PROGETTI V CALL 2024/2025

Titolo progetto:	Applicazione di metodi di Intelligenza Artificiale per la codifica automatica delle cause di morte
Descrizione	Con il presente progetto si intende sperimentare l'applicazione di metodi di Intelligenza Artificiale per la classificazione delle cause di morte. In particolare, si vuole migliorare la performance degli attuali metodi di codifica automatica utilizzati dall'Istat. Annualmente, l'Indagine sui decessi e le cause di morte censisce circa 670mila decessi ai quali deve essere attribuita la causa iniziale di morte secondo le disposizioni della Classificazione Internazionale delle Malattie (ICD-10) dell'OMS (Organizzazione Mondiale della Sanità). Attualmente il sistema utilizzato è Iris che raggiunge una performance pari a circa l'80% (decessi codificati automaticamente rispetto al totale). È necessario, quindi, codificare manualmente circa 135mila decessi all'anno, con un notevole impiego di risorse altamente formate. Iris è un software utilizzato in molti Paesi d'Europa e del mondo per la codifica automatica e manuale delle cause di morte. Queste cause sono certificate dai medici curanti dei deceduti o da medici necroscopi su una scheda cartacea con diverse righe in cui vanno descritte la sequenza di malattie o traumatismi che ha condotto direttamente a morte, gli altri stati morbosi rilevanti esclusi dalla sequenza principale e le circostanze che hanno provocato i traumatismi o avvelenamenti nel caso in cui essi siano stati i responsabili del decesso. Un primo modulo di Iris effettua il riconoscimento del testo riportato sulla scheda, attribuendo a ciascuna espressione diagnostica un codice ICD-10. Tale riconoscimento avviene grazie a un dizionario di termini medici sviluppato da esperti dell'Istat. Su ciascuna riga della scheda possono essere riportate più espressioni diagnostiche separate da segni di interpunzione o da espressioni di connessione come "dovuto a", "con", ecc. Talvolta oltre al codice base della Classificazione Iris aggiunge ulteriori descrittori al codice (flags, durate delle malattie, ecc.). Un secondo modulo di Iris si basa sull'insieme dei codici ICD-10 assegnati nella fase di riconoscimento del testo e, tenendo conto anche di informazioni di contesto, quali sesso e età del deceduto, durate e flag, e altre malattie riportate, elabora i codici fornendo una stringa coerente di codici ICD-10, utilizzata nelle statistiche di mortalità, chiamata "stringa delle cause multiple di decesso". Un ultimo modulo si basa sulla stringa di cause multiple e, applicando un dettagliato algoritmo descritto nell'ICD-10, seleziona una sola causa di morte, detta "causa iniziale", che viene utilizzata per le statistiche ufficiali e costituisce lo standard per i confronti internazionali di mortalità per causa. La maggior parte degli scarti di Iris è dovuta al mancato riconoscimento del testo medico (primo modulo di Iris) dovuto a incompletezza del dizionario, errato reporting da parte dei medici certificatori o inadeguata registrazione delle schede. Gli altri due moduli si basano su regole deterministiche e la loro performance è molto elevata.
Obiettivi	L'obiettivo specifico del progetto è migliorare la performance del modulo di riconoscimento del testo medico di Iris attraverso l'utilizzo di metodi di intelligenza artificiale per l'attribuzione di codici ICD-10 (con i relativi descrittori aggiuntivi) a ciascuna espressione diagnostica riportata sulle schede di morte. Si vorrebbe quindi arrivare a codificare, utilizzando metodi di intelligenza artificiale, almeno una parte delle espressioni che attualmente sono codificate manualmente.
Metodologia	L'obiettivo di classificazione mediante ICD-10 del testo medico, descrittivo delle cause di morte, è stato perseguito sperimentando diversi approcci metodologici, che vanno dal Machine Learning tradizionale al Deep Learning, per arrivare ai Large Language Models (LLM). Per la costruzione dei modelli sono stati forniti i seguenti dataset, contenenti testi medici etichettati con una classe presente in ICD-10. Nello specifico: • Dataset di Training: testi riportati sulle schede di morte codificate (in modalità completamente automatica da Iris o manualmente da un codificatore esperto) degli anni 2022 e 2023, esclusi i testi contenuti nel file di Test (n=4327927) • Dataset di Test: testi riportati sulle schede di morte scartate da Iris relative ad agosto e settembre 2023, esclusi i testi relativi a traumatismo e causa esterna (n=11909) • Gold standard: selezione di circa 1000 testi dal file di Test dei quali è stata verificata la correttezza della codifica

	Si è scelto di includere nel dataset di Test, e di conseguenza nel Gold standard, solo testi scartati da Iris, in quanto obiettivo del progetto è applicare metodi di AI per la codifica di questo sottoinsieme. Mediante questi test set si valuta la capacità dei modelli di AI di generalizzare su dati mai visti durante la fase di training. Nel dataset di Training sono stati inclusi anche testi codificati automaticamente da Iris, che sono la maggior parte, in quanto si è valutato che possono essere utili per l'addestramento del modello. Come enunciato nella parte descrittiva del progetto, i testi presenti nei dataset sono più o meno complessi: nel migliore dei casi ad una singola espressione diagnostica è associato univocamente un codice (etichetta), oppure ad n espressioni, presenti nella stessa riga e separate da un punto e virgola, corrispondono n codici (etichette) sempre separati da un punto e virgola; nel peggiore dei casi, ad una espressione corrispondono due o più codici, separati da punto e virgola o da "slash" Questo dipende dalla presenza, nel testo, di congiunzioni, preposizioni o locuzioni che vengono trattate come separatori e che, pertanto suddividono lo stesso testo in più entità diagnostiche, ovvero, più cause di morte. Per le ragioni di complessità di cui sopra, prima di tutto è stato effettuato un oneroso e sofisticato trattamento di pre-processing (comune a tutti gli algoritmi di IA utilizzati) di ciascun dataset, allo scopo di fornire agli algoritmi di classificazione un input il più possibile privo di ambiguità. Inoltre, il dataset di Training è stato arricchito dai testi etichettati estratti dall'indice analitico della classificazione ICD-10 che consiste in una lista alfabetica di malattie con il corrispondente codice ICD-10. Terminato il pre-processing, una serie di modelli di IA sono stati addestrati e ne sono state valutate le performance, mediante metriche standard (e.g. Accuracy, F1-score), rispetto a tutti i dataset di training e di test. Più in dettaglio, il modello di Machine Learning scelto è stato il **Random Forest** (***RF***), che da letteratura risulta piuttosto performante su task di classificazione, soprattutto su testi di breve lunghezza come quelli medici. Per quanto riguarda l'approccio Deep Learning, è stato sperimentato prima l'approccio non Foundational, ossia basato su modelli non pre-addestrati che non facciano uso di Transfer Learning. Pertanto è stato usato un modello basato su uno spazio Embedding Word2Vec a 300 dimensioni e una rete ricorrente di ultima generazione Bidirectional LSTM (Long Short Termi Memory), che è un modello bidirezionale come Bert in grado di comprendere sia il contesto di sinistra che di destra di ciascuna parola. In seguito si è passati all'uso di un Transfomer come **Bert** (*Bidirectional Encoder Representations from Transfomers*) che sfruttano il Transfer Learning, ovvero scegliendo diversi tipi di modelli pre-addestrati su testi in italiano generico (***XML-Roberta – Base e Large***) o su testi medici specialistici (***Bert Bio ITA***, ***MedBIT-r3-Plus)*** che presentanto un numero di parametri che varia da 270 a 550 milioni. Infine, è stato sperimentato l'uso di un LLM come **LLama base 1B** (i.e. circa 1 miliardo di parametri). Sia Bert che LLama sono stati successivamente sottoposti a *fine-tuning* sul dataset di Training, ovvero riaddestrati in modo da aggiornare i loro parametri per specializzarli alla classificazione nell'ambito del dominio di interesse. La sperimentazione è stata condotta addestrando tutti i modelli di IA selezionati con un campione di 300K testi, anziché con l'intero dataset di training, a causa dei limiti computazionali delle risorse hardware a disposizione che, soprattutto per modelli con tantissimi parametri come Llama, avrebbero richiesto maggiori capacità di RAM e processori più veloci. La piattaforma utilizzata per addestrare questi modelli è Azure Machine Learning con macchine virtuali equipaggiate di GPU Nvidia A100 con 80 Giga di Ram e 5000 unità di elaborazione parallele (Nvidia Cuda Cores). Per valutare le performance dei modelli è stato calcolato l'F1-score, che è la media armonica di altre due metriche molto conosciute: Precision e Recall. F1-Score è una metrica che misura l'accuratezza delle classificazioni ed è particolarmente adatta in presenza di dataset, come quello in esame, ove le classi sono fortemente "sbilanciate", ovvero il numero di esempi che rappresenta ogni classe è molto eterogeneo.
Risultati ottenuti	Dai risultati sperimentali emerge quanto segue: • rispetto al training set, i modelli che raggiungono l'accuratezza più alta sono i modelli Bert, che mostrano valori di F1-score intorno al 99% sul Training Set, con *XML-Roberta – Large* in testa con il **99,5%**. Invece l'accuratezza sul Test Set interno (20% estratto dal Training set) è del **97.7%**. In particolare il miglior Bert è quello dove abbiamo effettuato il retraining del suo spazio embedding su tutto il corpus testuale del training set, test set e gold standard. Il tempo di Training di questo XML-Roberta Large è stato di 4 ore su GPU. Il modello di Deep Learning non pre-trained basato su W2V+LSTM con retraining degli embeddings su tutto il corpus testuale (Training set+Test set+Gold Standard) ha totalizzato risultati

	competitivi anche con **97%** sul Training Set e **94.8%** su Test Set. • Il modello peggiore è il RF con **97,4%** e **94.2%** su test set, seguito da LLama con **98,5%** e **95.3%** sul test set. Da notare che il valore ottenuto con RF è comunque alto, trattandosi di un algoritmo addestrato esclusivamente sui testi presenti nel dataset fornito dai colleghi tematici e non pre-addestrato dunque su altri testi; • rispetto al dataset denominato "Gold Standard", i cui 1000 testi contengono espressioni diagnostiche scartate da Iris (pertanto terminologia medica spesso diversa da quella presente nel dataset di addestramento) i valori di F1-score ottenuti da tutti i modelli sono più bassi. Il migliore in questo caso risulta essere XML-Roberta Large con retraining degli embeddings con **53.5%** di F1-Score seguito da Llama, con il suo **51,3%**, seguito da *XML-Roberta – Base* con **50,2%**. Con RF e gli altri Bert si ottengono valori F1-score al di sotto del **50%**. Infine W2V+BiLSTM ha raggiunto un risultato deludente sul Gold Standard con **37%** di F1-Score. Da sottolineare che tali valori sono comunque da considerarsi soddisfacenti, vista la diversità dei contenuti di tale dataset rispetto a quella prevalente e nota dei testi nel training. In definitiva, il modello più promettente fino ad oggi è risultato *XML-Roberta – Large.* Più in generale, l'attività svolta nel presente laboratorio ha dimostrato che gli algoritmi di IA possono offrire un valido supporto all'indagine sulle Cause di morte, pertanto si intende portare avanti la sperimentazione con l'obiettivo di migliorare ancora di più l'accuratezza e di riuscire ad aumentare il numero di testi classificati, soprattutto di quelli contenenti termini nuovi.
Membri del Team	Referente progetto: Simone Navarra (DCSW) Referente informatico: Angela Pappagallo (DCME) Esperti tematici (DCSW): Tania Bracci, Francesco Grippo, Chiara Orsi Esperto informatico (DCME): Francesco Pugliese