possibile riconoscere automaticamente il livello di priorità a partire dalle unità lessicali contenute nei testi delle mail con le quali veniva richiesto l'intervento e quali fossero le operazioni di pre-trattamento dei testi (filtraggio, stemming, lemmatization, ecc...) necessarie a migliorare l'affidabilità del processo.

Articolazione della tesi

Sulla base degli obiettivi proposti la seguente tesi è stata suddivisa in quattro capitoli differenti:

Nel primo capitolo troviamo lo stato dell'arte, comprendente l'analisi della letteratura tecnica e della letteratura scientifica. A seguito di una analisi della normativa vigente sulla manutenzione in campo edilizio, viene condotta una analisi degli articoli scientifici maggiormente rappresentativi. Viene introdotto l'argomento della prioritizzazione degli interventi manutentivi spiegando in prima istanza il processo standard di manutenzione di un generale sistema edilizio e in seguito vedendo quali sono stati i lavori correlati che hanno interessato tale argomento. Infine, viene condotto un approfondimento sull'uso del text mining in campo della prioritizzazione degli interventi manutentivi facendo riferimento maggiormente alle tecniche di machine learning impiegate e descrivendo i principali lavori correlati riguardanti tale argomento.

Nel secondo capitolo vengono individuate le fasi operative intercorse per lo sviluppo del sistema automatico. Viene descritto il caso studio e viene analizzato il set di dati dove sono state raccolte le richieste manutentive. Successivamente vengono individuati i metodi applicati che hanno portato allo sviluppo del metodo.

Nel terzo capitolo sono presenti i risultati. Questi vengono visti in riferimento alle fasi operative intercorse durante lo sviluppo del metodo e facendo particolare riferimento alla variazione dei risultati nell'applicazioni dei tre casi sopra descritti.

Nel quarto troviamo le conclusioni. Vengono riassunti i risultati e viene presentata una prospettiva di ricerca in riferimento a quanto trattato nella tesi.

1 STATO DELL'ARTE

In questo capitolo viene presentato lo stato dell'arte degli argomenti trattati. In primo luogo viene introdotta la manutenzione nell'ambito edilizio, descrivendo quali sono le principali problematiche legate ad essa e successivamente individuando le norme e gli standard che la riguardano. Nella seconda parte viene fatta una rassegna della letteratura a riguardo della prioritizzazione degli interventi manutentivi. Infine si individuano le tecniche di possibile utilizzo a riguardo del tema trattato.

1.1 La manutenzione negli edifici

Gli edifici si deteriorano gradualmente nel tempo a causa di agenti atmosferici, eventi catastrofici e per azioni dell'uomo. Le attività di manutenzione sono necessarie per impedire questo processo di deterioramento e per garantire il funzionamento continuo degli edifici. La manutenzione è un processo sistematico per salvaguardare le caratteristiche e le risorse gestite a livelli ottimali per raggiungere gli standard e le prestazioni dell'edificio [1]. Negli ultimi decenni, con il rapido aumento dei costi delle nuove costruzioni, l'efficace manutenzione è diventata molto importante per gli edifici esistenti [2].

Il costo riguarda una delle problematiche principali su cui focalizzare la nostra attenzione. Stimando la totalità dei costi che caratterizzano il ciclo di vita di un edificio, attraverso la Life Cycle Cost [3] Analysis (LCCA) , vediamo come la manutenzione riguarda circa il 75% della spesa totale [4].

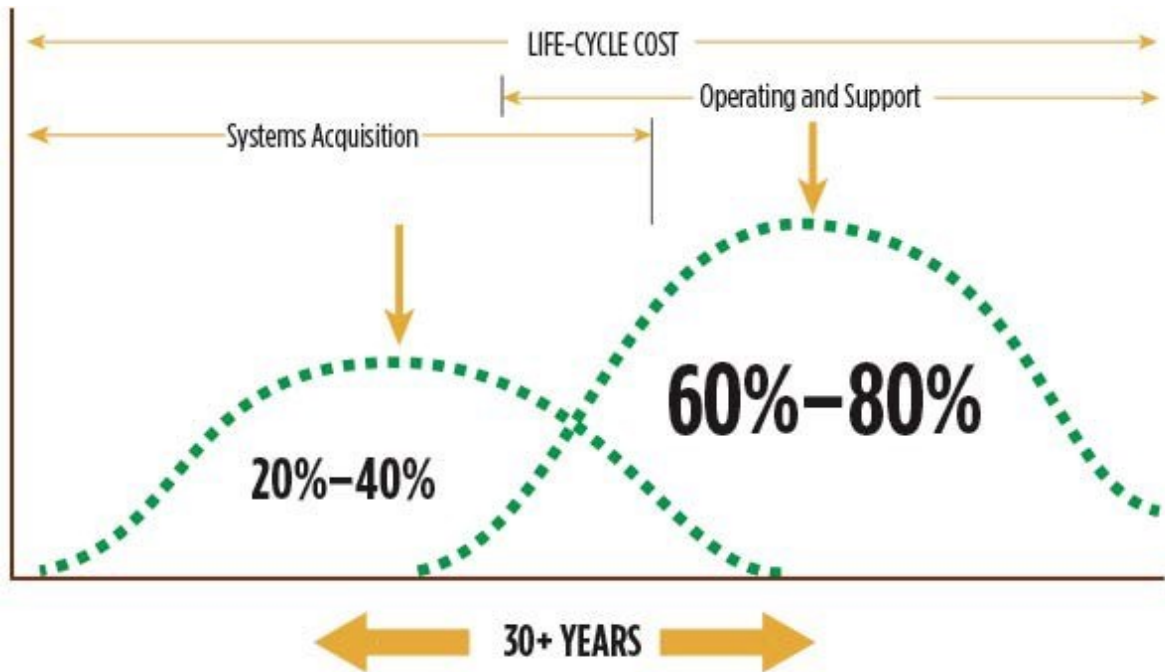


Figura 1.1.1 Life-Cycle cost[4]

Vista e considerata questa analisi, la manutenzione, anche se molto importante, risulta una voce di costo a cui non si presta molta attenzione, dedicando ad essa spesso la quota minima di budget rispetto agli altri costi compresi nell'intero ciclo di vita.

L'incremento dei costi di manutenzione negli edifici, deriva principalmente da una scorretta pianificazione della gestione di essa. Le attività manutentive svolte in tempi e modi non correttamente pianificati comportano una perdita di efficienza del sistema edilizio [5].

La selezione delle strategie di manutenzione efficienti e un utilizzo di approcci decisionali corretti possono portare a una migliore gestione del budget e della manutenzione. [6] [7].

1.1.1 Riferimenti normativi

Di seguito vengono presentate e commentate le definizioni relative alla manutenzione facendo riferimento alle norme UNI EN 13306, UNI EN 9910 e la UNI 10147

La norma tecnica UNI EN 13306 definisce la manutenzione nel modo seguente : *“combinazione di tutte le azioni tecniche, amministrative e gestionali, durante il ciclo di vita di un’entità, destinate a mantenerla o a riportarla in uno stato in cui possa eseguire la funzione richiesta”*.

Questa definizione considera la manutenzione come quell’insieme di operazioni che cercano di portare l’edificio ad uno standard accettabile, cioè il livello al quale fare riferimento per la determinazione dello stato di un oggetto e quindi alla sua funzionalità, all’estetica o alle prestazioni tecnologiche.

1.1.1.1 Le strategie manutentive

Al verificarsi di un guasto, un’avaria o una semplice deriva vengono impiegate delle strategie manutentive per la risoluzione di tali problematiche.

Di seguito vengono identificate le principali strategie manutentive in riferimento alle normative UNI EN 13306 UNI EN 9910 e la UNI 10147 .

Manutenzione correttiva (o manutenzione a guasto)

“Manutenzione eseguita a seguito dalla rilevazione di un avaria e volta a riportare l’entità in uno stato in cui possa eseguire la funzione richiesta”

Questa tipologia di manutenzione viene applicata ad un macchinario, componente o asset a seguito di un fermo o di una rottura con l'obiettivo di ripristinarlo. È una delle tipologie di manutenzione più costosa in quanto deve essere effettuata su di un elemento già compromesso. Questa caratteristica fa sì che la manutenzione correttiva può considerarsi efficace quando applicata a macchinari a basso costo tali da non compromettere il funzionamento di un intero impianto.

A sua volta la manutenzione correttiva si divide in altre categorie. Di seguito vengono elencate le tipologie con relativa definizione:

- Manutenzione differita: “ correttiva che non è eseguita immediatamente dopo la rilevazione di un'avaria, ma che è differita in conformità a determinate regole”;
- Manutenzione d'urgenza: “ correttiva eseguita senza indugio dopo la rilevazione di un guasto in modo da evitare conseguenze inaccettabili”.

La manutenzione migliorativa

“Insieme delle azioni di miglioramento o piccola modifica che non incrementano il valore patrimoniale del bene”

Questa tipologia di manutenzione viene applicata con lo scopo di ridurre o eliminare le possibilità di guasto. Questa strategia manutentiva è molto efficace ed è effettuata a basso costo in quanto, attraverso semplici migliorie possono essere risolti molti problemi ricorrenti. Con questa non si mira ad aumentare la capacità produttiva ma si incentivano punti di forza come l'affidabilità, la sicurezza e la qualità.

La manutenzione preventiva

“ Manutenzione eseguita ad intervalli predeterminati o secondo criteri prescritti e prevista per ridurre la probabilità di guasto o di degrado nel funzionamento di un’entità”

Questa tipologia di manutenzione ha l’obiettivo di ridurre il numero di guasti imprevisti e inaspettati andando a beneficiare sui costi eccessivi che devono essere impiegati per la risoluzione dei problemi. Un corretto studio sulle tempistiche e sulle modalità di guasto rendono l’applicazione di questo metodo molto efficace.

A sua volta la manutenzione preventiva si divide in altre categorie. Di seguito vengono elencate la tipologie con relativa definizione:

- Manutenzione ciclica: “ preventiva effettuata a intervalli di tempo stabiliti o un numero di unità di misura di utilizzo, ma senza una precedente indagine sulle condizioni dell’entità”;
- Manutenzione di opportunità: “preventiva eseguita in anticipo rispetto all’occorrenza, pianificata come conseguenza di un’attività imprevista che ne consente l’esecuzione a un minor costo ridotto o con un minor numero di risorse”;
- Manutenzione secondo condizione: “preventiva che include una combinazione di monitoraggio delle condizioni e/o ispezione e/o prove, analisi e le azioni di manutenzione che ne conseguono”.

Manutenzione predittiva

“Manutenzione su condizione eseguita in seguito a una previsione derivata dall’analisi ripetuta o da caratteristiche note e dalla valutazione dei parametri significativi afferenti il degrado dell’entità.”

In sostanza la manutenzione predittiva è una strategia manutentiva che utilizza dati di monitoraggio per prevedere le condizioni della macchina prendendo le decisioni in base a ad un modello predittivo[8].

La progressiva diffusione di innovazioni tecnologiche come IoT (Internet of Things) e CMMS, integrati con tecniche di ML permettono di recepire dati ed elaborarli per ottenere modelli predittivi [9][10]. Si deduce quindi che l’approccio di una manutenzione predittiva presenta sfide come la corretta estrazione dei dati e lo sviluppo di algoritmi predittivi accurati [11].

Per la valutazione delle condizioni dei sistemi edilizi, l’analisi delle richieste di manutenzione degli utenti finali fornisce una buona fonte di informazioni, dato che queste rappresentano la percezione che una caratteristica o un elemento dell’edificio è sottoperformante. L’analisi delle richieste manutentive riflette quindi, in maniera diretta le prestazioni dell’edificio [12].

Questo studio, opera in questo ambito, cercando di sfruttare le informazioni ricavabili dai testi di richieste manutentive così da essere utilizzate per lo sviluppo di un modello in grado di prioritarle autonomamente.

1.2 La prioritizzazione degli interventi manutentivi

Nel seguente paragrafo si intende esaminare quali sono state le principali metodologie utilizzate per dare priorità agli interventi manutentivi. In prima istanza viene descritto ed esaminato un processo di gestione standard all'interno di un edificio e successivamente viene fatta una rassegna sui principali studi effettuati riguardanti questo ambito.

1.2.1 Processo standard di manutenzione

All'interno delle strutture edilizie sono identificati processi di manutenzione che vedono come protagonisti gli operatori del gruppo di Facility Manager, gli utilizzatori delle strutture o il team di ispezione e gli addetti alla manutenzione. Questi operano in modo coordinato cercando di offrire un buon servizio di manutenzione.

Di seguito viene descritto in brevi passaggi il processo manutentivo di un edificio:

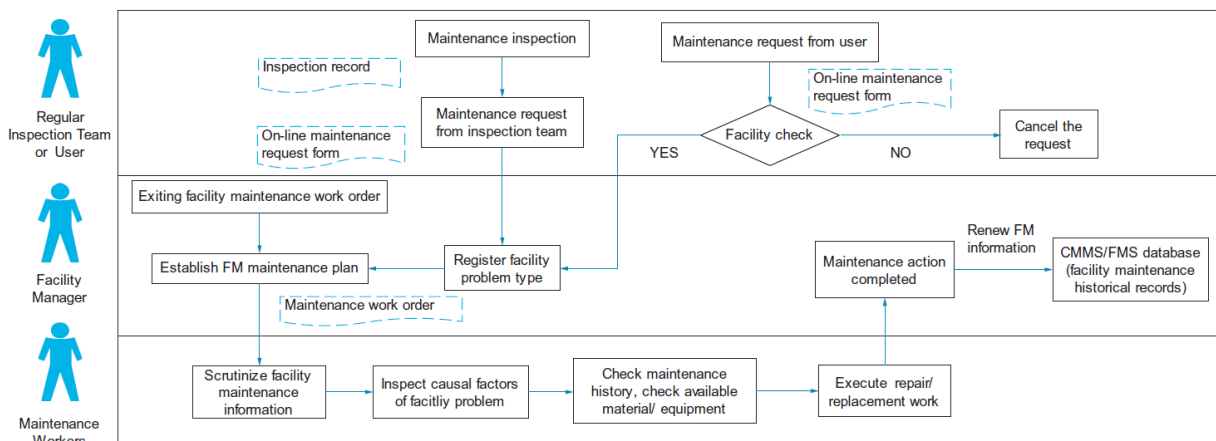


Figura 1.2.1: Processo di manutenzione [5]

Nell processo mostrato in figura 1.2.1 vediamo come un team di ispezione controlla abitualmente le condizioni dei componenti dell'edificio. Se uno di questi non funziona correttamente, il team di ispezione riferisce immediatamente il suo stato al facility manager attraverso una richiesta di manutenzione che viene esaminata e successivamente trasformata in un ordine di lavoro. Il facility manager registra il problema con l'ausilio di sistemi computerizzati come CMMS o FMC e assegna la priorità degli interventi manutentivi e agli addetti alla manutenzione i compiti da eseguire. Questi ultimi controllano la disponibilità di materiali e attrezzature, quindi riparano e sostituiscono il componente. Quando la manutenzione è stata completata, il facility manager aggiorna i registri di FM.

Tutte queste informazioni e le attività di manutenzione delle strutture possono essere supportate da sistemi computerizzati di gestione della manutenzione (CMMS) e sistemi di gestione delle strutture (FMS) per gestire il processo di manutenzione degli edifici e per fornire una piattaforma informativa meglio organizzata [7].

I CMMS sono tradizionalmente utilizzati dalle organizzazioni di manutenzione delle strutture per registrare, gestire e comunicare le operazioni quotidiane e possono essere distribuiti per la gestione delle risorse, il controllo dell'inventario, la generazione delle richieste di servizio, la generazione degli ordini di lavoro di diversi tipi e tracciare le risorse dei servizi e dei materiali utilizzati per completare gli ordini di lavoro [13]. I sistemi FMS o CAFM che integra moduli grafici CAD (Computer Aided Design) e software di database con lo scopo di fornire la capacità di gestione delle strutture, compresa quella dello spazio[14].

Purtroppo vengono individuati due principali problemi nel processo in cui questi sistemi computerizzati vengono utilizzati[7]:

- 1- Non possono fornire una programmazione automatica degli ordini di lavoro
- 2- Le informazioni raccolte non sono molto accurate

1.2.2 Rassegna della letteratura

Gli articoli in letteratura mostrano varie strategie dedicate alla prioritizzazione degli interventi manutentivi. Di seguito vengono identificati gli studi effettuati a riguardo dell'ambito trattato nella tesi facendo riferimento all'autore e alla tecnica utilizzata.

Il primo studio, [7] condotto da W. Chen et al., ha visto come obiettivo quello di organizzare gli ordini di lavoro attraverso una metodologia integrata BIM-FM. Per questo scopo è stato impiegato un algoritmo Dijkstra modificato, normalmente utilizzato per la ricerca dei cammini minimi in un determinato grafo. Nel processo sono stati considerati quattro differenti fattori delle richieste manutentive: distanza, tipo di problema, livello di emergenza e posizione.

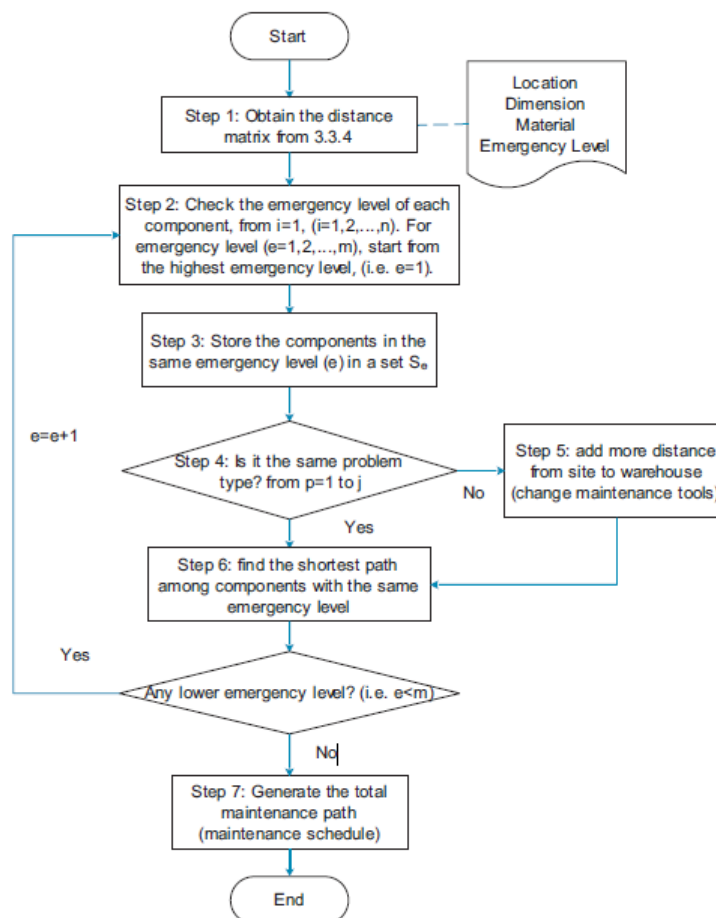


Figura 1.2.1 Processo di prioritizzazione[5]

Nella figura 1.2.1 è possibile individuare, come dopo l'assegnazione dell'emergenza ad ogni attività manutentiva (step 1), riferita ad un determinato componente, la priorità viene stabilita attraverso la scelta del percorso più breve tra i componenti dello stesso livello di emergenza (step 6).

Nel secondo studio [15] di Beauregard Michael A. and Steven K. Ayer è stato sviluppato uno strumento di supporto decisionale (DTS) per dare priorità ad ordini di lavoro derivanti da strutture istituzionali. Questo strumento offre la possibilità al gruppo di facility manager di valutare la gravità della richiesta con una compilazione manuale orientata verso quattro fattori differenti: Influenza (soggetto a cui è arrecato il danno), stato dell'edificio, uso dell'edificio e luogo dell'edificio.

Nella figura 1.2.2 viene mostrata l'applicazione di questo metodo.

IMPACT

Very High
High
Moderate
Low
Very Low

INFLUENCE	BUILDING STATUS	BUILDING USAGE	INSTITUTIONAL ENABLER(S)
Teacher's of their Students <input checked="" type="checkbox"/>	Site	Student Instruction	INSTRUCTIONAL SPACE PLAYGROUND TECHNOLOGY ZONES
Non-Teacher's who interface with Students	Building <input checked="" type="checkbox"/>	Study Space (non-classroom)	REFERENCE SPACE CREATIVE SPACE TEMP. CONTROL / COMFORT
District or School Admin.	Room	Special / General Use <input checked="" type="checkbox"/>	CAFETERIA OVERALL IMPRESSION ELECTRICAL BATHROOMS ACADEMIC MAGNET PLUMBING <input checked="" type="checkbox"/>
Other's who DO NOT interface with students		Administrative Offices	CIRCULATION ZONES VISUALIZATION PAINT / PATCH ACTIVITY ZONES SAFETY LANDSCAPING
OTHER: _____	OTHER: _____	OTHER: _____	LEARNING ZONES AIR QUALITY EQUIP. & FURNISHINGS SAFE PLACE ACOUSTICS HARDWARE QUIET SPACE

JUSTIFICATION: _____ BY: _____ DATE: ___ / ___ / ___

Figura 1.2.2 Multi-criterio decisionale [8]

Come mostrato in figura il gruppo di facility manager seleziona in base ai quattro fattori le caratteristiche relative ad ogni intervento manutentivo

stabilendo così la gravità secondo una scala. Il metodo risulta efficace in ambito della manutenzione ordinaria.

Un terzo studio, condotto da S. Madureira et al.[16] presenta una metodologia per implementare un piano di manutenzione per le facciate di edifici. In particolare lo studio si sviluppa su tre fasi:

- Ispezione dettagliata degli edifici esistenti con individuazione delle anomalie presenti e compilazione delle migliori strategie di intervento;
- Applicazione delle azioni correttive, secondo la diagnosi di ispezione, ristabilendo i livelli di prestazione dell'edificio;
- Impostazione di un piano di manutenzione proattivo necessario a prevenire e monitorare le prestazioni delle facciate;

In particolare, nella seconda fase, la prioritizzazione degli interventi è stata risolta attraverso l'applicazione di una metodologia che tiene conto di quattro criteri: esposizione ambientale (EE); estensione dell'anomalia (AE); conseguenza del mancato intervento (NI) e non conformità ai requisiti (NC). Questa tecnica è basata sul sistema dei fattori moltiplicatori Gasper e de Brito in cui ogni criterio è diviso in tre livelli di gravità: alto (3), medio (2) e basso (1). Inoltre, a ciascun criterio è stato stabilito un fattore moltiplicatore (da 1 a 4). La tabella mostra la struttura della classificazione dei criteri.

Criterion	Level	Description	Score	Multiplying factor
Environmental exposure (EE)	Low	Rural area	1	1
	Medium	Urban area	2	
	High	Coastal area	3	
Anomaly extent (AE)	Low	≤33%	1	2
	Medium	34 a 66%	2	
	High	≥67%	3	
Non-intervention consequences (NI)	Low	Minimal effects	1	3
	Medium	Degradation increase	2	
	High	MSE rupture	3	
Non-compliance with requirements (NC)	Low	Complies with minimal aesthetical, safety and functional requirements	1	4
	Medium	It does not comply with aesthetical requirements	2	
	High	It does not comply with functional requirements	3	
	Very high	It does not comply with safety requirements	4	

Tabella 1.2.1: Classificazione dei criteri di prioritizzazione

Successivamente sulla base del criterio descritto, viene determinato un indicatore di valutazione delle priorità (1).

$$P_{intervention} = \frac{P(anomaly)}{Max(P_{anomaly})} \times 100 = \frac{1EE+2EA+3NI+4NC}{34} \times 100 \quad (1)$$

Dove $P_{intervention}$ rappresenta l'indicatore di classificazione della priorità dell'intervento e $P_{anomaly}$ sta per il peso dell'anomalia rilevata.

Negli studi appena esaminati è possibile immaginare come la prioritizzazione degli ordini di lavoro sia fatta da richieste manutentive, precedentemente analizzate e corrette da eventuali errori grammaticali o

lessicali per facilitare il processo di comprensione. Questa pre-analisi del testo, che potrebbe impiegare risorse umane, temporali ed economiche può essere evitata applicando tecniche di text mining le quali permettono di analizzare un testo destrutturato ricavandone informazioni utili da utilizzare in modo più efficace. In riferimento a quanto detto di seguito viene presentato uno studio dove l'applicazione della tecnica sopra citata viene utilizzata nell'ambito trattato nella tesi.

In questo quarto studio, condotto da R. Bortolini et al. [12] vediamo quindi come la gestione delle richieste manutentive è stata effettuata attraverso un approccio basato su text mining. In particolare è stato proposto un metodo per classificare ed analizzare le richieste manutentive, le condizioni dei sistemi dell'edificio, determinare i problemi tipici che gli utenti finali lamentano e stabilire la loro gravità.

La metodologia di ricerca ha impiegato due fasi principali. Nella prima, è stato proposto un sistema di classificazione dei problemi sulla base dei risultati di un indagine condotta con cinquantatre esperti sul tema delle prestazioni degli edifici. I risultati di questa indagine sono stati utilizzati per sviluppare quindi un sistema di ispezione degli edifici oggettivo e standardizzato.

Nella seconda fase, è stato presentato un approccio di text mining per la gestione delle richieste manutentive. Inizialmente sono state estratte le richieste manutentive dal CMMS, creando un set di dati in formato .csv dove sono stati rimossi i segni di punteggiatura e gli spazi. Successivamente sono state classificate le richieste manutentive in "categorie" attraverso un'analisi di frequenza delle parole. Infine, è stata stabilita la gravità (basso, medio e alto) attraverso l'individuazione della frequenza relative alle parole

caratterizzanti di ogni livello. Le parole più frequenti relative all'alta gravità sono quelle che l'utente finale o il team di facility management usano quando è richiesta una riparazione o un'azione immediata (ad esempio, urgente, sicurezza, emergenza, allarme, incendio). Le parole relative alla bassa gravità sono quelle che l'utente finale o il team FM usano quando una riparazione o un'azione può essere posticipata e pianificata (ad esempio, dare un'occhiata, cambiare, verificare, pulire, dipingere). Le richieste non classificate in nessuna delle categorie precedenti sono definite di media gravità.

I risultati hanno mostrato come un rilevamento automatico dei guasti può essere utile per pianificare le azioni di manutenzione correttiva e per migliorare le prestazioni dei sistemi dell'edificio e quindi la soddisfazione degli utenti finali.

Nel corso degli anni la prioritizzazione di richieste manutentive ha riguardato diversi ambiti, oltre a quello delle costruzioni. Troviamo la prioritizzazione di richieste di riparazione per impianti industriali [17][18]. Xu et al.[19] hanno proposto un approccio orientato alla gestione di ordini di lavoro di interruttori. Hamdi et al.[20] hanno progettato un sistema di gestione per richieste di manutenzione di attrezzature mediche. Lewandowski et al. [14] hanno proposto una nuova metodologia per dare priorità alla manutenzione guidata dagli eventi utilizzando informazioni di sistemi di monitoraggio delle condizioni.

Nei casi studio appena descritti, vediamo come l'assegnazione delle priorità degli ordini di lavoro viene ricavata non in maniera automatica e diretta, ma con l'impiego di decisioni prese da chi genera il metodo. L'efficienza di tali

modelli sarà quindi dipendente della professionalità e dell'esperienza di chi lo genera.

Con l'ausilio di classificatori basati su algoritmi di Machine Learning è possibile addestrare un modello partendo direttamente da un set di dati iniziali in cui le informazioni sono già note. Questo permette al modello di classificare in maniera oggettiva in quanto viene data al computer la capacità di "imparare" senza esplicitamente essere programmato[21].

1.3 Text mining per la prioritizzazione di interventi manutentivi

In questo paragrafo viene presentata una rassegna della letteratura riguardante le tecniche utilizzate nel campo del text mining facendo particolare attenzione agli algoritmi di Machine Learning (ML) , impiegati in problemi simili a quello trattato in questa tesi. Questa indagine ha lo scopo di identificare gli algoritmi più promettenti da utilizzare per il raggiungimento di buoni risultati.

Gli approcci di apprendimento automatico sono suddivisi in tre categorie principali: supervisionato, non supervisionato e semi-supervisionato[22]. Nell'apprendimento supervisionato, vengono messi a disposizione campioni di dati etichettati, così da permettere all'algoritmo di ML di sviluppare ipotesi di classificazione basate sui dati iniziali. L'apprendimento non supervisionato scopre modelli nascosti nei dati senza che questi vengano etichettati. A metà strada tra questi due grandi gruppi troviamo l'apprendimento semi-supervisionato dove i dati di partenza sono solo in parte etichettati [23].

Nel caso studio trattato in questa tesi l'applicazione degli algoritmi di ML è avvenuta attraverso l'impiego dell'app *Classification Learner* di MATLAB che utilizza la tecnica *cross-validation* (convalida incrociata). In sostanza, una volta stabilito il numero di parti "k" in cui dividere il dataset, l'applicazione :

- 1- Divide i dati in k parti;
- 2- Per ogni parte di verifica:

- a- Addestra il modello usando le osservazioni della parte di addestramento (osservazioni non contenute nella parte di verifica)
 - b- Valuta le prestazioni del modello usando i dati della parte di verifica
- 3- Calcola l'errore medio di validazione su tutte le parti.

Nel nostro caso il dataset è stato diviso in 5 parti.

La classificazione automatica del testo avviene quindi attraverso l'applicazione di tecniche di ML in cui, attraverso l'addestramento di un algoritmo finalizzato ad estrarre caratteristiche da un insieme di documenti di testo pre-etichettati, ne classifica di nuovi in base al loro contenuto [23].

Baek et al.[24] hanno condotto una revisione completa sullo stato attuale del text mining nel mondo delle costruzioni. In questa ricerca sono state identificate le tre fasi principali che intercorrono nello studio di un testo: pre-elaborazione, generazione delle caratteristiche ed estrazione delle informazioni. La prima indica il lavoro preliminare sui dati di testo grezzo prima di codificarlo in un formato leggibile dalla macchina. La seconda fase si riferisce alla caratterizzazione dei dati di testo come espressioni decifrabili dalla macchina e la terza è l'estrazione delle informazioni significative dai dati di testo. Nella fase di estrazione delle informazione, lo studio condotto rileva un utilizzo frequente di algoritmi di ML del tipo supervisionato.

1.3.1 Tecniche di ML utilizzate

In questo paragrafo vengono indicati brevemente i principali tipo di algoritmi di ML supervisionati impiegati nel text mining facendo riferimento maggiormente al libro scritto da Jan Žižka et al.[22] :

1.3.1.1 *Classificatore bayesiano*

Il classificatore Naive Bayesian (NB) è uno dei metodi di classificazione più semplice ed efficace. La grande efficienza di questo metodo è dovuta principalmente all'assunzione di indipendenza condizionata tra le classi che vengono predette [25].

Consideriamo un set di dati suddiviso in due parti. La prima costituita dalla matrice delle caratteristiche e la seconda costituita dal vettore risposta (target). La matrice delle caratteristiche "X" contiene tutti i vettori (righe) del set di dati in cui ogni vettore consiste nel valore delle caratteristiche dipendenti. Il numero di caratteristiche è "d". Quindi : $X=(x_1,x_2,x_3,x_d)$. Il vettore risposta (target) "y" contiene il valore della classe (gruppo) per ogni riga della matrice della caratteristiche.

Le due principali assunzioni di Naive Bayes sono legate al fatto che ogni caratteristica della stessa classe contribuisce in maniera indipendente e uguale al risultato.

Prima di verificare l'applicazione del classificatore introduciamo il teorema su cui è basato tale classificatore.

Teorema di Bayes

Il teorema di Bayes permette di calcolare la probabilità di un determinato evento A considerando le informazioni disponibili sull'evento B. Il teorema di Bayes è enunciato matematicamente come segue

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)} \quad (2)$$

A, B = Eventi

$P(B|A)$ = Probabilità dell'evento B dato A

$P(A), P(B)$ = Probabilità indipendenti (a priori)

$P(A|B)$ = Probabilità dell'evento A dato B

Il modello di Naive Bayes

Considerando quindi una matrice di dati X e un vettore risposta y, introduciamo quanto segue:

$$P(y | X) = \frac{P(X | y)P(y)}{P(X)} \quad (3)$$

y è la variabile di classe e X è un vettore di caratteristiche dipendenti con dimensione d cioè $X = (x_1, x_2, x_3, x_d)$, dove d è il numero di variabili/caratteristiche del campione. $P(y|X)$ è la probabilità di osservare la classe Y dato il campione X.

Applicando le principali assunzioni prima dette, assumiamo che tutte le caratteristiche in X siano reciprocamente indipendenti, condizionate dalla categoria y.

$$P(y | x_1, \dots, x_d) = \frac{P(y) \prod_{i=1}^d P(x_i|y)}{P(x_1)P(x_2)\dots P(x_d)} \quad (4)$$

Considerando che il numeratore rimane costante:

$$P(y | x_1, \dots, x_d) \propto P(y) \prod_{i=1}^d P(x_i | y) \quad (5)$$

Infine, per trovare la probabilità di un dato campione per tutti i possibili valori della variabile di classe y , applichiamo:

$$y = \operatorname{argmax}_y P(y) \prod_{i=1}^n P(x_i | y) \quad (6)$$

1.3.1.2 *K-nearest neighbors*

Il classificatore K-Nearest Neighbor (KNN) è un metodo ampiamente utilizzato e viene spesso impiegato nel campo del text mining. In sostanza, avendo dei documenti etichettati in partenza, per classificare uno nuovo “x” l’algoritmo KNN trova i “k” vicini del documento etichettato [26].

Di seguito vengono descritte le caratteristiche operative principali dell’algoritmo:

- 1- Documenti etichettati ($\vec{d}_1, \vec{d}_2, \dots, \vec{d}_2$).
- 2- Quando deve essere classificato un documento sconosciuto \vec{x} , l’algoritmo trova i documenti simili a questo. Quindi, una volta ottenuti i “ k ” documenti più simili. La somiglianza viene calcolata con il coseno di similitudine come segue:

$$\operatorname{sim}(\vec{x}, \vec{d}_i) = \frac{\sum_{k=1}^N x_k \times d_{ik}}{\sqrt{(\sum_{k=1}^N x_k^2)(\sum_{k=1}^N d_{ik}^2)}}. \quad (7)$$

Dove N è il numero totale di elementi delle caratteristiche.

- 3- Con i k documenti più simili a \vec{x} , l'importanza di \vec{x} per ogni classe C_j può essere calcolata come segue:

$$p(\vec{x}, C_j) = \sum_{\vec{d}_i \in \text{KNN}(\vec{x})} \text{sim}(\vec{x}, \vec{d}_i) y(\vec{d}_i, C_j). \quad (8)$$

$\text{KNN}(\vec{x})$ denota il set di documenti k simili a \vec{x} . Mentre $y(\vec{d}_i, C_j)$:

$$y(\vec{d}_i, C_j) = \begin{cases} 1, & \vec{d}_i \in C_j \\ 0, & \vec{d}_i \notin C_j \end{cases}. \quad (9)$$

- 4- Infine, \vec{x} viene classificato calcolando l'importanza massima su $p(\vec{x}, C_j)$:

$$C = \arg \max_{C_j} (p(\vec{x}, C_j)). \quad (10)$$

1.3.1.3 Decision Trees

Decision Trees (DT) è un classificatore gerarchico ad albero. È composto da un insieme di nodi che simboleggiano le decisioni o i risultati finali e le foglie che rappresentano la suddivisione dei dati. Esso classifica un determinato campione di dati applicando una serie di regole. Ognuna di esse è rappresentata da un nodo che a sua volta ne produce un altro per ogni risultato della regola applicata.

Tale processo viene ripetuto fino a quando i dati campione sono ordinati in una classe seguendo il percorso del nodo iniziale a quello finale [27]

Nell'immagine mostrata vediamo un esempio di albero decisionale. Per la corretta predizione si parte da un nodo superiore. In seguito alla regola applicata abbiamo altri due diversi nodi. Il classificatore controlla i valori e decide quale ramo proseguire. Quando i rami raggiungono un nodo foglia, i dati sono classificati come 0 e 1.

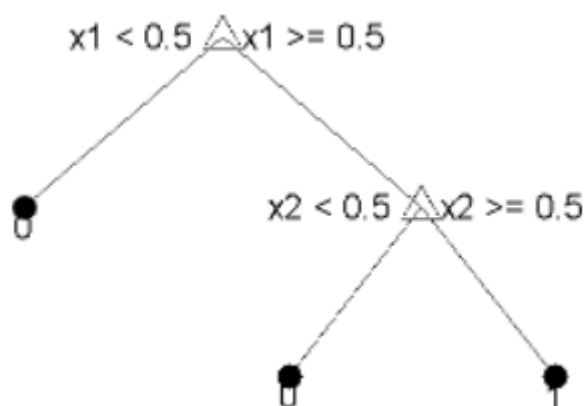


Figura 1.3.1: Esempio Decision Trees[32]

1.3.1.4 Random Forest

Random Forest è un metodo di classificazione basato sull'applicazione di un gruppo di alberi decisionali. L'idea di questo classificatore di basa sul tentativo di sopprimere l'overfitting del gruppo di alberi. Per overfitting si intende la situazione in cui l'algoritmo crea relazioni tra un attributo irrilevante e il risultato finale nel set di dati etichettati, portando ad un errore di apprendimento durante il processo di addestramento. Tutto questo porta il modello ad essere perfetto una volta applicati i dati etichettati e pessimo quando vengono impiegati dati di test. Con l'applicazione di algoritmi Random Forest questa particolare situazione viene evitata selezionando casualmente gli attributi per ogni albero ad ogni nodo di divisione permettendo all'algoritmo di correggere eventuali errori [28].

Le classificazione effettuate con Random Forest sono generalmente più robuste e possono raggiungere prestazioni migliori dei singoli Decision Tree.

1.3.1.5 *Support Vector Machine*

Support Vector Machine (SVM) è una tecnica di apprendimento potente molto utilizzata nel campo del text mining. Durante la fase di addestramento, considerato uno spazio vettoriale, l'algoritmo genera uno spazio linearmente separabile identificato come iperpiano. Questo delimita lo spazio che meglio separa le due classi di un determinato campione. Per uno spazio linearmente separabile, la superficie di decisione è un iperpiano che può essere scritto come:

$$wx + b = 0 \quad (11)$$

Dove x è un oggetto arbitrario da classificare, w il vettore ricavato dall'addestramento del set di dati e b una costante.

Per gli oggetti linearmente inseparabili, i dati di input originali vengono trasformati in uno spazio di dimensioni superiori utilizzando una mappatura non lineare e l'iperpiano di separazione lineare può essere trovato anche nel nuovo spazio senza aumentare la complessità di calcolo utilizzando la funzione kernel. L'applicazione di questa funzione ha lo scopo di ricavare le relazioni tra le caratteristiche multidimensionali rappresentate in uno spazio [29].

Nella figura seguente si nota un esempio di applicazione grafica del metodo

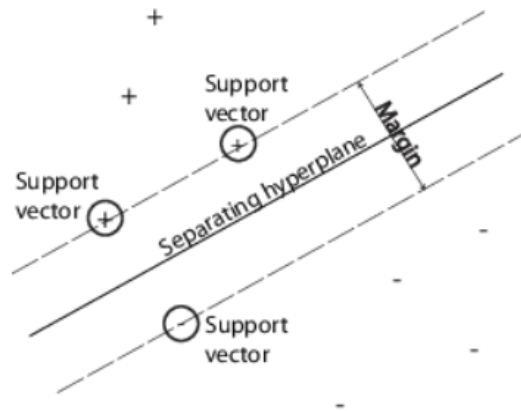


Figura 1.3.2: Esempio grafico SVM

1.3.1.6 Logistic Regression

La Logistic Regression è un particolare modello di regressione in cui la relazione tra la variabile obiettivo e la variabile indipendente è stimata attraverso una distribuzione logistica cumulativa.

Questo metodo può essere visto come un caso specifico del modello lineare generalizzato e quindi analogo alla regressione lineare.

Il classificatore logistico è spesso utilizzato perché computazionalmente leggero e possiede un discreto potere di generalizzazione. [30].

1.3.1.7 Neural Networks Classifiers

Le Neural Networks (Reti Neurali) sono algoritmi basati sul comportamento del cervello umano, imitando il modo in cui i neuroni biologici si inviano segnali. Il modello è composto da due strati di nodi, uno di input ed uno di output, e altri nodi intermedi facoltativi. Al crescere degli strati il modello diventa sempre più complesso. Ogni nodo ha un peso e una soglia associati. Le informazioni vengono trasferite da un nodo all'altro seguendo gli strati. Inizialmente le informazioni vengono trasmesse allo strato di ingresso (input) e successivamente elaborate. Se l'output di qualsiasi nodo è al di sopra della soglia specificata, tale nodo viene attivato, inviando le informazioni allo strato successivo[31]. Nell'immagine seguente vediamo una rappresentazione grafica del modello:

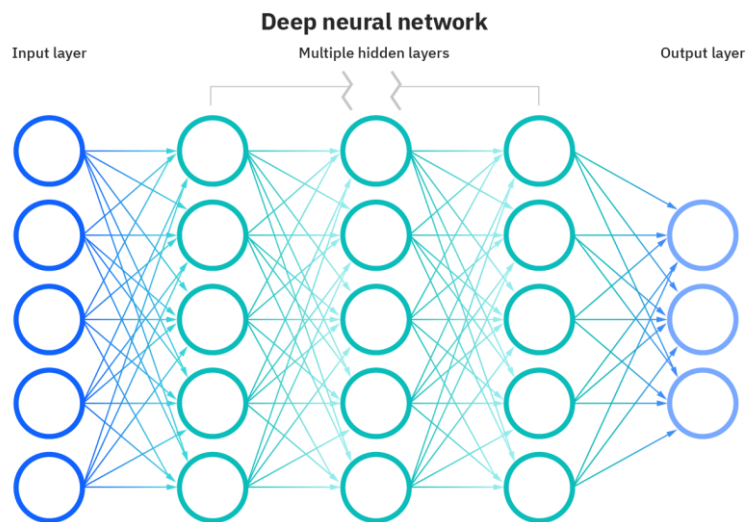


Figura 1.3.3: Modello Neural Network[31]

Successivamente ad un breve approfondimento riguardante le tipologie principali di algoritmi di tipo supervisionato, di seguito troviamo una tabella in cui vengono identificati i punti di vantaggio e svantaggio dei classificatori. Per l'individuazione delle caratteristiche elencate è stata presa in riferimento la guida MATLAB [32].








<i>Classificatore</i>	<i>Vantaggi</i>	<i>Svantaggi</i>
 Naive Bayes Classifiers	<ul style="list-style-type: none"> • Facilità di interpretazione • Veloce nella fase di training • Efficiente nella previsione di più classi 	<ul style="list-style-type: none"> • La condizione di indipendenza tra le diverse classi potrebbe portare a stime errate nella previsione dei dati
 K-Nearest Neighbor Classifier	<ul style="list-style-type: none"> • Buona accuratezza in spazi dimensionali limitati 	<ul style="list-style-type: none"> • Difficoltà di interpretazione
 Decision Trees	<ul style="list-style-type: none"> • Facilità di interpretazione • Veloce nella fase di training 	<ul style="list-style-type: none"> • Vulnerabile all'overfitting
 Random Forest	<ul style="list-style-type: none"> • Correzione di overfitting generati nei semplici alberi decisionali 	<ul style="list-style-type: none"> • Difficoltà di interpretazione
 Support Vector Machines	<ul style="list-style-type: none"> • Facilità di interpretazione • Efficiente nella previsione di più classi 	<ul style="list-style-type: none"> • Lento nella fase di addestramento se si utilizza un set di dati grande
 Logistic Regression	<ul style="list-style-type: none"> • Facilità di interpretazione 	<ul style="list-style-type: none"> • Vulnerabile all'overfitting
 Neural Network	<ul style="list-style-type: none"> • Efficiente nella previsione di più classi 	<ul style="list-style-type: none"> • Difficoltà di interpretazione

Tabella 1.3.1 : Classificatori

La tabella presentata sarà utile in seguito ad una rassegna sulla letteratura di lavori correlati, per identificare la tipologia di classificatori presi in considerazione per il nostro studio.

1.3.2 Rassegna della letteratura

Di seguito viene presentata una rassegna della letteratura riguardante due studi in cui sono stati applicati algoritmi di Machine Learning finalizzati all'assegnazione delle priorità di intervento.

Nel primo studio Vincente García-Díaz et al. [33] hanno condotto uno studio per migliorare l'accuratezza dei classificatori probabilistici per sistemi di supporto decisionale. In particolare presentano una piattaforma (DSocialL) per automatizzare l'elaborazione delle informazioni ottenute dalle reti sociali, basata sull'utilizzo di un classificatore Bayesiano.

Lo studio è stato eseguito attraverso l'uso di un campione di 387 tweet sia con sentimenti positivi che negativi verso film. È stato applicato un classificatore Naive Bayes usando k-fold cross-validation (con k=10) ottenendo una precisione finale di 0,789.

In conclusione è stato visto che i classificatori probabilistici sono affidabili ma dipendono fortemente dalle dimensioni del set di dati di allenamento e da una corretta pre-elaborazione del testo iniziale.

Nel secondo studio, condotto da Yunjeong Mo et al.[34], vediamo la creazione di un modello di Machine Learning che legge i testi delle richieste di servizio e assegna automaticamente la forza lavoro e la priorità attraverso la tecnica di elaborazione del linguaggio naturale (NLP). Il set di dati utilizzati per la modellazione di questo metodo contiene 82.106 richieste manutentive degli edifici per un periodo di tre anni.

Nella preparazione dei dati, è stato diviso l'insieme dei dati in tre diversi set: Un set di addestramento (60%) , un set per lo sviluppo (20%) e uno per

la valutazione finale delle prestazioni (20%). Per la fase di apprendimento sono stati utilizzati tre algoritmi di apprendimento automatico: Logistic Regression (LR), Naive Bayes (NB) e la Support Vector Machines (SVM).

I risultati di questo studio mostrano un accuratezza massima del 77% per la previsione della forza lavoro e un accuratezza massima dell' 88% per la previsione della priorità utilizzando per entrambi l'algoritmo LR.

In riferimento agli studi descritti notiamo come le recenti applicazioni di Natural Language Processing (NLP) utilizzano algoritmi di ML per raggiungere prestazioni di classificazione più elevati.

Considerando la Tabella 1 sopra inserita ed i casi studio in riferimento al tema trattato nella tesi, sono stati scelti cinque tipologie di classificatori supervisionati: Naive Bayes Classifier, Support Vector Machine, Decision Trees, Neural Network e Logistic Regression.

Naive Bayes Classifier è una tipologia di classificatore di facile interpretazione con un'ottima velocità nella fase di training. Inoltre risulta efficiente nella previsione di più classi. Support Vector Machine attraverso l'individuazione di separatori lineari e polinomiali riesce a raggiungere livelli elevati di classificazione. Logistic regression risulta essere di facile interpretazione e raggiunge livelli elevati di accuratezza nella previsioni di più classi. Decision Trees risulta essere veloce nella fase di training e Neural Network grazie alla facilità di implementazione a più livelli risulta essere ottimo per la classificazione a più vie.

Questi algoritmi, oltre ai vantaggi appena descritti, risultano essere molto efficaci nella classificazione di tipo Bag-of-word (BOW) [34] [35] in cui un singolo testo, nel nostro caso una richiesta manutentiva, viene identificata

come un insieme di parole in cui l'ordine non ha importanza. Questa particolare tecnica sarà descritta nei capitoli successivi.

1.4 Conclusione sullo stato dell'arte

In riferimento ai precedenti paragrafi, si nota come l'assegnazione delle priorità di intervento nella manutenzione del patrimonio immobiliare risultino essere un problema abbastanza discusso.

In conclusione, considerando la rassegna sulla letteratura degli argomenti trattati, è possibile notare come lo sviluppo di modelli che utilizzano algoritmi di ML per la prioritizzazione delle richieste sono orientati al raggiungimento di un buon risultato e non a quali devono essere i criteri che influenzano lo stesso.

In vari studi abbiamo visto che il modo in cui viene elaborato il testo, prima che esso venga utilizzato per lo sviluppo dei modelli, influenza la bontà dei risultati finali.

È necessario quindi comprendere quali sono le variazioni dei risultati in riferimento all'applicazione di diversi algoritmi di ML e considerando diverse metodologie di elaborazione testuale.

2 FASI, MATERIALI E METODI

In questo capitolo vengono descritte in prima istanza le fasi che hanno condotto allo sviluppo del metodo. In seguito viene presentato il caso dell'Università Politecnica delle Marche, introducendo il sistema manutentivo presente all'interno di esso e descrivendo il set di dati raccolto e utilizzato per il nostro caso studio. Infine vengono presentate le metodologie impiegate per lo sviluppo del metodo.

2.1 Le fasi operative

La tesi in questione presenta cinque fasi principali:

- 1) *Raccolta delle richieste manutentive* : In questa fase sono state raccolte le richieste di intervento manutentivo del complesso dell'Università Politecnica delle Marche composto da 23 edifici. Si tratta di ca. 12.000 richieste manutentive raccolte nell'arco di 34 mesi direttamente estrapolate da testi di e-mail redatti da lavoratori del complesso;
- 2) *Human manual annotation (HMA annotation)*: In questa fase è stato effettuato un processo di annotazione manuale per l'assegnazione di priorità seguendo il Best-Worst-Scale (BWS) method (approfondito in seguito) ed è stato chiesto a 40 annotatori, appartenenti a tre classi di esperienza differente (limitata, media ed elevata), di effettuare una compilazione manuale attraverso questo metodo. In particolare sono stati considerati:
 - Annotatori con esperienza "*limitata*": Laureati/laureandi magistrali del campo delle costruzioni (Ingegneria Edile, Civile o Edile-Architettura);

- Annotatori con esperienza “*media*”: Ingegneri, architetti o geometri operanti nel campo delle costruzioni ;
- Annotatori con esperienza “*elevata*”: Personale operante nel campo specifico del FM;

Lo scopo di questa fase è stato quello di raccogliere gli score attribuiti per ogni sentenza così da essere utilizzati successivamente per la fase di training dei classificatori ML. Per lo sviluppo del modello sono state selezionate 149 richieste manutentive dal dataset totale (rappresentative del 91,3% della TF del dataset originale)

3) *Creazione delle Document Term Matrix (DTM) con assegnazione classi*

: In questa fase, eseguita con l’ambiente di sviluppo per l’analisi statistica dei dati R, è stato elaborato il testo attraverso 3 combinazioni differenti di pre-elaborazione testuale impiegati nella text-mining che in seguito saranno descritti nel dettaglio. I testi elaborati sono stati il punto di partenza per la creazione delle DTM. Queste sono particolari matrici che descrivono la frequenza dei termini presenti in un determinato documento (nel nostro caso richieste manutentive) dove le righe corrispondono al documento di riferimento e le colonne ai termini.

Dalle tre diverse metodologie di elaborazione testuale sono state ricavate tre differenti DTM:

- DTM_sel : Matrice in cui sono presenti solo i termini selezionati da annotatori esperti (sono state selezionate, per ciascuna sentenza, 2-3 parole significative nel rappresentare problema e rilevanza da annotatori esperti);
- DTM_ste : Matrice formata da generali processi di elaborazione testuale incluso lo stemming;
- DTM_noste : Matrice formata da processi di elaborazione testuale escluso lo stemming;

Inoltre in questa fase sono state assegnate le classi riferite ad ogni richiesta manutentiva aggiungendo una colonna al termine di ogni matrice. Ogni classe corrisponde ad un intervallo di score attribuito;

- 4) *Applicazione Classificatori*: In questa fase, attraverso l'utilizzo della piattaforma per l'analisi dei dati MATLAB(classification learned), sono stati impiegate le cinque tipologie di classificatori menzionate nei capitoli precedenti (Classificatore Naive Bayes, Support Vector Machine, Decision Trees, Neural Network e Logistic Regression) per ogni tipologia di DTM.
- 5) *Analisi risultati*: In questa fase sono state plottate le confusion matrix, quindi Accuracy Precision, Recall ed F1, sono stati impiegati per un'analisi dettagliata dei risultati (globali e riferiti a ciascuna classe). Questi, riferiti ad ogni singola DTM per ogni classificatore impiegato hanno permesso di confrontare i diversi metodi di elaborazione testuale e confrontare le scelte operative.

2.2 Materiali

2.2.1 Caso studio: le richieste manutentive all'interno dell'Università Politecnica delle Marche

L'Università Politecnica delle Marche è situata ad Ancona e comprende 23 edifici utilizzati per scopi didattici e amministrativi. Ospita circa 16.000 studenti e 1000 lavoratori.

La gestione dell'attività manutentiva all'interno del complesso dell'Università Politecnica delle Marche viene garantita da un ente esterno attraverso un contratto di outsourcing finalizzato ad assicurare un efficiente servizio manutentivo della struttura.

Le categorie di elementi a cui è soggetta la manutenzione sono: componenti dell'edificio, sistema di allarme, impianto elettrico, impianto di elevazione, impianto antincendio, HVAC ed impianto idrico.

L'appaltatore all'interno delle strutture svolge due tipologie di attività manutentive:

Manutenzione periodica (*planned*)

L'attività manutentiva viene garantita attraverso dei controlli (check) periodici riguardanti tutti gli apparati. Il servizio prevede una riparazione o sostituzione di un elemento ogni volta che vi è un'interruzione del normale funzionamento degli apparecchi.

Manutenzione a richiesta (*on demand*)

L'attività manutentiva, in questo caso, viene eseguita attraverso delle riparazioni ogni volta che vi è la rottura di un determinato elemento.

Nel caso della manutenzione *on demand* vengono elaborate delle richieste manutentive dagli utilizzatori dei servizi in cui si descrivono le caratteristiche del problema rilevato.

L'appaltatore riceve molteplici richieste manutentive e spesso risulta complicato gestire nel migliore dei modi gli interventi da effettuare a seguito di queste. Risulta quindi estremamente utile un servizio di prioritizzazione automatica che vada a garantire una gestione regolare del servizio manutentivo.

Nello schema seguente troviamo riassunto il servizio di gestione del caso studio analizzato.

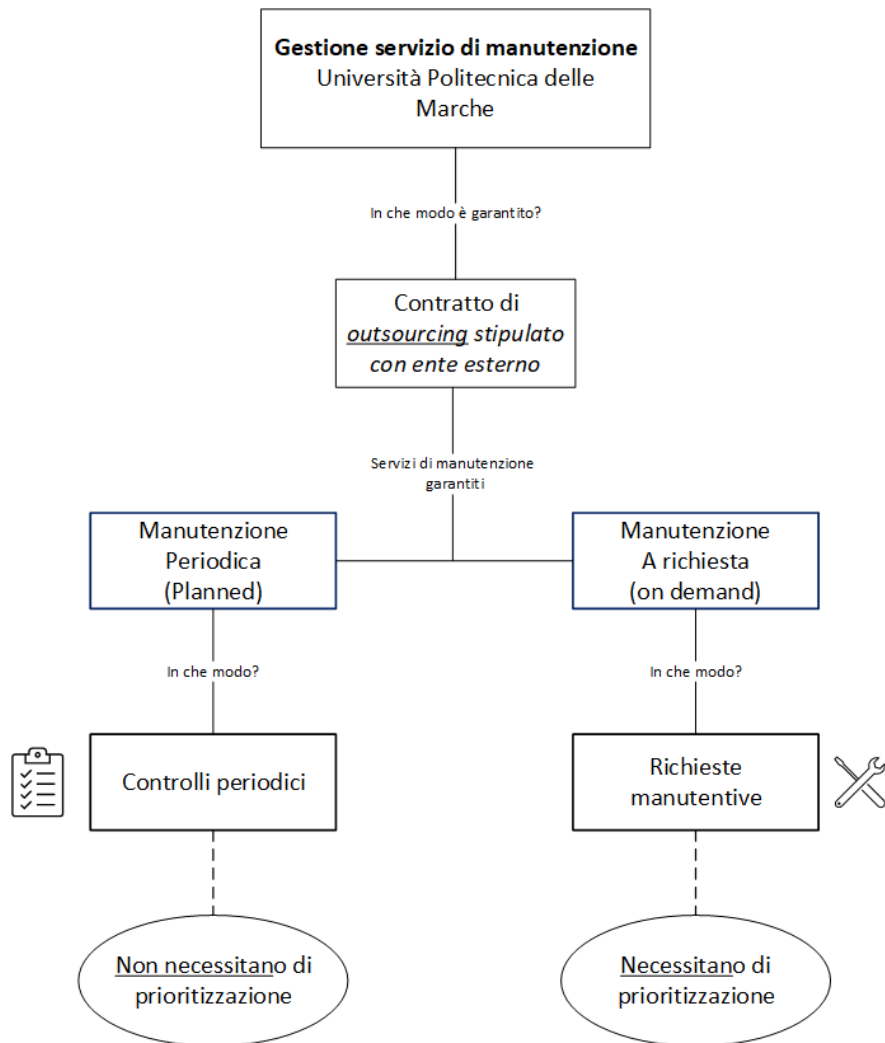


Figura 2.2.1: Servizio manutentivo UNIVPM

2.2.2 Analisi del set di dati raccolti

Il modello predittivo è stato sviluppato sulla base di un dataset formato da richieste di intervento manutentivo generate nell'arco di 34 mesi nei 23 edifici dell'Università Politecnica delle Marche.

Le richieste di manutenzione sviluppate dagli utilizzatori finali consistono in brevi testi inviati via e-mail ed elaborati dai tecnici dell'ente gestore della manutenzione.

Il dataset comprende comunicazioni su anomalie e guasti dei componenti dell'edificio . In particolare sono state individuate sette categorie di problematiche principali : impianto elettrico (illuminazione, sistemi di alimentazione, connessione LAN e WLAN), componenti dell'edificio (pareti, porte, finestre, controsoffitti, ecc.), HVAC (tubazioni di riscaldamento, ventilazione e raffreddamento), impianto idraulico, antincendio, sistema di allarme ed impianto di elevazione.

Di seguito vengono illustrate il numero di richieste manutentive appartenenti ad ogni categoria:

Categoria	Num. richieste
Componenti edificio	2255
Sistema di allarme	346
Elettrico	3229
Elevazione	357
Antincendio	505
HVAC	1971
Idraulico	1617

Tabella 2.2.1 : Richieste Manutentive

Di seguito vengono presentati 4 esempi di richieste manutentive presenti all'interno del dataset:

Edificio 2, q150 bagni, lavandino con perdita da rubinetto e sifone.

Manca corrente elettrica generale.

Si segnalano 3 o 4 neon non funzionanti siti lungo il corridoio di Farmacologia 3°p.polo Eustachio scala B

Si chiede il controllo dell'impianto di condizionamento della Biblioteca

Come è possibile notare, il testo delle richieste risulta essere strutturato secondo un ordine diverso a seconda di ogni sentenza e spesso vengono indicate informazioni di poca importanza considerando lo scopo di dare priorità a queste.

Lo studio in questione cerca quindi di elaborare il testo delle sentenze, al fine di individuare il gruppo di parole che più hanno rappresentanza della richiesta manutentiva.

Per lo sviluppo del sistema automatico sono state estratte 149 richieste manutentive ed essendo redatte in lingua inglese sono state tradotte anche in lingua italiana.

2.3 Metodi

2.3.1 HMA annotation

2.3.1.1 *Best-Worst Scaling (BWS)*

Nel caso studio presentato in questa tesi l'annotazione manuale è stata una fase di fondamentale importanza per ricavare gli score di riferimento di ogni singola sentenza. Mediante l'associazione di ciascuna richiesta manutentiva con un relativo punteggio è stato possibile ricavare i dati di partenza (dati di training) per l'applicazione dei classificatori Machine Learning. Questi infatti necessitano di dati di training per essere addestrati.

A tale proposito è stato sviluppato un modello di annotazione manuale basato sul metodo Best-Worst Scaling (BWS)[36][37] . Questo risulta essere un metodo efficace rispetto ad altri approcci di annotazione proposti.

In questo metodo agli annotatori vengono presentati n -elementi alla volta (dove $n > 1$ e tipicamente $n = 4$) dove viene chiesto loro di indicare quale elemento sia il migliore (il più alto secondo la propria percezione) e quale sia il peggiore (il più basso secondo la propria percezione). Tutti gli elementi soggetti a valutazione sono organizzati in un insieme di m 4-tuple ($m \geq N$, dove N è il numero di elementi) in modo che ogni elemento sia valutato più volte in diverse 4-tuple.

In questo studio, agli annotatori sono state presentate diverse quartine (4-tuple) ed è stato chiesto loro di selezionare il caso migliore e quello peggiore (meno urgente e più urgente). Ogni quartina è costituita da un sottoinsieme di sentenze estratte dal dataset di partenza. In totale sono state ricavate 150 quartine così da avere ogni frase ripetuta 5 volte.

Di seguito vediamo un esempio di quartina presentata agli annotatori.

Num	NOTE	CODE	BEST	WORST
1	polo monte dago palazzina prove materiali e strutture ruota 155 bagno uomini perdita idrica	11		1
2	edificio 2 q150 bagni, lavandino con perdita da rubinetto e sifone	17	1	
3	Vi segnaliamo che nel bagno uomini di fronte scuole di specializzazione il primo lavandino è otturato.	107		
4	Pulizia corpi illuminanti locali ex Segreteria Studenti	113		

WORST=Caso più urgente
BEST=Caso meno urgente

Figura 2.3.1 Estratto quartina

Nella quartina mostrata in figura notiamo che per la valutazione del caso peggiore e migliore (più urgente e meno urgente), è stato chiesto agli annotatori di indicare la propria valutazione assegnando un “1” dove ritenevano più opportuno.

Infine lo score di ogni sentenza è stato ricavato calcolando, come indicato in [36]:

$$Score = \frac{Somma\ di\ Best - Somma\ di\ Worst}{N\ ripetizioni\ della\ sentenza} \quad (1)$$

Quindi sottraendo al numero di volte che la sentenza è stata valutata come migliore, il numero di volte risultata peggiore e dividendo per quante volte è stata ripetuta la sentenza.

Gli annotatori sono stati distinti in tre gruppi differenti, ognuno avente diverso livello di esperienza:

- Annotatori con esperienza “*limitata*”: Laureati/laureandi magistrali del campo delle costruzioni (Ingegneria Edile, Civile o Edile-Architettura);
- Annotatori con esperienza “*media*”: Ingegneri, architetti o geometri ;
- Annotatori con esperienza “*elevata*”: Personale operante nel campo del FM;

In totale sono state raccolte 49 annotazioni appartenenti a 40 annotatori differenti. A 9 di loro è stato chiesto di compilare il modello sia in lingua inglese che in lingua italiana.

Di seguito è illustrato uno schema riassuntivo in cui vengono indicati i dettagli degli annotatori che hanno preso parte alla compilazione:

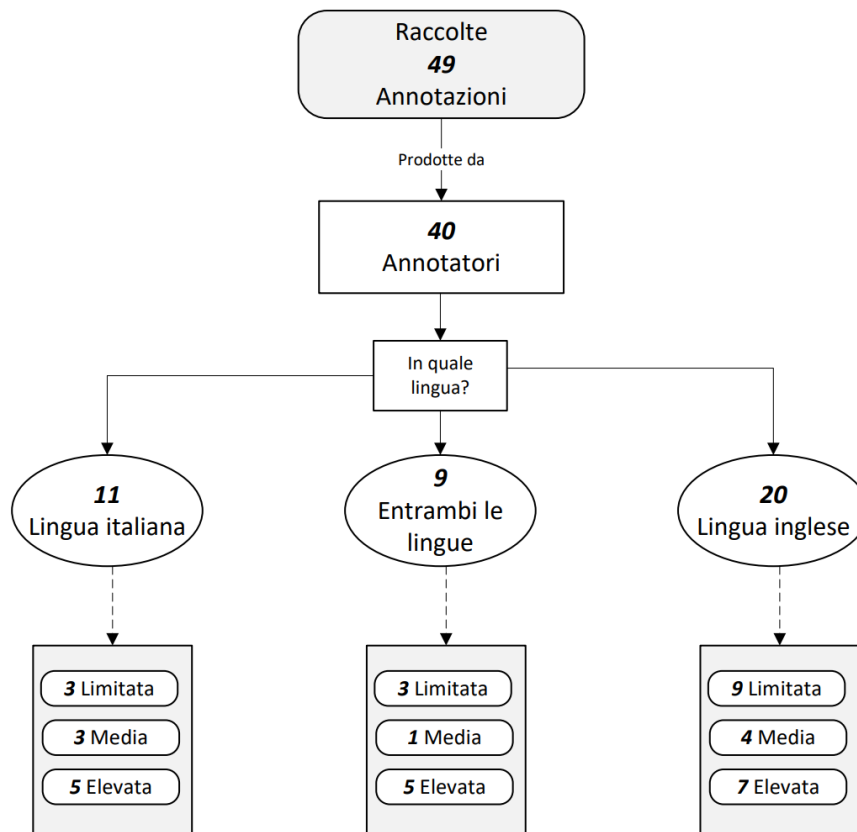


Figura 2.3.2: Annotazioni raccolte

Dopo aver calcolato gli score ricavati da ogni annotazione, questi sono stati organizzati in un file Excel indicando per ogni sentenza il codice identificativo, il testo in lingua inglese, il testo in lingua italiana e il rispettivo punteggio.

Tale file è stato utilizzato per l'assegnazione delle classi, indicative di quattro livelli di urgenza differenti, successivamente alla creazione delle DTM attraverso l'ambiente *R studio*.

Di seguito vediamo un estratto della tabella utilizzata per la raccolta degli score ricavati da ogni annotazione:

Code	testi italiano	testi inglesi	1	2	3	4
			A	A	B	B
			italiano	inglese	italiano	inglese
			Limitata	Limitata	Limitata	Limitata
11	polo monte dago palazzina prove materiali e strutture ruota 155 bagno uomini perdita idrica	polo monte dago building tests materials and structures wheelq155 men's bathroom water loss	-0,2	0		-0,6
17	edificio 2 q150 bagni, lavandino con perdita da rubinetto e sifone	dises - bathroom p002_029 - near scale 1 - sink faucet where no water flows	0,2	0,2		0,6
107	Vi segnaliamo che nel bagno uomini di fronte scuole di specializzazione il primo lavandino è otturato.	we would like to point out that in the men's room in front of the graduate school the first sink is blocked.	0	0,4		0
113	Pulizia corpi illuminanti locali ex Segreteria Studenti	cleaning of local lighting fixtures ex student secretariat	0,4	0,8		1
171	ripristinare il funzionamento dei lettori installati presso il Core Facility della Facoltà di Medicina (M09 e M10) i quali non hanno attive le comunicazioni con la rete internet.	restore operation of the readers installed at the core facility of the faculty of medicine (m09 and m10) which do not have active communications with the internet.	-0,4	-0,6		-0,6
174	Si segnala plafoniera spenta nella biblioteca del centro EBI n*tel 220-6277 - riferimento dr.ssa Scrami .	please note that the ceiling light is off in the ebi center library n tel 220-6277 - reference dr. scrami.	0,2	-0,2		0,4
277	edificio 2/4 q155 SIMAU locale n. 174 lavabo con tubi scarico scollegati riferimento Mara Formica	b2 building 2-4 q155 simau local no. 174 washbasin with disconnected drain pipes reference mara formica	-0,4	-0,2		0,6
309	edificio PMS Q132 OFFICINA finestra vasistas guasta - riferimento Stefano Bugarini	pms building q132 workshop transom window broken - reference stefano bufarini	0,8	-0,4		-0,2
426	Si chiede il ripristino dell'erogazione dell'energia elettrica (LAVORO GIA' ESEGUITO) nello stabulario del polo Eustachio.	the restoration of the electricity supply (work already executed) in the eustachian pole enclosure is requested.	1	0		-0,4
678	Si segnalano 3 o 4 neon non funzionanti siti lungo il corridoio di Farmacologia 3°p.polo Eustachio scala B	there are 3 or 4 non-functioning neon sites along the pharmacology corridor 3 eustachian pole scale b	0,8	0,6		0,2

Figura 2.3.3: Estratto tabella annotazioni

2.3.1.2 *L'Alpha di Krippendorff*

L'Alpha di Krippendorff misura il grado di accordo tra un insieme di dati. Nel nostro caso il suo impiego è stato necessario per verificare la coerenza delle annotazioni appartenenti ai tre gruppi distinti (esperienza limitata, media ed elevata)[38].

Il coefficiente è applicabile a qualsiasi dimensione del campione.

Gli intervalli che identificano il grado di accordo tra l'insieme dei dati sono:

$\alpha \geq 0.800$ *Elevato*

$0.667 > \alpha \geq 0.667$ *Medio*

$\alpha < 0.667$ *Basso*

2.3.1.3 *Test di Spearman*

Il test di Spearman consente di individuare il grado di relazione tra due variabili attraverso l'indice di correlazione (*rho*). Diversamente dal test di Pearson questo è non parametrico. Nel nostro caso è stato utilizzato per l'individuazione delle annotazioni che più si discostavano dal risultato medio degli score ricavati[39].

Il coefficiente di correlazione di Spearman assume valori tra -1 e +1 indicando nel segno e nel valore il tipo e la forza della correlazione. Il segno positivo indica una correlazione direttamente proporzionale mentre il segno negativo indica una correlazione inversamente proporzionale. Quindi:

$-1 < rho < 0$ *correlazione inversamente proporzionale*

$\rho = 0$ *correlazione nulla*

$0 < \rho < 1$ *correlazione direttamente proporzionale*

$\rho = \pm 1$ *correlazione perfetta*

2.3.2 Creazione DTM

In questo paragrafo vengono descritti i processi di elaborazione testuale che hanno condotto alla creazione delle *Document Term Matrix (DTM)* utilizzate in seguito per l'applicazione dei classificatori.

Ogni DTM è stata generata attraverso tre combinazioni di processi testuali differenti. In particolare:

- DTM_sel : selezione di parole da parte di annotatori esperti;stemming;
- DTM_ste : rimozione di numeri e punteggiatura, sostituzione dei sinonimi con parola di riferimento, rimozione stop-words, stemming e pos tagging;
- DTM_noste : rimozione di numeri e punteggiatura, sostituzione dei sinonimi con parola di riferimento, rimozione stop-words e pos tagging;

I processi di pre-elaborazione testuale sono stati effettuati sui testi in lingua italiana.

Di seguito è mostrata una tabella riassuntiva in cui sono indicati i processi di elaborazione testuale impiegati per ogni DTM. Nei paragrafi successivi sono descritti nel dettaglio.

Processo di elaborazione	Matrice con parole selezionate DTM_sel	Matrice con incluso stemming DTM_ste	Matrice con escluso stemming DTM_noste
Selezione parole	✓		
Rimozione numeri e punteggiatura		✓	✓
Sostituzione sinonimi		✓	✓
Stop-words		✓	✓
Stemming	✓	✓	
Pos tagging		✓	✓

Tabella 2.3.1: Processi di Elaborazione testuale

Successivamente alla creazione delle DTM, generate dai termini rimanenti dei processi di analisi testuale, sono state assegnate le classi aggiungendo una colonna alle DTM con gli score ricavati dagli annotatori.

In particolare, considerando la non perfetta concordanza dei risultati degli score attribuiti dai diversi annotatori alle sentenze, non è stata effettuata la media tra di loro ma sono state replicate le frasi, con relativo punteggio, tante volte quanti sono stati gli annotatori.

Il confronto delle tre differenti metodologie impiegate per la pulizia del testo sono state di fondamentale aiuto per individuare i punti fondamentali che incidono sulla bontà dei risultati di un classificatore.

Successivamente vengono descritte le tecniche di elaborazione testuale impiegate:

2.3.2.1 Stemming

Lo stemming è il processo tramite il quale viene indentificata la radice delle parole trasformandola nella sua forma flessa e più generale. Il termine stemming deriva dall'inglese stem ossia stelo. Si considera la radice di una parola come uno stelo da cui si diramano tutte le parole varianti. Nello stemming la radice è un elemento linguistico irriducibile[40].

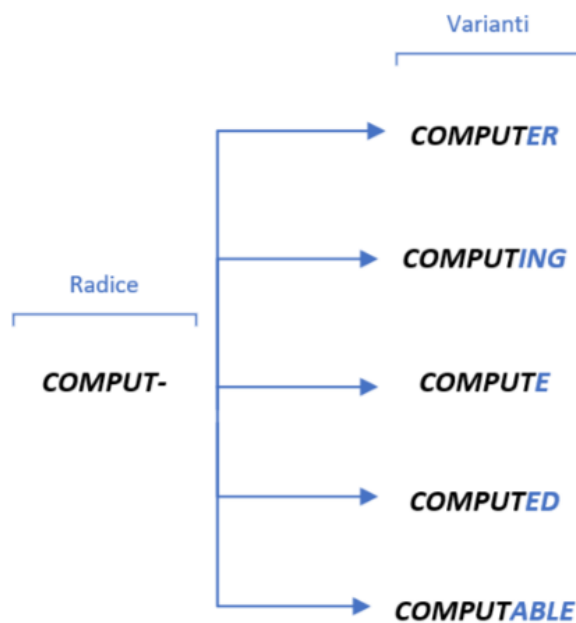


Figura 2.3.4 Stemming

Nel processo di stemming la riduzione a radice può avvenire per diversi gradi. È bene optare quindi per gradi bassi scegliendo una radice lunga in quanto, scegliendone una corta aumenterebbero le famiglie di variabili che sono composte da quella radice.

All'interno di un testo questo processo permette di ottenerne uno nuovo con lo stesso numero di termini ma con meno varianti.

2.3.2.2 *Part-of-Speech tagging (PoS tagging)*

Il processo di POS tagging consiste nell'assegnare un tag a ciascuna parola di un documento. Il tag definisce la *classe* della parola fornendo informazioni per determinare il ruolo della parola stessa e quindi scartare quelle che non sono di nostro interesse[22].

In genere le classi considerate possono essere:

Chiuse: classi che hanno membri fissi, dove a determinati termini non vi è la possibilità di aggiungerne di altri con la stessa funzione logica testuale (articoli, pronomi, preposizioni e congiunzioni).

Aperte: classi soggette all'aggiunta di nuovi termini (sostantivi, verbi, avverbi, aggettivi).

Negli anni sono stati molti i progetti che hanno visto la creazione di tagset utilizzati per l'assegnazione delle classi in un documento. Di seguito vengono presentati i tag del progetto Penn Treebank.

CC	Coordinating conj.	TO	infinitival <i>to</i>
CD	Cardinal number	UH	Interjection
DT	Determiner	VB	Verb, base form
EX	Existential there	VBD	Verb, past tense
FW	Foreign word	VBG	Verb, gerund/present pple
IN	Preposition	VBN	Verb, past participle
JJ	Adjective	VBP	Verb, non-3rd ps. sg. present
JJR	Adjective, comparative	VBZ	Verb, 3rd ps. sg. present
JJS	Adjective, superlative	WDT	Wh-determiner
LS	List item marker	WP	Wh-pronoun
MD	Modal	WPS	Possessive <i>wh</i> -pronoun
NN	Noun, singular or mass	WRB	Wh-adverb
NNS	Noun, plural	#	Pound sign
NNP	Proper noun, singular	\$	Dollar sign
NNPS	Proper noun, plural	.	Sentence-final punctuation
PDT	Predeterminer	,	Comma
POS	Possessive ending	:	Colon, semi-colon
PRP	Personal pronoun	(Left bracket character
PPS	Possessive pronoun)	Right bracket character
RB	Adverb	"	Straight double quote
RBR	Adverb, comparative	'	Left open single quote
RBS	Adverb, superlative	“	Left open double quote
RP	Particle	'	Right close single quote
SYM	Symbol	”	Right close double quote

Figura 2.3.5 Tagset progetto Penn Treebank[4]

Nel nostro caso studio, questo processo è stato utile per rimuovere spazi congiunzioni e articoli. In particolare è stato eseguito con la funzione *spacy_parse()*.

2.3.2.3 Stop-words

Le stop-words sono delle parole comuni che in un testo non contribuiscono ad arricchirne il significato. Queste possono variare dalla lingua in cui è scritto un testo. In lingua italiana gli articoli, le proposizioni, le congiunzioni o gli aggettivi sono considerate stop-words. Nella seguente tabella vediamo alcune di queste:

a	ben	ecco	qua	sembra	sui	tempo
ad	buono	fra	quale	senza	sulla	tra
alla	che	gli	quasi	si	suo	vostr

Tabella 2.3.3 Stop-words

Nell'applicazione di algoritmi di analisi testuale le stop-words consumano notevoli risorse computazionali (spazio di memoria, tempo di esecuzione dell'algoritmo) non apportando alcun vantaggio. Considerando l'applicazione di un algoritmo di ML con lo scopo di classificare in più categorie un testo, la presenza di queste parole contribuisce negativamente nella iniziale fase di training . Infatti l'algoritmo trovando queste ripetuti in più frasi appartenenti a classi differenti, tenderà a dare un'errata valutazione.

Tuttavia, l'eliminazione delle stop-words deve essere eseguita in modo prudente in quanto potrebbero essere eliminate parole che contribuiscono a dare un valore semantico alle frasi. Rimane importante quindi verificare quali liste di queste vengono applicate ad un testo.

Nel nostro caso le stop-words presenti all'interno delle richieste manutentive sono state rimosse attraverso la funzione `removeWord()` del package `tm` presente nell'ambiente *R studio*. La lista delle stop-words è stata arricchita con parole individuate all'interno del dataset come nomi di persona, nomi di luoghi e parole non ritenute opportune per il training dell'algoritmo.

2.3.2.4 *Applicazione dei sinonimi*

In un normale testo vengono impiegate diverse parole che hanno lo stesso significato. Sostituire queste con una singola parola di riferimento all'interno di un testo facilita notevolmente l'applicazione dell'algoritmo da impiegare.

A tale proposito per l'analisi delle richieste manutentive sono state individuate e sostituite delle parole con dei sinonimi rendendo il testo più semplice.

Di seguito vengono presentate una lista di parole che sono state sostituite dal vocabolo di riferimento:

Parole sostituite	Parola di riferimento
forza motrice, corrente, corrente elettrica, energia, energia elettrica, rete elettrica, alimentazione elettrica	elettricità
vasistas, persiana, lucernari	finestra
calare, far scendere, ribassare, tirare giù, sbassare	abbassare

Tabella 2.3.4 Sinonimi

2.3.2.5 *Il modello Bag of Word*

Nel modello Bag of Word ogni documento (sentenza) viene rappresentato da un gruppo di parole ignorando l'ordine di queste. L'insieme dei gruppi di parole, appartenenti a più documenti, vengono trasformati in vettori di lunghezza fissa indicando quante volte ogni parola appare nel documento[22].

Prendendo in esempio tre sentenze :

- 1) *Urgente Riparare condotto ventilazione*
- 2) *Urgente riparare tubazioni*
- 3) *Riparare condotto*

Individuando il gruppo di parole da cui è composta ogni sentenza e trasformandoli in vettori di uguale lunghezza otteniamo:

- | | |
|--|---------------|
| 1) <i>Urgente Riparare condotto ventilazione</i> | $[1,1,1,1,0]$ |
| 2) <i>Urgente riparare tubazioni</i> | $[1,1,0,0,1]$ |
| 3) <i>Riparare condotto</i> | $[0,1,1,0,0]$ |

Unendo questi vettori ricaviamo una matrice in cui sono indicati il numero di ripetizioni di ogni singolo termine appartenente a ogni documento (Document Term Matrix):

Sent.	urgente	riparare	condotto	ventilazione	tubazioni
Sent.1	1	1	1	1	0
Sent.2	1	1	0	0	1
Sent.3	0	1	1	0	0

Tabella 2.3.5 DTM esempio

Tale matrice è applicabile ad algoritmi di Machine Learning permettendo un confronto tra diversi documenti.

Nel nostro caso, successivamente ai processi di elaborazione testuale descritti in precedenza, questo è stato l'ultimo processo che ha portato alla creazione delle DTM impiegate per l'applicazione degli algoritmi.

2.3.3 Applicazione Classificatori

In seguito alla creazione delle DTM e dell' assegnazione delle classi sono stati applicati i classificatori.

In particolare ogni DTM generata da 3 diverse combinazioni di processi testuali è stata applicata ai cinque classificatori di Machine Learning precedentemente descritti (Naive Bayes Classifier, Support Vector Machine ,Decision Tree, Neural Network e Logistic Regression). La scelta della tipologia della distribuzione impiegata è stata eseguita confrontando i risultati ottenuti da quelle disponibili per ogni classificatore, optando per quella con i risultati migliori.

Attraverso l'applicazione "*Classification Learned*" presente nella piattaforma *MATLAB* è stato possibile estrapolare le confusion matrix con le quali è stato possibile confrontare il comportamento dei classificatori al variare delle tre metodologie impiegate

In seguito vediamo uno schema riassuntivo che descrive in breve le metodologie di analisi per ogni DTM con indicati la tipologia di classificatore impiegato e le distribuzioni disponibili per ognuno.

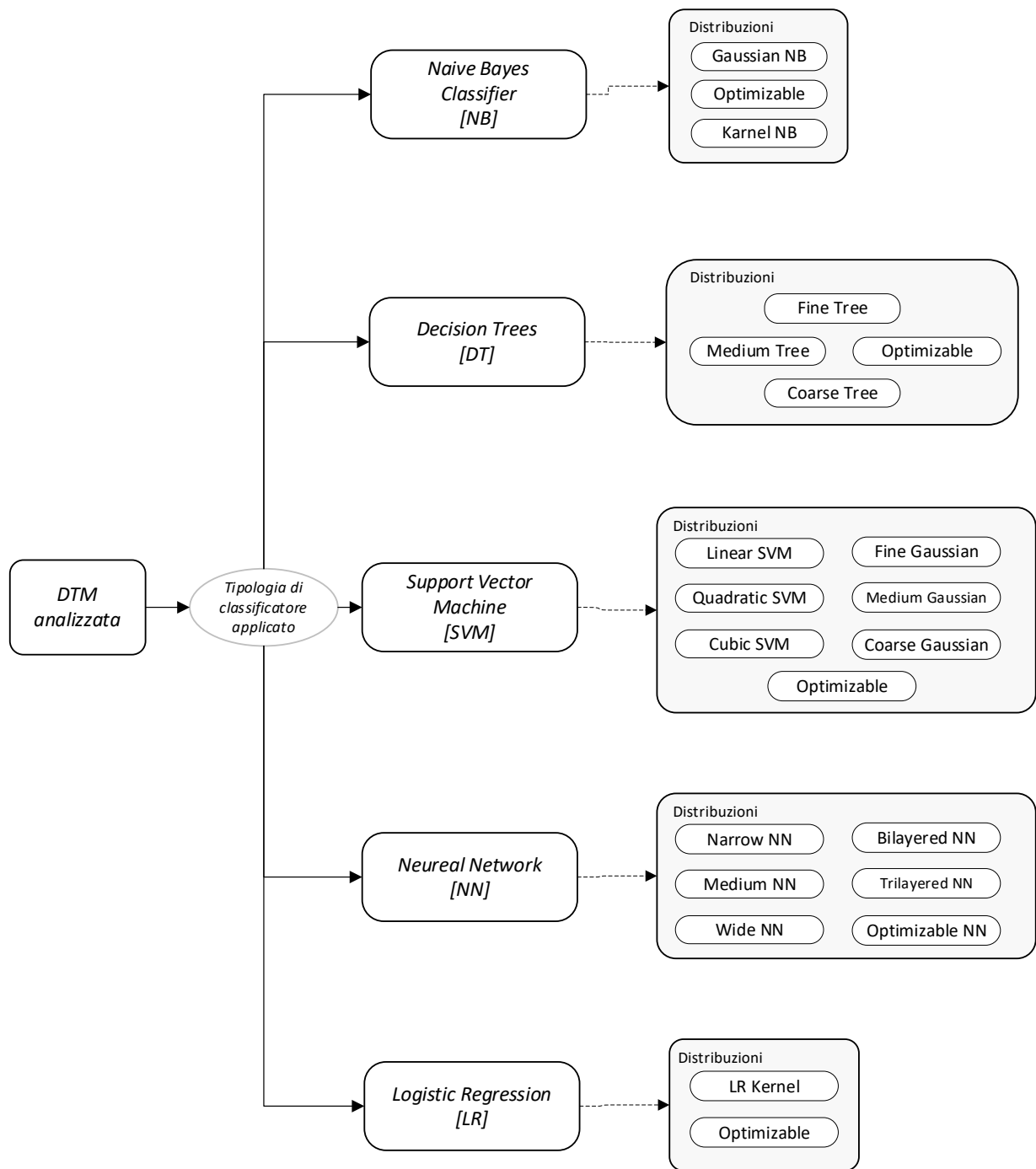


Figura 2.3.5 Applicazione classificatore

2.3.4 Analisi dei risultati

In questo paragrafo vengono presentati i metodi di valutazione e confronto delle diverse metodologie applicate. In particolare viene introdotta la *confusion matrix* e vengono descritti i coefficienti impiegati per la misura della bontà dei classificatori.

2.3.4.1 Confusion Matrix

La *matrice di confusione* (*confusion matrix*) permette di analizzare gli errori effettuati da un modello di Machine Learning. In particolare possono essere rilevati i campi dove il modello sbaglia maggiormente una predizione[41].

Considerando una classificazione binaria:

		Classi previste	
		Classe 1 (negativo)	Classe 2 (positivo)
Classi reali	Classe 1 (negativo)	True Negative	False Positive
	Classe 2 (positivo)	False Negative	True Positive

Figura 2.3.6 Confusion Matrix

Notiamo che nella diagonale vengono individuate le risposte corrette, quindi il caso in cui la classe predetta risulta essere quella reale. Negli spazi esterni alla diagonale vengono individuate le risposte in cui il modello effettua degli

errori. Questo è il caso in cui il modello predice una classe che non corrisponde a quella reale.

Questa particolare matrice può essere impiegata per la predizioni di più classi come mostrato in figura:

		Classi previste		
		Classe 1	Classe 2	Classe 3
Classi reali	Classe 1	True (Classe 1)	False (Classe 2)	False (Classe 3)
	Classe 2	False (Classe 1)	True (Classe 2)	False (Classe 3)
	Classe 3	False (Classe 1)	False (Classe 2)	True (Classe 3)

Figura 2.3.7 Confusion Matrix

Attraverso l'analisi di queste matrici è possibile calcolare dei coefficienti utilizzabili per ricavare la bontà dei classificatori che sono descritti in seguito:

2.3.4.2 Accuracy

L'*accuratezza* (*accuracy*) misura la percentuale di classificazioni corrette[41]. Questo coefficiente viene calcolato come segue:

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total}$$

Dove per True Positive (TP) si intende il numero veri positivi e per True Negative (TN) il numero di veri negativi.

2.3.4.3 Precision

La *precisione* (*precision*) misura la percentuale di positivi trovati[41]. Viene calcolata nel modo seguente:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Dove per False Positive(FP) viene indicato il numero di falsi positivi

2.3.4.4 Recall

Il *richiamo* (*recall*) indica la percentuale di positivi riconosciuti correttamente [41]. Viene calcolata come segue:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

2.3.4.5 F1 Score

Il *punteggio F1* (*F1 Score*) è la media armonica tra le *Precision* il *Recall* [41]. In particolare:

$$F1\ Score = \frac{2\ Precision * Recall}{Precision + Recall}$$

3 RISULTATI

3.1 HMA annotation

3.1.1 Verifica di correlazione tra annotazioni

In questo paragrafo si illustrano i risultati delle analisi di correlazione ottenuti confrontando gli score ricavati dalle *Human Manual Annotation (HMA annotation)* dei diversi annotatori. In particolare l'analisi si è svolta mettendo a confronto due gruppi principali di annotazioni (in lingua italiana ed in lingua inglese) e dal confronto dei gruppi appartenenti ai tre livelli di esperienza (limitata, media ed elevata). Lo scopo di questa analisi è stato quello di selezionare le annotazioni che più si discostavano del risultato medio e successivamente scartarle per l'ottenimento di un migliore risultato del modello.

Il grado di relazione tra i diversi dati è stato misurato attraverso l'indice di correlazione di *Spearman* descritto in precedenza.

Di seguito vengono mostrate delle immagini in cui sono presenti i grafici di correlazione con relativo coefficiente. L'analisi è stata svolta in prima istanza confrontando i gruppi di annotatori appartenenti ai diversi livelli di esperienza e successivamente confrontando gli score delle annotazioni ottenute in lingua inglese e in lingua italiana.

3.1.1.1 Annotazioni complessive

Num. Annotazione	Coefficiente di Spearman	Num. Annotazione2	Coefficiente di Spearman2
1	0,72	26	0,82
2	0,76	27	0,8
3	-0,58	28	0,71
4	0,18	29	0,79
5	0,64	30	0,76
6	0,81	31	0,78
7	0,79	32	0,69
8	0,81	33	0,86
9	0,8	34	0,054
10	0,82	35	0,69
11	0,8	36	0,69
12	0,79	37	0,79
13	0,73	38	0,74
14	0,068	39	0,79
15	0,87	40	0,77
16	0,7	41	0,8
17	0,77	42	0,74
18	0,74	43	-0,82
19	0,85	44	-0,81
20	0,,44	45	0,79
21	0,77	46	0,84
22	0,82	47	0,73
23	0,77	48	0,81
24	0,77	49	0,81
25	0,84		

Figura 3.1.1: Coefficiente di Spearman. Annotatori complessivi

Nella figura rappresentata si mostrano i coefficienti di *Spearman* ottenuti dalla relazione di 49 annotazioni e la media complessiva. Si nota come molti di questi assumano un valore non elevato (< 0.8) e a volte negativo.

Le annotazioni con tale caratteristiche risultano essere in disaccordo con la media dei valori complessivi.

In seguito vengono mostrati i risultati di analisi confrontando i diversi gruppi di annotatori con esperienza differente.

3.1.1.2 Annotatori con esperienza elevata

In questo paragrafo vengono illustrati i risultati ottenuti ponendo in relazione gli score ricavati dalle annotazioni eseguite da persone con esperienza elevata e la media effettuata ricavata da tali dati.

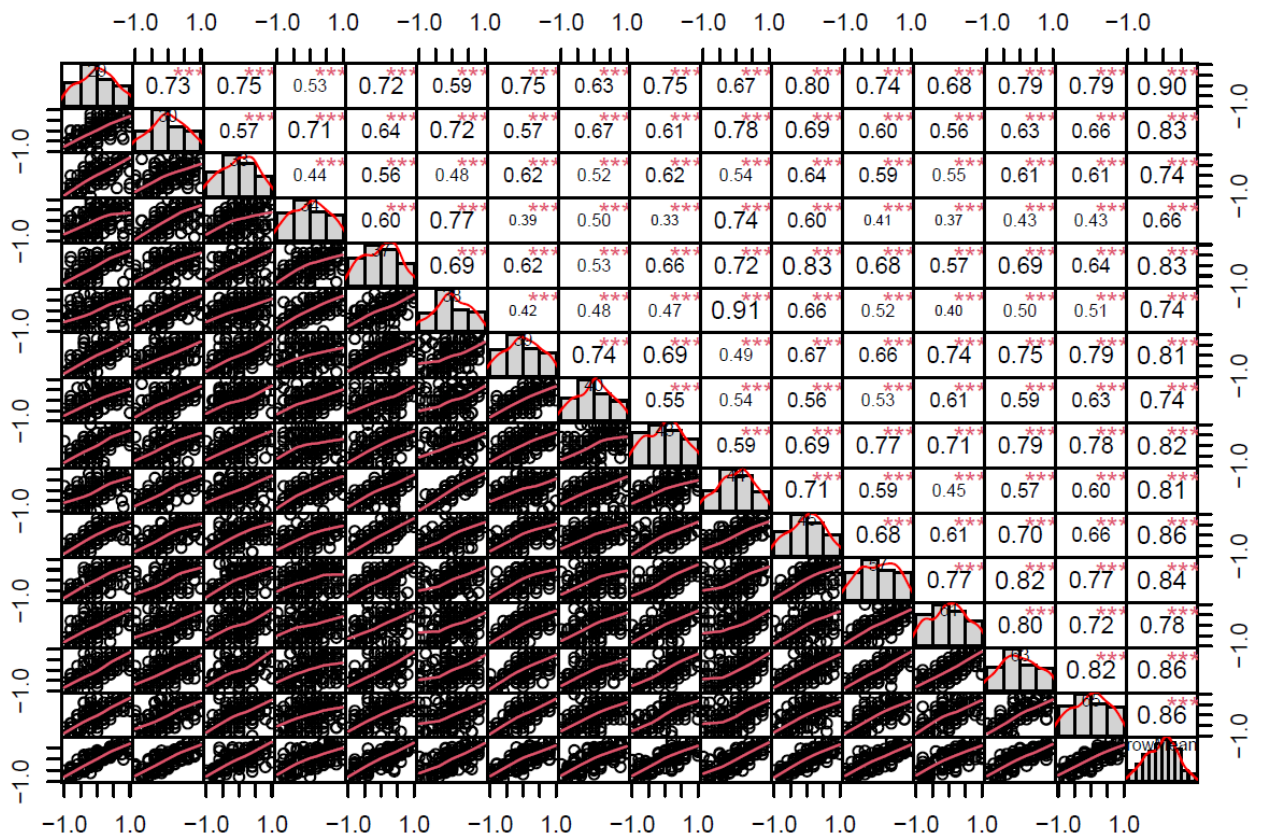


Figura 3.1.2 Coefficiente di Spearman. Esperienza elevata

I risultati presenti in figura sono comprensivi dei risultati delle annotazioni sia in lingua inglese che in lingua italiana.

Osservando l'ultima colonna, è possibile notare come molti coefficienti risultano essere non molto elevati e quindi in disaccordo con l'andamento medio complessivo.

3.1.1.3 Annotatori con esperienza media

Di seguito si illustrano i risultati ottenuti ponendo in relazione gli score di annotazioni effettuate da annotatori con esperienza media e la loro media.

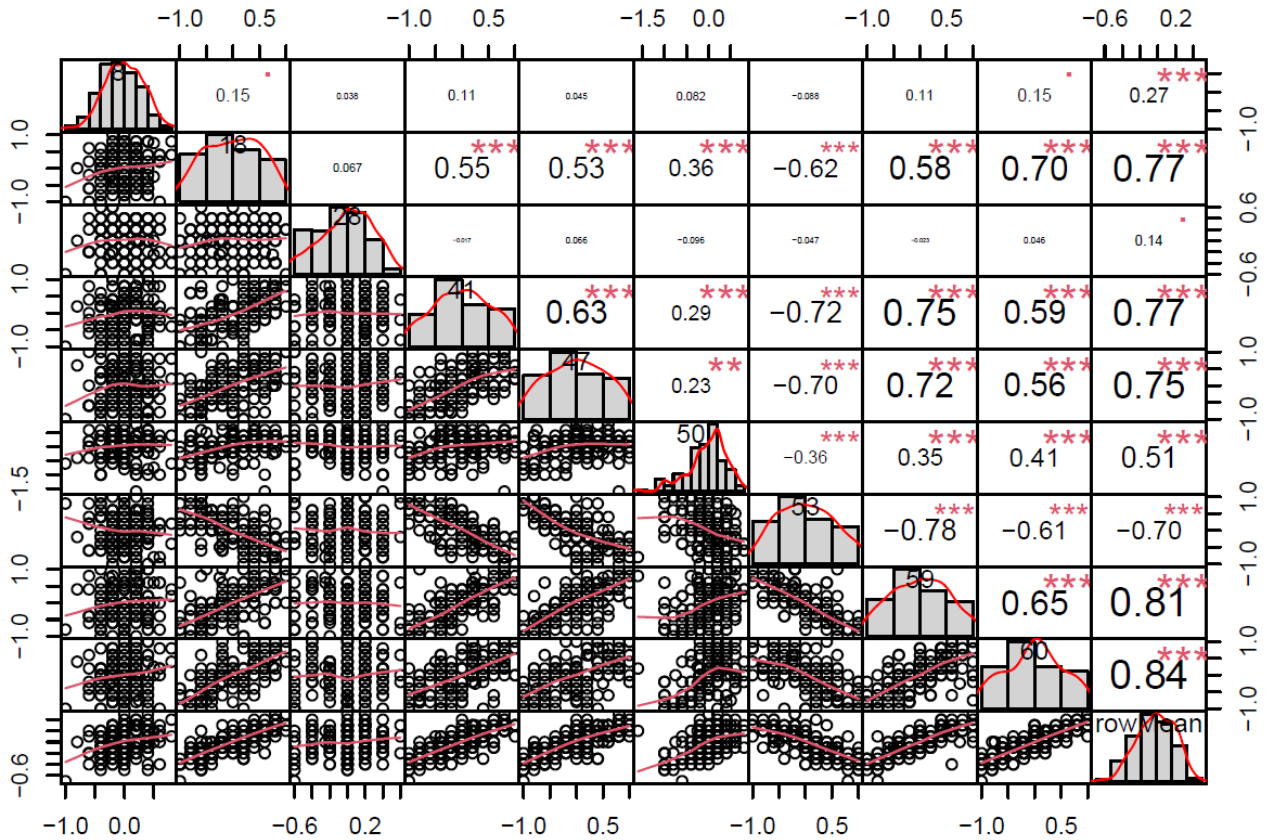


Figura 3.1.3 Coefficiente di Spearman. Esperienza media

Il risultato del confronto degli score delle 9 annotazioni rappresentato in figura mostra come gran parte di loro assuma valori < 0.8 e a volte negativi.

3.1.1.4 Annotatori con esperienza limitata

In questo paragrafo si mostrano i risultati ottenuti ponendo in relazione i dati ricavati da annotazioni effettuate da persone con esperienza limitata e la loro media. Anche in questo caso, nell'ultima colonna, è possibile notare la presenza di valori <0.8 e talvolta negativi.

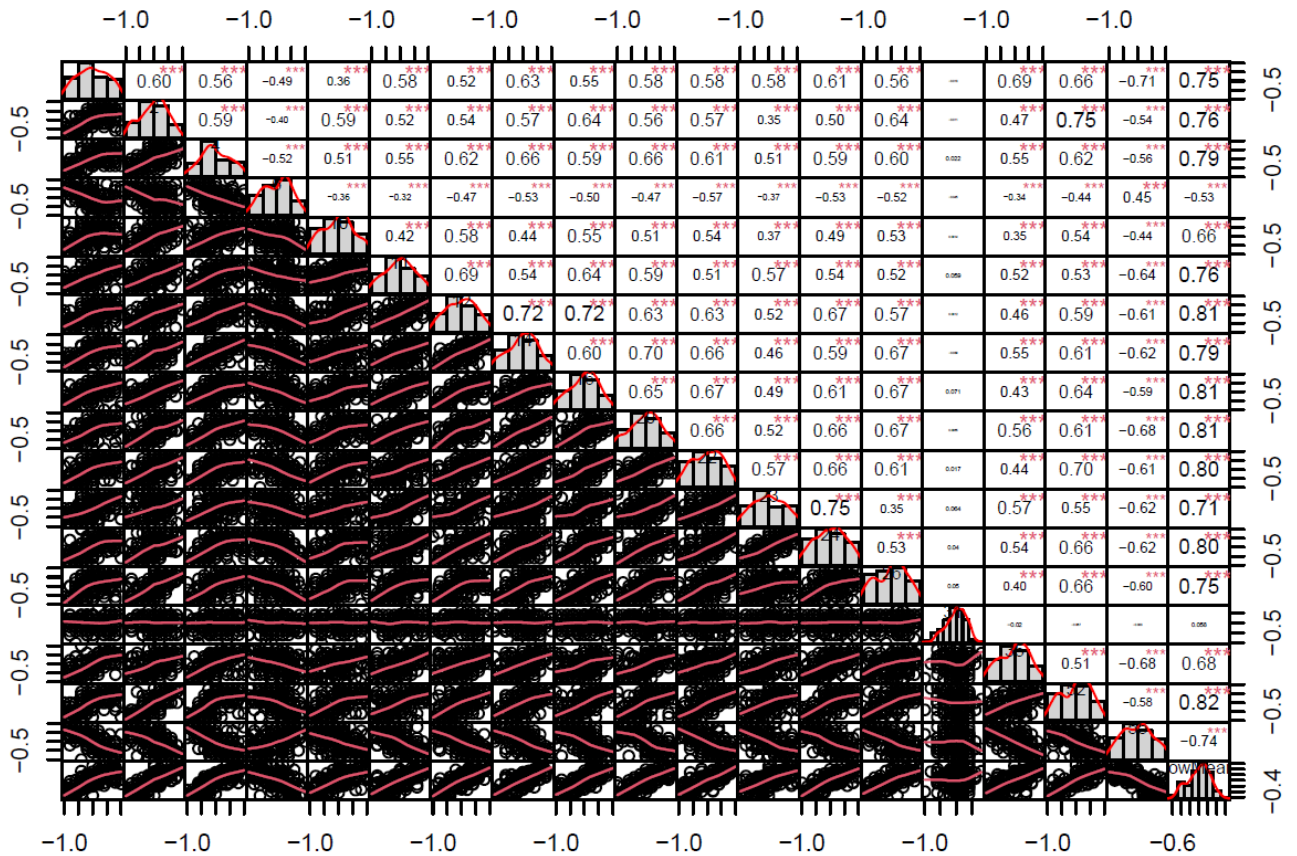


Figura 3.1.4 Coefficiente di Spearman. Esperienza limitata

Osservando i coefficienti di Spearman dei diversi gruppi con differenti livelli di esperienza si nota come il gruppo con esperienza limitata sia quello con i più bassi valori.

Questo dato risulta essere dettato dalla maggiore consapevolezza di attribuzione di grado di urgenza da parte degli annotatori con esperienza elevata.

Successivamente alla verifica da parte delle annotazioni effettuate da persone appartenenti a diversi livelli di esperienza sono stati analizzati i dati appartenenti alle annotazioni effettuate in lingua italiana e in lingua inglese.

Attraverso questa analisi, mettendo insieme gli annotatori di ogni livello di esperienza, sono stati individuati quelli che più si discostavano dal valore medio e sono stati eliminati. In particolare sono state scartate 20 annotazioni.

Di seguito si mostrano le relative immagini.

3.1.1.5 Annotazioni in lingua italiana

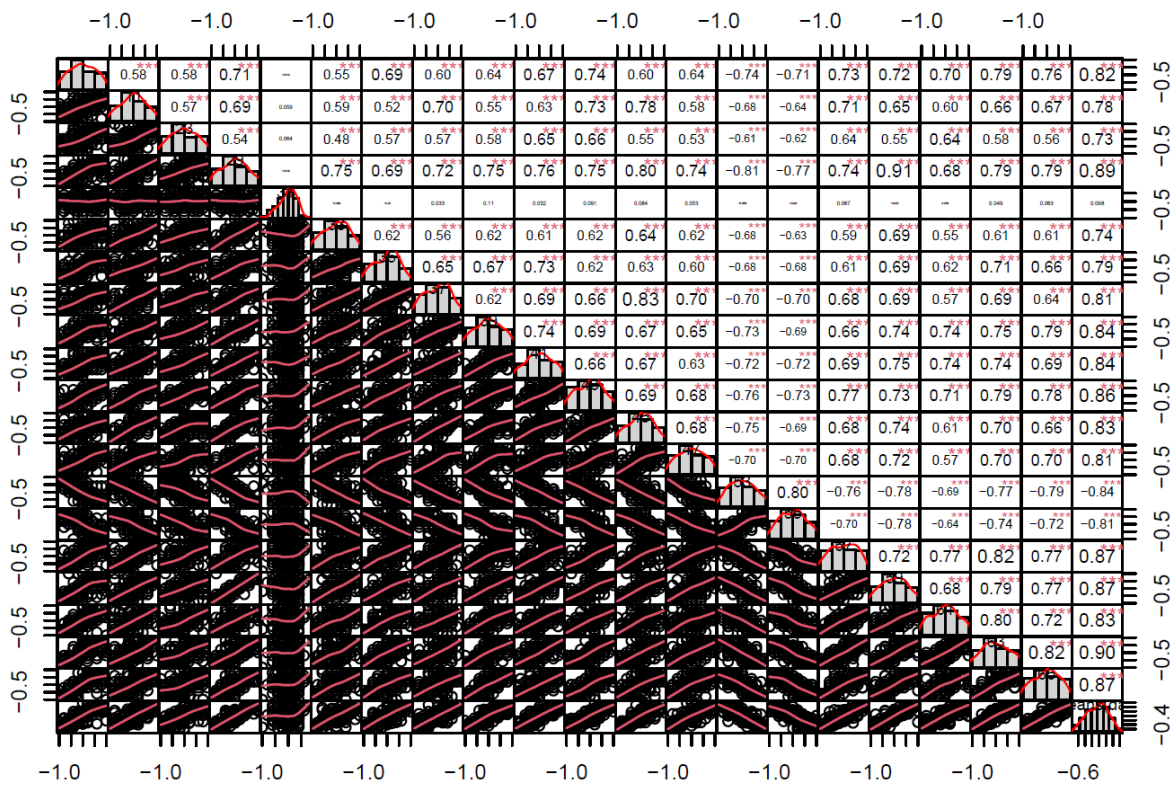


Figura 3.1.5 : Coefficiente di Spearman. Annotatori con sentenze in lingua italiana

Nella figura illustrata sono rappresentati i risultati ottenuti ponendo in relazione 20 annotazioni e la media calcolata tra i loro valori. Osservando l'ultima colonna, in cui sono presenti i coefficienti di correlazione tra la media ed ogni singola annotazione, è possibile individuare anche in questo caso la presenza di valori negativi o molto bassi.

Per una maggiore accuratezza e precisione dei risultati finali, sono stati scartati gli annotatori con un coefficiente minore di 0,8.

Di seguito viene rappresentata un'immagine in cui sono presenti i coefficienti di correlazione ottenuti dall'analisi delle sole annotazioni selezionate.

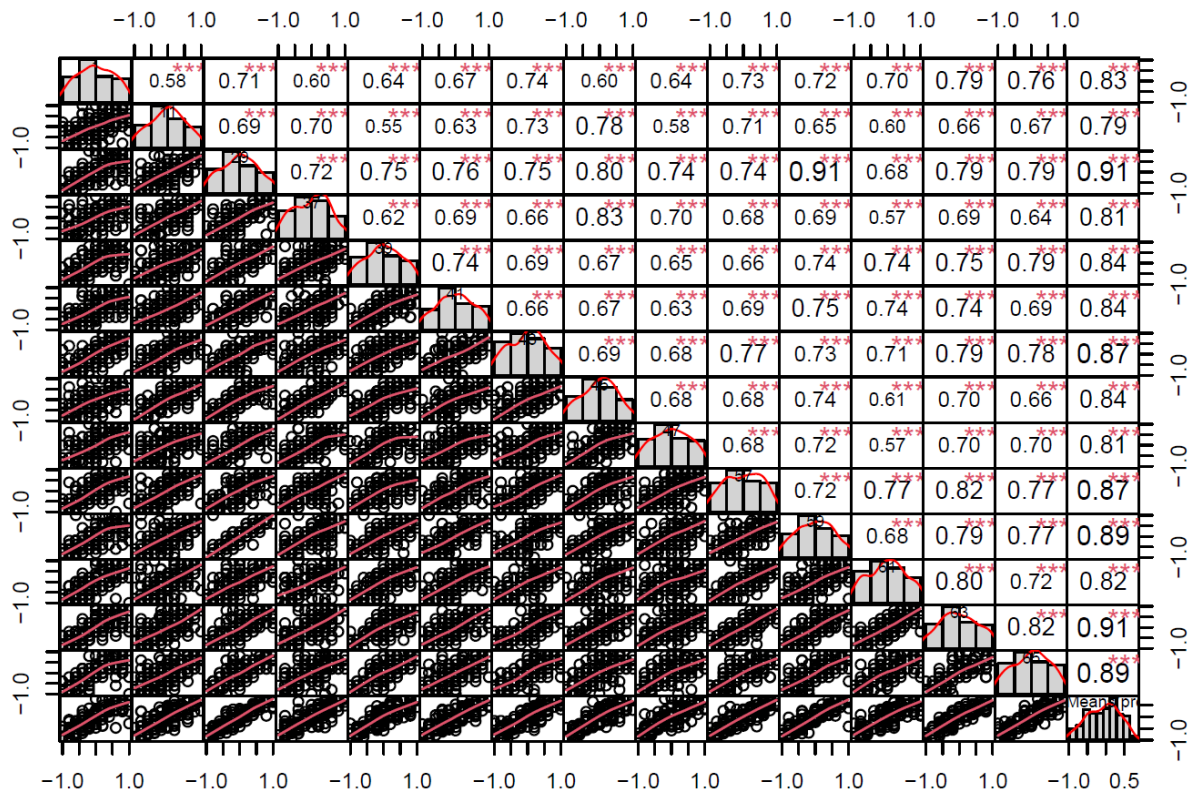


Figura 3.1.6 : Coefficiente di Spearman < 0.8. Annotatori con sentenze in lingua italiana

Si nota come ogni coefficiente, presente nell'ultima colonna, sia vicino al valore 0,8 o maggiore.

3.1.1.6 Annotazioni in lingua inglese

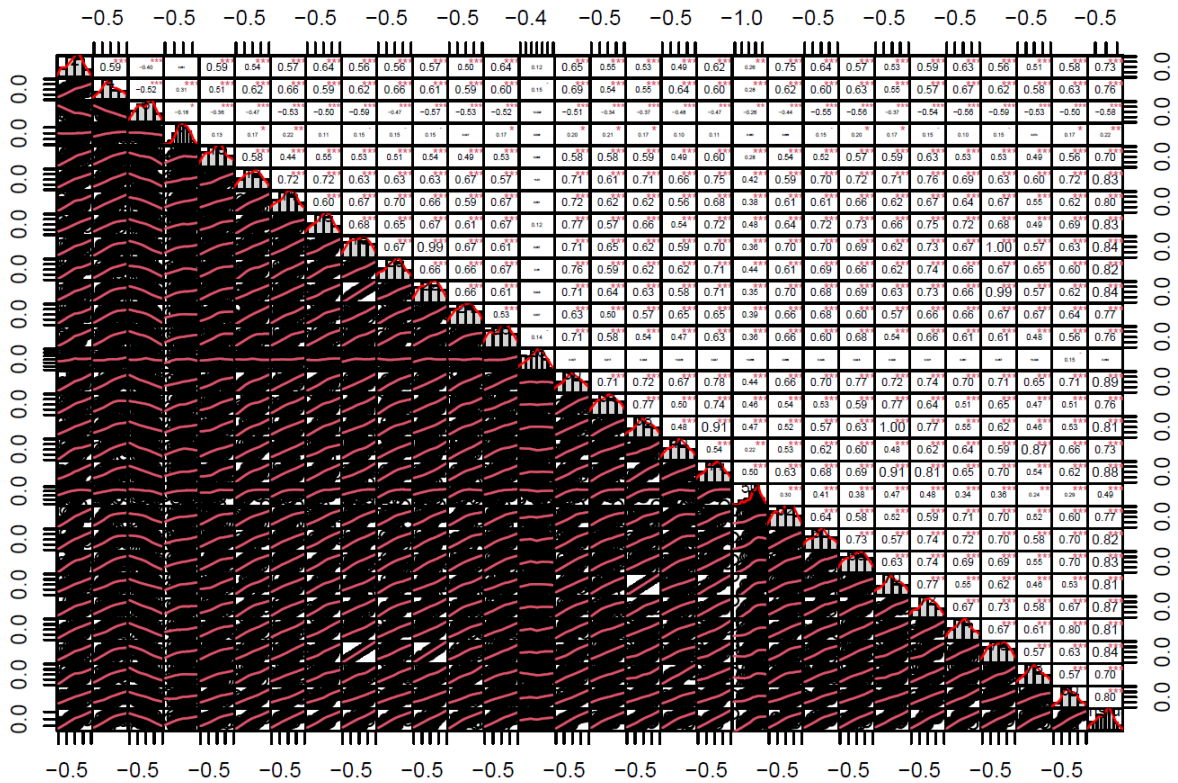


Figura 3.1.7: Coefficiente di Spearman. Annotatori con sentenze in lingua inglese

Nella figura precedente si osserva, come nel caso dei dati delle annotazioni con sentenze in italiano, la presenza di coefficienti con valore negativo o relativamente basso. Anche in questo caso sono stati scartate le annotazioni con tale caratteristica.

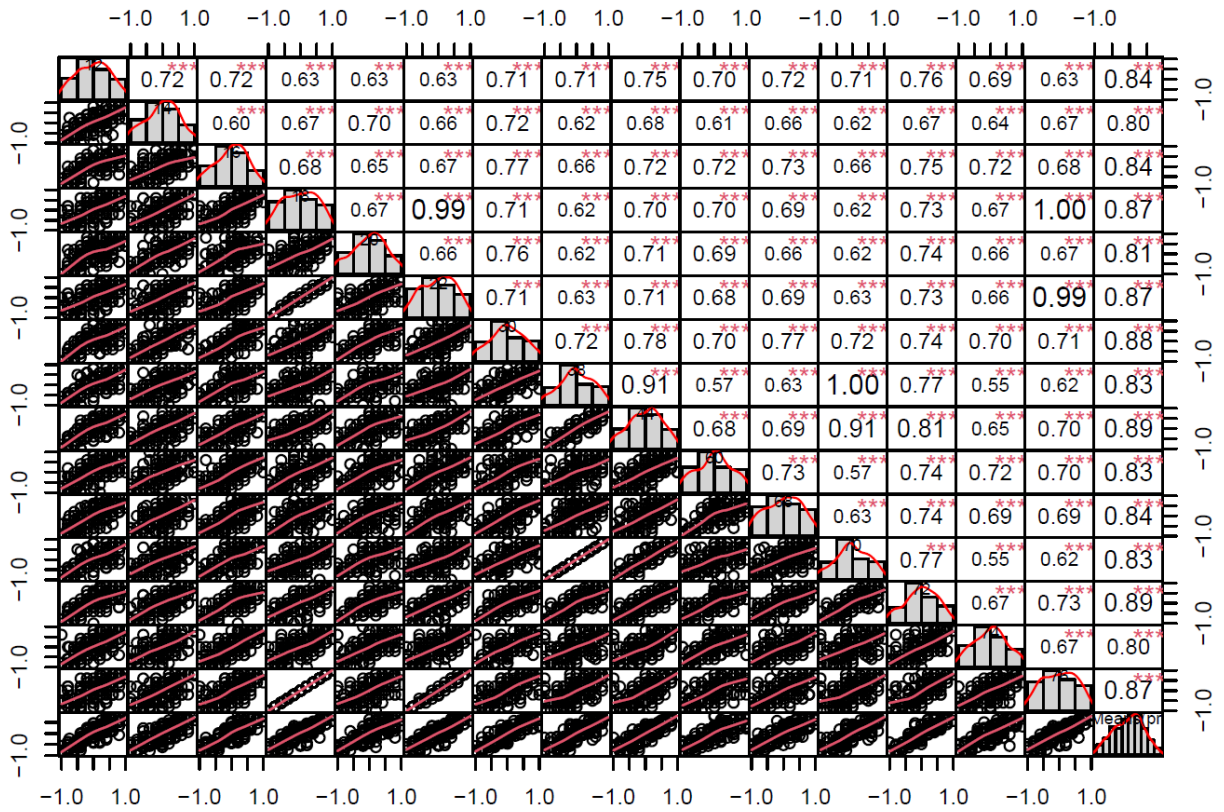


Figura 3.1.8: Coefficiente di Spearman < 0,8 : Annotatori con sentenze in lingua inglese

Una volta eliminate le annotazioni con coefficienti di correlazione minore di 0,8 è stata eseguita nuovamente la verifica di correlazione come mostrato in figura. Anche in questo caso i dati degli annotatori rimanenti mostrano un valore maggiore di 0,8.

3.1.2 Verifica di concordanza tra annotatori selezionati

Selezionati gli annotatori è stato verificato il loro livello di coerenza analizzando il gruppo di annotazioni in lingua italiana e quello in lingua inglese. Di seguito vediamo una tabella con i relativi coefficiente di *Krippendorff*:

Annotazioni	Krippendorff's alpha
Lingua italiana	0,708
Lingua Inglese	0,698

Tabella 3.1.1: Krippendorff. Annotatori in lingua italiana e inglese

Si nota come sia per le annotazioni in lingua italiana che per quelle in lingua inglese otteniamo un buon livello di accordo.

3.1.3 Verifica di correlazione tra annotatori selezionati

Di seguito è illustrato il grafico relativo alla verifica di correlazione ottenuto analizzando la media dei valori del gruppo di annotatori in lingua italiana e in lingua inglese.

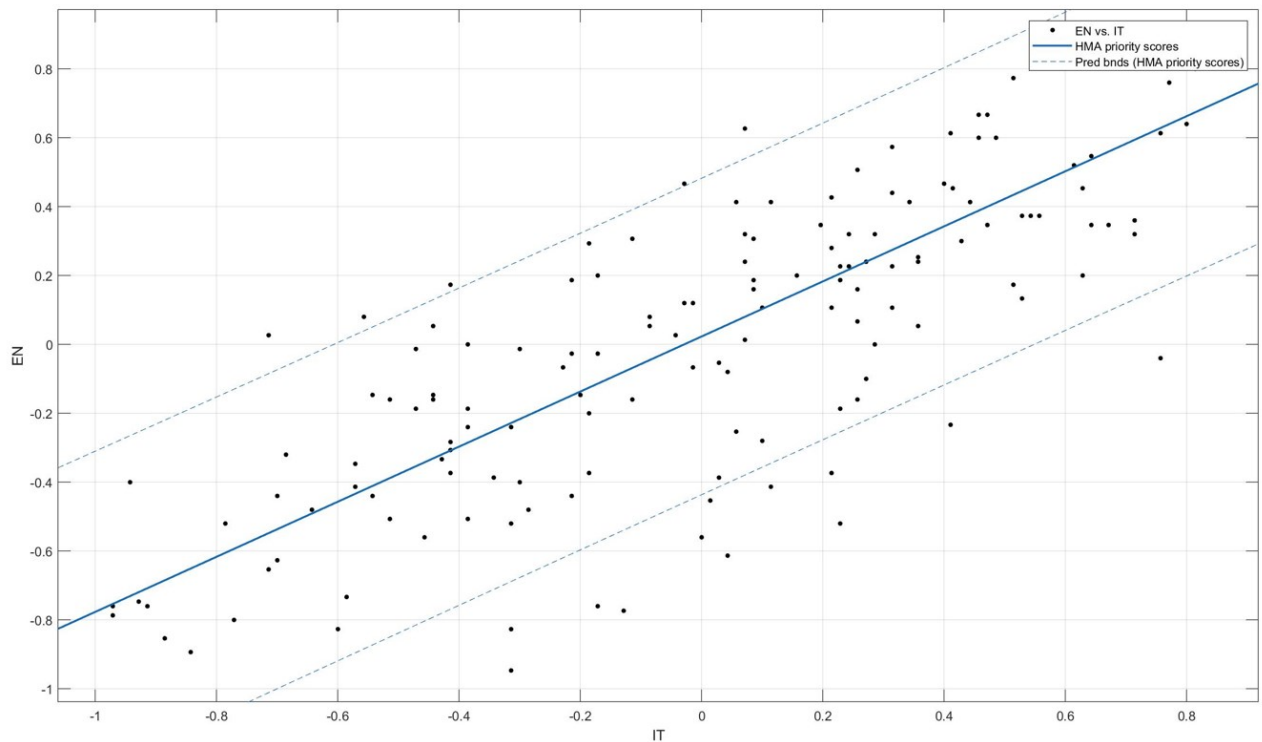


Figura 3.1.9: Verifica correlazione media valori

Lo scopo di questa analisi è stata di verificare la coerenza dei risultati ottenuti dalle annotazioni in lingua italiana e quelli ottenuti dalle annotazioni in lingua inglese.

La nuvola di punti rappresentata in figura dimostra una correlazione tra i due valori di dati analizzati.

Una volta analizzate e selezionate le annotazioni nelle diverse lingue il modello predittivo è stato sviluppato per l'elaborazione del testo in lingua italiana.

3.2 Applicazione classificatori

3.2.1 Presentazione dei risultati

In questo capitolo si illustrano i risultati ottenuti dall'applicazione dei 5 classificatori ML (Naive Bayes Classifier, Decision tree, Logistic Regression, Support Vector Machine e Neural Network) per ogni DTM ottenuta dai tre diversi processi di elaborazione testuale (DTM_sel, DTM_ste e DTM_noste).

In particolare sono presentate le *confusion matrix* di ogni metodologia di analisi del testo in due tipologie di classificazione : binaria e a 4 classi.

Nella rappresentazione dei risultati ogni classe indica il livello di urgenza. In particolare per quanto riguarda la classificazione binaria:

- Classe 1 : *massima urgenza*;
- Classe 2 : *minima urgenza*.

Mentre nella classificazione a quattro classi:

- Classe 1: *massima urgenza*;
- Classe 2: *media urgenza*;
- Classe 3: *limitata urgenza*;
- Classe 4: *minima urgenza*.

L'applicazione del classificatore Naive Bayes è risultata fallimentare per entrambe le tipologie di applicazioni (due classi e quattro classi).

3.2.2 Elaborazione del testo con selezione di parole (DTM_sel)

3.2.2.1 Support Vector Machine

Numero di classi : 4

Tipologia di distribuzione: Linear SVM

		Predetto			
		1	2	3	4
Osservato	1	307	134	4	1
	2	86	467	143	22
	3	9	156	262	88
	4	3	43	134	227

Accuracy	0,61
-----------------	------

Classe	1	2	3	4
Precision	0,76	0,58	0,48	0,67
Recall	0,69	0,65	0,51	0,56
F1	0,72	0,62	0,50	0,61

Figura 3.2.1: SVM, Linear SVM

Numero di classi : 2

Tipologia di distribuzione: Coarse Gaussian

		Predetto	
		1	2
Osservato	1	962	202
	2	165	757

Accuracy	0,82
-----------------	------

Classe	1	2
Precision	0,85	0,79
Recall	0,83	0,82
F1	0,84	0,80

Figura 3.2.2: SVM, Coarse Gaussian

3.2.2.2 Decision Tree

Numero di classi : 4

Tipologia di distribuzione: Optimizable Tree

		Predetto			
		1	2	3	4
Osservato	1	314	127	4	1
	2	107	453	135	23
	3	10	169	248	88
	4	3	49	135	220

Accuracy	0,59
-----------------	------

Classe	1	2	3	4
Precision	0,72	0,57	0,48	0,66
Recall	0,70	0,63	0,48	0,54
F1	0,71	0,60	0,48	0,60

Figura 3.3.3: DT, Optimizable Tree

Numero di classi : 2

Tipologia di distribuzione: Optimizable Tree

		Predetto	
		1	2
Osservato	1	970	194
	2	184	738

Accuracy	0,82
-----------------	------

Classe	1	2
Precision	0,84	0,79
Recall	0,83	0,80
F1	0,84	0,80

Figura 3.2.4: DT, Optimizable Tree

3.2.2.3 Neural Network Classifier

Numero di classi: 4

Tipologia di distribuzione: Trilayered NN

		Predetto			
		1	2	3	4
Osservato	1	316	124	4	2
	2	110	451	127	30
	3	10	162	242	101
	4	2	49	124	232

Accuracy	0,59
-----------------	------

Classe	1	2	3	4
Precision	0,72	0,57	0,49	0,64
Recall	0,71	0,63	0,47	0,57
F1	0,71	0,60	0,48	0,60

Figura 3.2.5: NN, Trilayered NN

Numero di classi: 2

Tipologia di distribuzione: Narrow NN

		Predetto	
		1	2
Osservato	1	977	187
	2	187	735

Accuracy	0,82
-----------------	------

Classe	1	2
Precision	0,84	0,80
Recall	0,84	0,80
F1	0,84	0,80

Figura 3.2.6 : NN, Narrow

3.2.2.4 Logistic Regression

Numero di classi: 4

Tipologia di distribuzione : Kernel LR

		Predetto			
		1	2	3	4
Osservato	1	269	173	4	0
	2	56	538	106	18
	3	6	207	232	70
	4	2	60	135	210

Accuracy	0,60
-----------------	------

Classe	1	2	3	4
Precision	0,81	0,55	0,49	0,70
Recall	0,60	0,75	0,45	0,52
F1	0,69	0,63	0,47	0,60

Figura 3.2.7: LR, Kernal

Numero di classi: 2

Tipologia di distribuzione : Kernel LR

		Predetto	
		1	2
Osservato	1	977	187
	2	194	728

Accuracy	0,82
-----------------	------

Classe	1	2
Precision	0,83	0,80
Recall	0,84	0,79
F1	0,84	0,79

Figura 3.2.8: LR, Kernal

3.2.3 Elaborazione del testo incluso lo stemming (DTM_ste)

3.2.3.1 Support Vector Machine

Numero di classi: 4

Tipologia di distribuzione: Fine Gaussian

		Predetto			
		1	2	3	4
Osservato	1	300	141	4	1
	2	92	479	120	27
	3	9	172	251	83
	4	2	49	134	222

Accuracy	0,60
-----------------	------

Classe	1	2	3	4
Precision	0,74	0,57	0,49	0,67
Recall	0,67	0,67	0,49	0,55
F1	0,71	0,61	0,49	0,60

Figura 3.2.9: SVM, fine gaussian

Numero di classi: 2

Tipologia di distribuzione: Linear SVM

		Predetto	
		1	2
Osservato	1	993	171
	2	195	727

Accuracy	0,82
-----------------	------

Classe	1	2
Precision	0,84	0,81
Recall	0,85	0,79
F1	0,84	0,80

Figura 3.2.10 SVM, linear SVM

3.2.3.2 Decision Tree

Numero di classi: 4

Tipologia di distribuzione: Fine Tree

		Predetto			
		1	2	3	4
Osservato	1	306	134	5	1
	2	92	506	100	20
	3	10	196	221	88
	4	2	57	122	226

Accuracy	0,60
-----------------	------

Classe	1	2	3	4
Precision	0,75	0,57	0,49	0,67
Recall	0,69	0,70	0,43	0,56
F1	0,71	0,63	0,46	0,61

Figura 3.2.11: DT, fine tree

Numero di classi : 2

Tipologia di distribuzione : Fine Tree

		Predetto	
		1	2
Osservato	1	989	180
	2	187	735

Accuracy	0,82
-----------------	------

Classe	1	2
Precision	0,84	0,80
Recall	0,85	0,80
F1	0,84	0,80

Figura 3.2.12:DT,Fine tree

3.2.3.3 Neural Network Classifier

Numero di classi : 4

Tipologia di distribuzione : Bilayered NN

		Predetto			
		1	2	3	4
Osservato	1	309	129	6	2
	2	113	450	126	29
	3	8	159	248	100
	4	2	50	124	231

Accuracy	0,59
-----------------	------

Classe	1	2	3	4
Precision	0,72	0,57	0,49	0,64
Recall	0,69	0,63	0,48	0,57
F1	0,70	0,60	0,49	0,60

Figura 3.2.13 : NN, Bilayered

Numero di classi: 2

Tipologia di distribuzione: Wide NN

		Predetto	
		1	2
Osservato	1	985	179
	2	198	724

Accuracy	0,82
-----------------	------

Classe	1	2
Precision	0,83	0,80
Recall	0,85	0,79
F1	0,84	0,79

Figura 3.2.14: NN, Wide

3.2.3.4 Logistic Regression

Numero di classi: 4

Tipologia di distribuzione: Kernel

		Predetto			
		1	2	3	4
Osservato	1	261	181	4	0
	2	51	540	112	15
	3	5	207	239	64
	4	2	64	132	209

Accuracy	0,60
-----------------	------

Classe	1	2	3	4
Precision	0,82	0,54	0,49	0,73
Recall	0,59	0,75	0,46	0,51
F1	0,68	0,63	0,48	0,60

Figura 3.2.15: LR, karnel

Numero di classe: 2

Tipologia di distribuzione: Kernel

		Predetto	
		1	2
Osservato	1	989	175
	2	203	719

Accuracy	0,82
-----------------	------

Classe	1	2
Precision	0,83	0,80
Recall	0,85	0,78
F1	0,84	0,79

Figura 2.2.16 : LR, karnel

3.2.4 Elaborazione del testo escluso lo stemming (DTM_noste)

3.2.4.1 Support Vector Machine

Numeri di classi: 4

Tipologia di distribuzione: Medium Gaussian

		Predetto			
		1	2	3	4
Osservato	1	300	141	4	1
	2	93	480	120	25
	3	7	172	232	104
	4	2	52	128	225

Accuracy	0,59
-----------------	------

Classe	1	2	3	4
Precision	0,75	0,57	0,48	0,63
Recall	0,67	0,67	0,45	0,55
F1	0,71	0,61	0,46	0,59

Figura 3.2.17: SVM, Medium Gaussian

Numero di classi : 2

Tipologia di distribuzione: Linear SVM

		Predetto	
		1	2
Osservato	1	986	178
	2	208	714

Accuracy	0,81
-----------------	------

Classe	1	2
Precision	0,83	0,80
Recall	0,85	0,77
F1	0,84	0,79

Figura 3.2.18: SVM, Linear

3.2.4.2 Decison Tree

Numero di classi : 4

Tipologia di distribuzione: Optimizable Tree

		Predetto			
		1	2	3	4
Osservato	1	302	139	5	0
	2	85	487	126	20
	3	6	196	229	84
	4	4	59	120	224

Accuracy	0,60
-----------------	------

Classe	1	2	3	4
Precision	0,76	0,55	0,48	0,68
Recall	0,68	0,68	0,44	0,55
F1	0,72	0,61	0,46	0,61

Figura 3.2.19: DT, Optimizable Tree

Numero di classi: 2

Tipologia di distribuzione : Fine Tree

		Predetto	
		1	2
Osservato	1	980	184
	2	204	718

Accuracy	0,81
-----------------	------

Classe	1	2
Precision	0,83	0,80
Recall	0,84	0,78
F1	0,83	0,79

Figura 3.2.20: Fine tree

3.2.4.3 Neural Network Classifier

Numero di classi: 4

Tipologia di distribuzione: Medium NN

		Predetto			
		1	2	3	4
Osservato	1	301	140	4	1
	2	102	455	132	29
	3	9	179	242	85
	4	2	60	120	225

Accuracy	0,59
-----------------	------

Classe	1	2	3	4
Precision	0,73	0,55	0,49	0,66
Recall	0,67	0,63	0,47	0,55
F1	0,70	0,59	0,48	0,60

Figura 3.2.21: NN, Medium

Numero di classi : 2

Tipologia di distribuzione: Narrow

		Predetto	
		1	2
Osservato	1	969	195
	2	191	731

Accuracy	0,81
-----------------	------

Classe	1	2
Precision	0,84	0,79
Recall	0,83	0,79
F1	0,83	0,79

Figura 3.2.22: NN, Narrow

3.2.4.4 Logistic Regression

Numero di classi: 4

Tipologia di distribuzione: Kernel

		Predetto			
		1	2	3	4
Osservato	1	269	173	4	0
	2	56	538	106	18
	3	6	207	232	70
	4	2	60	135	210

Accuracy	0,60
-----------------	------

Classe	1	2	3	4
Precision	0,81	0,55	0,49	0,70
Recall	0,60	0,75	0,45	0,52
F1	0,69	0,63	0,47	0,60

Figura 3.2.23:LR, Kernal

Numero di classi : 2

Tipologia di distribuzione: Kernel

		Predetto	
		1	2
Osservato	1	989	175
	2	206	716

Accuracy	0,82
-----------------	------

Classe	1	2
Precision	0,83	0,80
Recall	0,85	0,78
F1	0,84	0,79

Figura 3.2.24: LR, Kernal

Di seguito vengono mostrate delle tabelle riepilogative in cui sono raccolti i risultati per ogni applicazione dei diversi classificatori.

Classificatore	Distribuzione	Classe	DTM_sel				DTM_ste				DTM_noste			
			Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
Support Vector Machine	Linear SVM	1	0,61	0,76	0,69	0,72								
		2		0,58	0,65	0,62								
		3		0,48	0,51	0,5								
		4		0,67	0,56	0,61								
	Fine Gaussian	1					0,6	0,74	0,67	0,71				
		2						0,57	0,67	0,61				
		3						0,49	0,49	0,49				
		4						0,67	0,55	0,6				
	Medium Gaussian	1									0,59	0,75	0,67	0,71
		2										0,57	0,67	0,61
		3										0,48	0,45	0,46
		4										0,63	0,55	0,59
Decision Trees	Optimizabile	1	0,59	0,72	0,7	0,71					0,6	0,76	0,68	0,72
		2		0,57	0,63	0,6						0,55	0,68	0,61
		3		0,48	0,48	0,48						0,48	0,44	0,46
		4		0,66	0,54	0,6						0,68	0,55	0,61
	Fine Tree	1					0,6	0,75	0,69	0,71				
		2						0,57	0,7	0,63				
		3						0,49	0,43	0,46				
		4						0,67	0,56	0,61				
Neural Network	Trilayered	1	0,59	0,72	0,71	0,71								
		2		0,57	0,63	0,6								
		3		0,49	0,47	0,48								
		4		0,64	0,57	0,6								
	Bilayered	1					0,59	0,72	0,69	0,7				
		2						0,57	0,63	0,6				
		3						0,49	0,48	0,49				
		4						0,64	0,57	0,6				
	Medium	1									0,59	0,73	0,67	0,7
		2										0,55	0,63	0,59
		3										0,49	0,47	0,48
		4										0,66	0,55	0,6
Logistic Regression	Kernel	1	0,6	0,81	0,6	0,69	0,6	0,82	0,59	0,68	0,6	0,81	0,6	0,69
		2		0,55	0,85	0,63		0,54	0,75	0,63		0,55	0,75	0,63
		3		0,49	0,45	0,47		0,49	0,46	0,48		0,49	0,45	0,47
		4		0,7	0,52	0,6		0,73	0,51	0,6		0,7	0,52	0,6

Tabella 3.2.1: Risultati applicazione 4 classi

Distribuzione	Classe	DTM_sel				DTM_ste				DTM_noste			
		Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
Linear SVM	1					0,82	0,84	0,85	0,84	0,81	0,83	0,85	0,84
	2						0,81	0,79	0,8		0,8	0,8	0,77
Coarse Gaussian	1	0,81	0,85	0,83	0,84								
	2		0,79	0,82	0,8								
Optimizable	1	0,82	0,84	0,83	0,84								
	2		0,79	0,8	0,8								
Fine Tree	1					0,82	0,84	0,85	0,84	0,81	0,83	0,84	0,83
	2						0,8	0,8	0,8		0,8	0,78	0,79
Narrow	1	0,82	0,84	0,84	0,84					0,81	0,84	0,83	0,83
	2		0,8	0,8	0,8						0,79	0,79	0,79
Bilayered	1												
	2												
Wide	1					0,82	0,83	0,85	0,84				
	2						0,8	0,79	0,79				
Kernel	1	0,82	0,83	0,84	0,84	0,82	0,83	0,85	0,84	0,82	0,83	0,85	0,84
	2		0,8	0,79	0,79		0,8	0,78	0,79		0,8	0,78	0,79

Tabella 3.2.2 : Risultati applicazione 2 classi

3.3 Discussione dei risultati

Di seguito vengono discussi i risultati raccolti facendo particolare attenzione a due aspetti che hanno condizionato maggiormente i valori ottenuti: valutazione di influenza del filtraggio dei termini e capacità del classificatore di predire classi specifiche.

3.3.1 Valutazione di influenza del filtraggio dei termini

In questo paragrafo si intende descrivere l'influenza del filtraggio dei termini in riferimento alla capacità dei classificatori di comprendere la severità delle singole sentenze analizzate. L'influenza risulta essere particolarmente accentuata in riferimento a due attività condotte durante l'elaborazione del testo. La prima legata al filtraggio dei termini scelti manualmente e la seconda inerente al filtraggio di parole con bassa frequenza. Nella tabella sottostante sono indicati i valori di accuracy, precision, recall e F1 Score inerenti all'applicazione delle azioni introdotte sopra. Questi sono stati estratti dall'applicazione del classificatore Decision Tree.

	Rimozione punteggiatura e stop-words	Inclusione di termini con frequenza =0	Inclusione di termini significativi	Numero termini	Accuracy	Precision		Recall		F1 Score	
						1	2	1	2	1	2
M1	si			197	0,59	0,75	0,54	0,65	0,71	0,70	0,61
M2	si	si		644	0,60	0,74	0,56	0,68	0,67	0,71	0,61
M3	si	si	si	651	0,60	0,76	0,55	0,68	0,68	0,72	0,61

Tabella 3.2.3: Valori filtraggio testo

È possibile notare come all'aumentare dei termini analizzati aumentino i valori. I risultati ottenuti sono coerenti e confrontabili con lo studio di Mo. et al.[34] in cui viene evidenziata l'importanza del numero di termini nel testo.

Di seguito sono descritte le azioni presentate in merito alle conseguenze avute nell'elaborazione del testo e dei risultati ottenuti presentati nella tabella 3.2.3.

Influenza delle stop-words aggiuntive selezionate

La prima attività descritta è legata alla scelta delle stop-words aggiuntive, eliminate nel testo e individuate manualmente in riferimento ai termini che apparentemente risultavano irrilevanti per l'assegnazione di priorità delle sentenze. Questo particolare filtraggio è stato applicato nella metodologia di elaborazione testuale in cui è stato utilizzato il processo di stemming (DTM_ste) e quella esclusa lo stemming (DTM_noste).

In sostanza si è visto come la rimozione di termini, apparentemente non importanti come “*bagno*” o “*aule*”, incidessero negativamente sulla capacità del classificatore di assegnare la giusta priorità. Il motivo è legato al fatto che questi termini sono spesso presenti in sentenze che appartengono ad uno stesso problema lamentato e quindi spesso allo stesso livello di urgenza.

Le conseguenze di questa particolare azione sono visibili nella tabella 3.2.3. Si nota come l'inclusione dei termini significativi generi un aumento dell'andamento medio dei valori. Infatti per la metodologia M3 (Accuracy=0,60; Precision=0,655; Recall=0,68; F1 score=0,665) abbiamo risultati maggiori della M2 (Accuracy=0,60; Precision=0,65; Recall=0,67; F1 score=0,66).

Influenza delle parole con bassa frequenze relativa

Il secondo aspetto legato all'influenza dei termini è la scelta di eliminazione di parole con bassa frequenza.

Il classificatore, durante la sua applicazione, considera il set di dati diviso tra dati utilizzati per la fase di addestramento (training set) e dati utilizzati per la verifica (test set). Se vengono impiegati un numero ridotto di termini si rischia che durante l' applicazione del classificatore, questo vada a pescare termini facenti parte solo di un gruppo dei due set.

Le conseguenze di questa particolare azione, anche in questo caso, sono visibili nella tabella 3.2.3. Infatti l'analisi del testo con parole a bassa frequenza generano un aumento dei valori . In particolare per la metodologia M2 (Accuracy=0,60; Precision=0,65; Recall=0,67: F1 score=0,66) abbiamo risultati maggiori della M1 (Accuracy=0,59; Precision=0,645; Recall=0,68: F1 score=0,655).

In definitiva, l'inclusione di termini apparentemente non importanti e l'aggiunta di termini con bassa frequenza, favoriscono un raggiungimento di buoni risultati in quanto il testo analizzato risulta essere più ricco di termini.

3.3.2 Valutazione della capacità della predizione di classi specifiche

In questo paragrafo viene descritta l'efficacia dei classificatori di risolvere il problema della prioritizzazione di interventi manutentivi.

Prendendo in considerazione i risultati ottenuti ed osservando le tabelle riepilogative in cui sono rappresentati i coefficiente riferiti ad ogni singola classe, si nota come quelle negative presentano dei coefficienti più elevati rispetto alle altre.

Per quanto riguarda i risultati che illustrano i coefficienti ottenuti con l'applicazione delle quattro classi, si nota come quelle indicate con *massima urgenza* ed *urgenza media (1 e 2)* hanno coefficienti più elevati. Mentre per l'applicazione con due classi è possibile notare come per quella di *massima urgenza*, anche in questo caso, risulta avere coefficienti più elevati rispetto all'altra. In particolare prendendo in riferimento i risultati ottenuti con l'impiego del classificatore SVM nel caso della DTM_sel vediamo come la media dei valori delle due classi più urgenti (Precision=0,67; Recall=0,67: F1 score=0,67) risultino essere maggiori di quelle meno urgenti (Precision=0,575; Recall=0,535: F1 score=0,555). Questo vale anche per la classificazione a due vie, considerando il caso del classificatore SVM, notiamo dei risultati maggiori per la classe 1 (Precision=0,85; Recall=0,83: F1 score=0,84) piuttosto che per la classe 2 (Precision=0,79; Recall=0,82: F1 score=0,80). Questo particolare fenomeno è individuabile nell'applicazione di tutti i classificatori. Infatti, nell'applicazione a quattro vie si nota come i coefficienti con un valore maggiore o uguale di 0,70 (celle colorate tabella 3.2.1) siano presenti nella maggior parte dei casi nelle prime

due classi. Mentre per la classificazione a due vie i valori maggiori o uguali di 0,84 (celle colorate tabella 3.2.2) sono presenti nella prima classe.

Il modello sviluppato risulta essere ottimo per la predizione di richieste manutentive di maggiore urgenza.

4 CONCLUSIONI

L'obiettivo della tesi ha previsto lo sviluppo di un sistema automatizzato, basato su tecniche di Machine Learning, applicate al linguaggio naturale per l'assegnazione delle priorità di intervento nella manutenzione di patrimoni immobiliari, facendo particolare attenzione a quali fossero le operazioni di pre- trattamento del testo necessarie a migliorare l'affidabilità del processo.

Per condurre lo studio sono state analizzate delle richieste di manutenzione generate nell'arco di 34 mesi nei 23 edifici dell'Università Politecnica delle Marche (12.000 interventi ca.). Per l'ottenimento di score di riferimento utili a sviluppare il metodo è stato chiesto a 40 annotatori, appartenenti a 3 classi di esperienza differente (Limitata, Media ed Elevata), di attribuire manualmente il livello di urgenza su 149 richieste manutentive secondo il metodo *Best-Worst Scale*. Per analizzare i testi da cui sono composte le richieste manutentive sono state scelte tre combinazioni di metodologie di pre-elaborazione testuale. In particolare nella prima il testo delle richieste è stato studiato prendendo in riferimento delle parole scelte da esperti da loro considerate significative. Nella seconda il testo è stato trattato da processi come: eliminazione di numeri e punteggiatura, sostituzione dei sinonimi con una parola di riferimento, rimozione di stop-words, stemming e pos-tagging. Mentre nell'ultima combinazione di pre-elaborazione , il testo è stato trattato con processi uguali alla seconda ma escludendo lo stemming.

I testi ottenuti dalle combinazione di metodologia di trattamento di pre-elaborazione testale sono stati il punto di partenza per la creazione delle Document Term Matrix. In particolare sono state create 3 diverse DTM per ogni combinazione descritta sopra (DTM_sel, DTM_ste, DTM_noste).

In seguito, ogni matrice è stata utilizzata per l'applicazione dei 5 classificatori di Machine Learning selezionati (Naive Bayes Classifier, Support Vector Machine, Logistic Regression, Neural Network e Decision Trees) e sono stati ricavati i coefficienti necessari per la valutazione delle diverse strategie utilizzate per ogni differente distribuzione del classificatore scelta. Ogni classificazione è stata effettuata con quattro e due classi dove il livello di urgenza massimo, nel primo caso è stato indicato con le prime due classi mentre nella classificazione binaria con la prima.

Analizzando i risultati, si nota come il classificatore Naive Bayes sia risultato poco efficiente per ogni tipologia di DTM applicata. In riferimento agli altri classificatori sono stati raggiunti livelli di accuratezza del circa 60% per l'applicazione con 4 classi e del circa 80% per le due classi. Il classificatore Support Vector Machine è risultato il migliore vista un accuratezza del 61% (4 classi) e del 82% (2 classi).

Inoltre, grazie ai risultati ottenuti tra la comparazione dei diversi testi trattati e analizzati, in prima istanza si è visto come processi di pre-elaborazione testuale, finalizzati alla riduzione dei termini rappresentativi di un testo, portino ad una diminuzione dell'efficienza dei classificatore. Il motivo è legato al fatto che più il numero di parole è elevato e più vi è la probabilità che un determinato termine sia presente nei set di dati utili all'apprendimento (training set e test set). In secondo luogo, grazie alla comparazione dei coefficienti (precision, recall ed F1) i modelli di classificazione ottenuti risultano essere particolarmente efficienti per la predizioni di classi più urgenti.

È possibile concludere che il modello sviluppato con tecniche di ML, estrapolando informazioni da testi scritti, risulta essere un metodo efficace

per facilitare l'assegnazione di priorità di intervento, riducendo il carico di lavoro del gruppo di Facility Manager che si occupa di gestire la struttura.

Inevitabilmente lo studio presenta delle limitazioni, tra cui la necessità di ampliare, in termini di dimensioni, il dataset quale è stato fatto il processo di HMA per migliorare l'accuratezza nel caso di classificazione a quattro vie. Studi futuri potrebbero essere mirati all'implementazione del modello con altri metodi di analisi testuali come la costruzione di un lessico specifico che porterebbe ad un'analisi più dettagliata delle richieste manutentive. Inoltre, il modello sviluppato potrebbe essere abilitato all'utilizzo con la lingua inglese così da rendere più accessibile tale sistema.

5 BIBLIOGRAFIA E SITOGRAFIA

- [1] N. Nabilah, I. Syahrizal, N. Izieadiana, F. Mat, N. Zainab, and N. Azizan, “Review on maintenance issues toward building maintenance management best practices,” *J. Build. Eng.*, vol. 44, no. October 2020, p. 102985, 2021, doi: 10.1016/j.jobe.2021.102985.
- [2] C. Chen and L. Tang, “BIM-based integrated management workflow design for schedule and cost planning of building fabric maintenance,” *Autom. Constr.*, vol. 107, no. July, p. 102944, 2019, doi: 10.1016/j.autcon.2019.102944.
- [3] “Life Cycle Costing in Construction: Reduce your building’s lifetime costs.” [Online]. Available: <https://medium.com/@oneclicklca/life-cycle-costing-in-construction-reduce-your-buildings-lifetime-costs-3c83775ff43e>.
- [4] M. A. El-Haram and R. M. W. Horner, “Application of the principles of ILS to the development of cost effective maintenance strategies for existing building stock,” *Constr. Manag. Econ.*, vol. 21, no. 3, pp. 283–296, Apr. 2003, doi: 10.1080/0144619032000093774.
- [5] B. Chanter and P. Swallow, *Building Maintenance Management: Second Edition*. 2008.
- [6] I. Flores-Colen and J. De Brito, “A systematic approach for maintenance budgeting of buildings faades based on predictive and preventive strategies,” *Constr. Build. Mater.*, vol. 24, no. 9, pp. 1718–1729, 2010, doi: 10.1016/j.conbuildmat.2010.02.017.

- [7] W. Chen, K. Chen, J. C. P. Cheng, Q. Wang, and V. J. L. Gan, “BIM-based framework for automatic scheduling of facility maintenance work orders,” *Autom. Constr.*, vol. 91, no. March, pp. 15–30, 2018, doi: 10.1016/j.autcon.2018.03.007.
- [8] B. Schmidt and L. Wang, “Cloud-enhanced predictive maintenance,” *Int. J. Adv. Manuf. Technol.*, vol. 99, no. 1–4, pp. 5–13, 2018, doi: 10.1007/s00170-016-8983-8.
- [9] C. Talamo and M. Bonanomi, *Knowledge management and information tools for building maintenance and facility management*. 2015.
- [10] M. Kuhn and K. Johnson, *Applied predictive modeling*. 2013.
- [11] Y. Bouabdallaoui, Z. Lafhaj, P. Yim, L. Ducoulombier, and B. Bennadji, “Predictive maintenance in building facilities: A machine learning-based approach,” *Sensors (Switzerland)*, vol. 21, no. 4, pp. 1–15, 2021, doi: 10.3390/s21041044.
- [12] R. Bortolini and N. Forcada, “Analysis of building maintenance requests using a text mining approach: building services evaluation,” *Build. Res. Inf.*, vol. 48, no. 2, pp. 207–217, 2020, doi: 10.1080/09613218.2019.1609291.
- [13] P. Parsanezhad and J. Dimyadi, “Effective Facility Management and Operations via a BIM-based Integrated Information System EFFECTIVE FACILITY MANAGEMENT AND OPERATIONS VIA A BIM-BASED INTEGRATED INFORMATION SYSTEM,” no. May, 2014, doi: 10.13140/2.1.2298.9764.

- [14] “Maintenance Work Order Based on Event-Driven Information of Condition-Monitoring Systems—A Genetic Algorithm for Scheduling and Rescheduling of Dynamic Maintenance Orders.” [Online]. Available: <https://www.wbdg.org/facilities-operations-maintenance>.
- [15] M. A. Beauregard and S. K. Ayer, “Leveraging previously reported research to create a decision support tool for institutional facility maintenance,” *J. Facil. Manag.*, vol. 17, no. 3, pp. 249–266, Jan. 2019, doi: 10.1108/JFM-08-2018-0051.
- [16] S. Madureira, I. Flores-Colen, J. de Brito, and C. Pereira, “Maintenance planning of facades in current buildings,” *Constr. Build. Mater.*, vol. 147, pp. 790–802, 2017, doi: 10.1016/j.conbuildmat.2017.04.195.
- [17] A. Grigoriev, “Modeling and solving the periodic maintenance problem,” vol. 172, pp. 783–797, 2006, doi: 10.1016/j.ejor.2004.11.013.
- [18] J. Pongpech and D. N. P. Murthy, “Optimal periodic preventive maintenance policy for leased equipment,” vol. 91, pp. 772–777, 2006, doi: 10.1016/j.res.2005.07.005.
- [19] X. Xu, S. Member, M. Kezunovic, and D. Wong, “Agent-Oriented Approach to Work Order Management for Circuit Breaker Maintenance,” pp. 1–6.
- [20] N. Hamdi, R. Oweis, H. A. Zraiq, and D. A. Sammour, “An Intelligent Healthcare Management System: A New Approach in Work-order Prioritization for Medical Equipment Maintenance Requests,” pp. 557–567, 2012, doi: 10.1007/s10916-010-9501-4.

- [21] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Found. Trends® Inf. Retr.*, vol. 2, no. 1–2, pp. 1–135, 2008, doi: 10.1561/1500000011.
- [22] J. Žižka, F. Dařena, and A. Svoboda, *Text Mining with Machine Learning*. 2019.
- [23] J. J. McArthur, N. Shahbazi, R. Fok, C. Raghubar, B. Bortoluzzi, and A. An, "Machine learning and BIM visualization for maintenance issue classification and enhanced data collection," *Adv. Eng. Informatics*, vol. 38, no. June, pp. 101–112, 2018, doi: 10.1016/j.aei.2018.06.007.
- [24] S. Baek, W. Jung, and S. H. Han, "A critical review of text-based research in construction: Data source, analysis method, and implications," *Autom. Constr.*, vol. 132, no. September, p. 103915, 2021, doi: 10.1016/j.autcon.2021.103915.
- [25] "Text Classification Using Naive Bayes: Theory & A Working Example," [Online]. Available: <https://towardsdatascience.com/text-classification-using-naive-bayes-theory-a-working-example-2ef4b7eb7d5a>.
- [26] F. Lu and Q. Bai, "A refined weighted K-nearest neighbors algorithm for text categorization," *Proc. 2010 IEEE Int. Conf. Intell. Syst. Knowl. Eng. ISKE 2010*, pp. 326–330, 2010, doi: 10.1109/ISKE.2010.5680854.
- [27] C. J. (1984). Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, "Classification And Regression Trees," [Online]. Available: <https://doi.org/10.1201/9781315139470>.

- [28] Z. Jin, J. Shang, Q. Zhu, C. Ling, W. Xie, and B. Qiang, “RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12343 LNCS, pp. 503–515, 2020, doi: 10.1007/978-3-030-62008-0_35.
- [29] W. Zhang, T. Yoshida, and X. Tang, “Text classification based on multi-word with support vector machine,” *Knowledge-Based Syst.*, vol. 21, no. 8, pp. 879–886, 2008, doi: 10.1016/j.knosys.2008.03.044.
- [30] A. Genkin, D. D. Lewis, and D. Madigan, “Sparse Logistic Regression for Text Categorization,” *Analysis*, p. 354, 2005, [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.89.3158&rep=rep1&type=pdf>.
- [31] IBM, “Reti Neurali.” 2020, [Online]. Available: <https://www.ibm.com/it-it/cloud/learn/neural-networks>.
- [32] “Choose Classifier Options.,” [Online]. Available: <https://it.mathworks.com/help/stats/choose-a-classifier.html#bunt0m5>.
- [33] V. García-Díaz, J. P. Espada, R. G. Crespo, B. C. Pelayo G-Bustelo, and J. M. Cueva Lovelle, “An approach to improve the accuracy of probabilistic classifiers for decision support systems in sentiment analysis,” *Appl. Soft Comput. J.*, vol. 67, pp. 822–833, 2018, doi: 10.1016/j.asoc.2017.05.038.
- [34] Y. Mo, D. Zhao, J. Du, M. Syal, A. Aziz, and H. Li, “Automated staff assignment for building maintenance using natural language processing,” *Autom. Constr.*, vol. 113, no. November 2019, p. 103150,

2020, doi: 10.1016/j.autcon.2020.103150.

- [35] S. Assaf and I. Srour, “Using a data driven neural network approach to forecast building occupant complaints,” *Build. Environ.*, vol. 200, no. March, p. 107972, 2021, doi: 10.1016/j.buildenv.2021.107972.
- [36] J. J. Louviere, T. N. Flynn, and A. A. J. Marley, *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press, 2015.
- [37] S. Kiritchenko and S. M. Mohammad, “Best–Worst scaling more reliable than rating scales: A case study on sentiment intensity annotation,” *ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 2, no. August, pp. 465–470, 2017, doi: 10.18653/v1/P17-2074.
- [38] K. Gwet, “On Krippendorff’s Alpha Coefficient,” Oct. 2015.
- [39] P. Sedgwick, “Spearman’s rank correlation coefficient,” *BMJ*, vol. 349, p. g7327, Nov. 2014, doi: 10.1136/bmj.g7327.
- [40] A. Minini, “Algoritmo di stemming.” [Online]. Available: <https://www.andreaminini.com/ir/stemming/>.
- [41] K. K. Al-jabery, T. Obafemi-Ajayi, G. R. Olbricht, and D. C. Wunsch II, “9 - Data analysis and machine learning tools in MATLAB and Python,” K. K. Al-jabery, T. Obafemi-Ajayi, G. R. Olbricht, and D. C. Wunsch II, Eds. Academic Press, 2020, pp. 231–290.