Università degli Studi di Bari Aldo Moro

Dipartimento di Informatica

Feature Engineering con Ontologie Cliniche

Progetto di Ingegneria della Conoscenza (ICON)

Componenti del gruppo

Giovanni Discanno Matricola 777311

g.discanno2@studenti.uniba.it

Repository GitHub:

https://github.com/Taip03/ICON-2425

Anno Accademico 2024/2025

Indice

Capitolo 1

Introduzione

In questo progetto l'obiettivo è mostrare come una **Knowledge Base** (la Disease Ontology e altre ontologie cliniche) possa essere utilizzata per arricchire un dataset medico con nuove feature derivate dal ragionamento ontologico, migliorando le prestazioni dei modelli di apprendimento supervisionato.

L'obiettivo del progetto è confrontare le performance ottenute su:

- dataset originale (*raw*);
- dataset arricchito con categorie ontologiche e regole cliniche (*moderato*);
- dataset arricchito con ulteriori regole cliniche (*avanzato*).

Requisiti funzionali

Per lo sviluppo del progetto è stato utilizzato l'IDE Visual Studio Code. Il linguaggio utilizzato nel progetto è stato Python nella sua versione 3.11, in quanto offre una ricca scelta di librerie adatte all'uso di modelli di apprendimento supervisionato e al lavoro con ontologie accessibili via web. In particolare le librerie utilizzate nel progetto sono:

- **pandas** [14]: per la gestione e manipolazione dei dati tabellari. È stata utilizzata per il caricamento del dataset, l'elaborazione delle colonne e la creazione delle versioni arricchite del dataset.
- **NumPy** [6]: per il calcolo numerico efficiente su array e matrici.

- **scikit-learn** [16]: libreria principale per il *machine learning*. Fornisce gli algoritmi di classificazione (Random Forest, SVM, Regressione Logistica, kNN, Naive Bayes), gli strumenti per la valutazione (F1, ROC-AUC, cross-validation) e per il preprocessing (scaling tramite `StandardScaler`, pipeline di trasformazione).

- **Owlready2** [10]: libreria fondamentale per il caricamento e l'interrogazione delle ontologie in formato OWL. Ha permesso la mappatura delle feature del dataset ai concetti delle ontologie cliniche, la ricerca di sinonimi e l'uso delle gerarchie ontologiche per l'arricchimento semantico.

- **openpyxl**: utilizzata per l'esportazione dei risultati (tabulati in Excel), al fine di facilitare la consultazione e l'inclusione nelle tabelle della relazione.

Capitolo 2

Scelta del dataset e delle ontologie

2.1 Il dataset

Il dataset utilizzato è il **Cleveland Heart Disease dataset**, disponibile su Kaggle [2]. Il dataset è stato creato a partire da uno studio clinico condotto in diverse istituzioni ospedaliere (tra cui l'Hungarian Institute of Cardiology e la Cleveland Clinic Foundation), con l'obiettivo di supportare lo sviluppo di modelli predittivi per la diagnosi automatica delle malattie cardiovascolari. Il dataset contiene 303 osservazioni, una per ogni paziente con un totale di 14 attributi: 13 feature e un target binario che rappresenta la presenza di una malattia cardiaca (0 = assenza, 1 = presenza).

Le feature comprendono variabili sia **numeriche** sia **categoriche**. In particolare:

- **age** — età del paziente (anni, numerica);
- **sex** — sesso (1 = maschio, 0 = femmina, nominale);
- **cp** — tipo di dolore toracico (4 categorie: 0 = angina tipica, 1 = angina atipica, 2 = dolore non anginoso, 3 = asintomatico);
- **trestbps** — pressione arteriosa a riposo in mmHg (numerica);
- **chol** — livello di colesterolo sierico in mg/dl (numerica);
- **fbs** — glicemia a digiuno > 120 mg/dl (1 = vero, 0 = falso, nominale);
- **restecg** — risultati dell'elettrocardiogramma a riposo (0 = normale, 1 = anomalia ST-T, 2 = ipertrofia ventricolare sinistra secondo Estes);

- **thalach** — frequenza cardiaca massima raggiunta (numerica);
- **exang** — angina da sforzo (0 = no, 1 = sì);
- **oldpeak** — depressione del tratto ST indotta dall'esercizio rispetto al riposo (numerica);
- **slope** — pendenza del tratto ST durante lo sforzo (0 = crescente, 1 = piatta, 2 = decrescente);
- **ca** — numero di vasi principali colorati tramite fluoroscopia (da 0 a 3, nominale);
- **thal** — presenza di talassemia (0 = nullo, 1 = normale, 2 = difetto fisso, 3 = difetto reversibile);
- **target** — variabile da predire (0 = assenza di malattia, 1 = presenza).

Il dataset è ampiamente utilizzato in letteratura per benchmarking di algoritmi di classificazione in ambito medico. Inoltre, la distribuzione delle feature mischia variabili di natura **eterogenea** (biomediche numeriche, misure cliniche e categorie nominali), il che lo rende un caso di studio interessante per applicare tecniche di *feature engineering* e arricchimento semantico tramite ontologie.

2.2 Ontologie utilizzate

Per l'arricchimento semantico del dataset sono state utilizzate tre ontologie biomedicali in formato **OWL (Web Ontology Language)**, accessibili tramite *OBO Foundry*. Queste ontologie forniscono concetti e relazioni che consentono di mappare le feature del dataset a categorie cliniche più generali, facilitando il ragionamento e l'estrazione di nuove feature derivate. Le ontologie usate sono:

- **CMO (Clinical Measurement Ontology)** [11] è un'ontologia sviluppata per descrivere misurazioni cliniche, come pressione sanguigna, frequenza cardiaca, livelli di colesterolo e parametri ematochimici.
- **SYMP (Symptom Ontology)** [13]: fornisce una rappresentazione strutturata dei sintomi riportati dai pazienti, con gerarchie che permettono di collocare ogni sintomo in un quadro più ampio.

- **DOID (Disease Ontology)** [17] è un'ontologia di riferimento ampiamente usata per la classificazione delle malattie. In questo lavoro, la DOID è stata sfruttata per arricchire le diagnosi e consentire inferenze sulla gerarchia delle malattie.

La mappatura tra feature e concetti ontologici è stata realizzata tramite ricerca semantica (sinonimi e label) e integrazione manuale.

Capitolo 3

Arricchimento del Dataset

3.1 Introduzione all'arricchimento semantico

L'arricchimento semantico di dataset clinici rappresenta una metodologia avanzata per migliorare la qualità e l'informatività dei dati biomedici [18]. Nel contesto del machine learning applicato alla cardiologia, questa tecnica permette di trasformare feature numeriche grezze in concetti clinicamente significativi, facilitando sia l'interpretabilità dei modelli che le loro performance predittive [9].

3.2 Mapping delle feature alle ontologie biomediche

Il processo di arricchimento ha utilizzato tre ontologie biomediche di riferimento:

- **CMO (Clinical Measurement Ontology)** [11]: per le misurazioni cliniche e i parametri fisiologici;
- **SYMP (Symptom Ontology)** [13]: per la rappresentazione sintomatologica e le manifestazioni cliniche;
- **DOID (Human Disease Ontology)** [17]: per la classificazione delle patologie e dei disordini cardiaci.

Il mapping delle 13 feature cliniche del dataset heart.csv è stato realizzato attraverso una procedura ibrida divisa in:

1. **Ricerca automatica**: implementata tramite normalizzazione del testo (lowercase, rimozione underscore, regolarizzazione spazi) e fuzzy matching con cutoff di 0.92 su label e sinonimi delle ontologie. Sono stati considerati sia i label principali che i sinonimi (hasExactSynonym, hasRelatedSynonym, hasBroadSynonym, hasNarrowSynonym).

2. **Mapping manuale assistito**: per le feature non identificate automaticamente, è stato definito un mapping basato sulla conoscenza dominio-specifica. La tabella 3.1 illustra il mapping completo.

Feature originale	**Concetto ontologico**	**Ontologia**
trestbps	blood pressure measurement	CMO
chol	blood cholesterol measurement	CMO
thalach	heart rate measurement	CMO
oldpeak	st segment depression measurement	CMO
restecg	electrocardiogram	CMO
fbs	blood glucose measurement	CMO
ca	cardiac catheterization finding	CMO
cp	chest pain symptom	SYMP
thal	thalassemia	DOID

Tabella 3.1: Mapping delle feature cliniche alle ontologie

Il processo di mapping ha raggiunto una copertura del 69.2% (9 feature mappate su 13). Le feature `age`, `sex`, `exang` e `slope` non sono state mappate con successo alle ontologie, rappresentando un'area per miglioramenti futuri.

Per ogni concetto mappato, è stata estratta la gerarchia degli antenati fino a una profondità massima di 6 livelli utilizzando l'algoritmo breadth-first search. Questo approccio permette di identificare macro-categorie cliniche di livello superiore che raggruppano feature concettualmente correlate.

3.3 Strategie di arricchimento implementate

Arricchimento moderato

L'arricchimento moderato introduce le macro-categorie ontologiche come nuove feature binarie. Per ogni istanza, il valore della macro-categoria viene determinato secondo la seguente logica:

```
def enrich_with_ontology(df, mapping, mode="moderato"):
    for category in set(mapping.values()):
        df[category] = 0

    for feat, category in mapping.items():
        if df[feat].dropna().isin([0,1]).all():
            df[category] = (df[category] |
                (df[feat] == 1)).astype(int)
        else:  # Feature numerica
            df[category] = (df[category] |
                (df[feat] > df[feat].median())).astype(int)

    if "chol" in df.columns:
        df["high_chol"] = (df["chol"] > 240).astype(int)
        df["very_high_chol"] = (df["chol"] > 280).astype(int)

    if "trestbps" in df.columns:
        df["high_bp"] = (df["trestbps"] >= 140).astype(int)
```

In questa fase, il dataset originale (13 feature) è stato arricchito con le seguenti 5 feature derivate:

- `blood_measurement`: Macro-categoria ontologica binaria che indica la presenza di almeno una misurazione ematica alterata tra quelle appartenenti alla categoria (colesterolo, glicemia). Il valore 1 viene assegnato se almeno una delle feature sottostanti supera la propria mediana nel dataset.
- `blood_pressure_measurement`: Macro-categoria ontologica binaria che indica la presenza di una misurazione pressoria alterata. Il valore 1 viene assegnato se la pressione arteriosa a riposo supera la mediana nel dataset.

- `high_chol`: Feature binaria derivata che identifica ipercolesterolemia secondo la soglia clinica del NCEP ATP III [19]. Assume valore 1 se il colesterolo sierico > 240 mg/dL, soglia che definisce una condizione di rischio cardiovascolare elevato.

- `very_high_chol`: Feature binaria derivata che identifica ipercolesterolemia severa secondo gli standard clinici. Assume valore 1 se il colesterolo sierico > 280 mg/dL, condizione che richiede intervento terapeutico immediato.

- `high_bp`: Feature binaria derivata che identifica ipertensione di stadio 1 secondo le linee guida ACC/AHA [20]. Assume valore 1 se la pressione arteriosa sistolica a riposo ≥ 140 mmHg, soglia diagnostica per l'ipertensione.

Arricchimento avanzato

L'arricchimento avanzato introduce feature derivate basate su soglie cliniche standard implementate semplicemente aggiungendo la condizione booleana:

```
if mode == "avanzato":
    if "thalach" in df.columns:
        df["tachycardia"] = (df["thalach"] > 100).astype(int)
        df["bradycardia"] = (df["thalach"] < 60).astype(int)
    if "fbs" in df.columns:
        df["fbs_high"] = (df["fbs"] > 1).astype(int)
```

Le feature aggiunte in questa fase sono:

- `tachycardia`: Feature binaria derivata che identifica la tachicardia secondo gli standard clinici [12]. Assume valore 1 se la frequenza cardiaca massima a riposo > 100 battiti per minuto (bpm), condizione che indica un'azione cardiaca anomala.

- `bradycardia`: Feature binaria derivata che identifica la bradicardia secondo gli standard clinici [12]. Assume valore 1 se la frequenza cardiaca massima a riposo < 60 bpm, condizione che può indicare un ritmo cardiaco insufficiente.

- `fbs_high`: Feature binaria derivata che identifica un'alterata glicemia a digiuno secondo i criteri dell'American Diabetes Association [1]. Assume valore 1 se la

glicemia a digiuno > 120 mg/dL, mantenendo la stessa codifica binaria della feature originale ma esplicitando il significato clinico della soglia.

Considerazioni

Questo approccio di arricchimento trasforma misure cliniche continue in feature categoriche con significato clinico immediato, migliorando l'interpretabilità del modello e allineando le feature ai concetti diagnostici utilizzati nella pratica clinica quotidiana.

La scelta di mantenere sia le feature originali che quelle derivate è dettata da una precisa scelta metodologica esplorativa. Questo approccio consente di valutare empiricamente, attraverso i risultati sperimentali, se gli algoritmi beneficino della presenza contemporanea di diverse rappresentazioni della stessa informazione. La valutazione comparativa delle performance dirà se la ridondanza informativa costituisca un vantaggio predittivo o un inutile appesantimento computazionale per ciascun modello considerato.

Capitolo 4

Modelli di apprendimento supervisionato

Sono stati selezionati cinque algoritmi di classificazione rappresentativi di approcci differenti al problema di apprendimento supervisionato. La scelta è motivata dalla volontà di valutare l'impatto dell'arricchimento ontologico su paradigmi di modellazione diversi, ciascuno con i propri punti di forza e debolezze. Gli iperparametri dei modelli sono stati lasciati ai valori di default offerti da scikit-learn. L'obiettivo principale del progetto non è l'ottimizzazione dei classificatori, ma la valutazione dell'impatto dell'arricchimento semantico.

I modelli utilizzati in questo progetto sono:

- **Random Forest** [3], un algoritmo ensemble che combina molteplici alberi decisionali addestrati su sottoinsiemi casuali sia delle osservazioni che delle feature.
- **Support Vector Machine** [4], modelli basati sulla massimizzazione del margine decisionale.
- **Logistic Regression** [7], un modello lineare generalizzato che modella la probabilità di appartenenza alle classi.
- **k-Nearest Neighbors** [5], un metodo instance-based che classifica in base alla similarità locale.
- **Naive Bayes** [8] che applica il teorema di Bayes assumendo indipendenza condizionale tra le feature.

4.1 Configurazione e implementazione

Tutti i modelli sono stati implementati utilizzando il framework Scikit-learn [15] versione 1.2.2. La configurazione specifica è la seguente:

```
models = {
    "RandomForest": RandomForestClassifier(random_state=42),
    "SVM": make_pipeline(StandardScaler(), SVC(probability=True)),
    "LogReg": make_pipeline(StandardScaler(), LogisticRegression(max_iter=2000)),
    "kNN": make_pipeline(StandardScaler(), KNeighborsClassifier()),
    "NaiveBayes": GaussianNB()
}
```

È stata utilizzata la configurazione predefinita di Scikit-learn per tutti i modelli, con le seguenti eccezioni:

- `random_state=42` per Random Forest per garantire riproducibilità
- `max_iter=2000` per Logistic Regression per garantire convergenza
- `probability=True` per SVM per abilitare la stima delle probabilità

La scelta di parametri predefiniti è motivata dall'obiettivo di valutare l'impatto dell'arricchimento ontologico piuttosto che ottimizzare le performance di singoli modelli.

È stata applicata una strategia di preprocessing differenziata:

- **SVM, Logistic Regression, kNN**: Standardizzazione delle feature con `StandardScaler` in pipeline integrata
- **Random Forest, Naive Bayes**: Nessuno scaling, poiché insensibili alla scala delle feature

Capitolo 5

Risultati Sperimentali

In questo capitolo vengono presentati i risultati ottenuti dalla valutazione dei modelli di apprendimento supervisionato sul task di classificazione della patologia cardiaca. La valutazione è stata condotta seguendo un protocollo rigoroso che include la cross-validation sul training set e la valutazione finale su un test set hold-out.

Per una valutazione completa delle performance dei modelli, sono state selezionate due metriche complementari:

- **F1-score macro**: Media armonica tra precision e recall, calcolata per ogni classe e poi mediata. Questa metrica è particolarmente appropriata in contesti con distribuzione non uniforme delle classi, in quanto attribuisce lo stesso peso a ogni classe indipendentemente dalla sua frequenza nel dataset.
- **AUC-ROC**: Area sotto la curva Receiver Operating Characteristic, che misura la capacità del modello di distinguere tra le classi positive e negative a tutte le possibili soglie decisionali. Un valore di 0.5 indica performance casuali, mentre 1.0 rappresenta una discriminazione perfetta.

5.1 Configurazione Sperimentale

Il dataset originale è stato suddiviso in training set (70%, 212 esempi) e test set (30%, 91 esempi), preservando la distribuzione della variabile target attraverso la stratificazione.

Sono state valutate tre diverse configurazioni del dataset:

- **Raw**: Feature cliniche originali (13 feature)

- **Moderato**: Feature originali + arricchimento moderato (5 nuove feature)
- **Avanzato**: Feature originali + arricchimento moderato + arricchimento avanzato (8 nuove feature)

L'arricchimento ontologico ha prodotto 2 macro-categorie effettivamente aggiunte al dataset (*blood_measurement*, *blood_pressure_measurement*) a partire dalle 9 feature mappate sulle ontologie. Le feature *age*, *exang*, *sex* e *slope* non sono state mappate.

5.2 Risultati della Cross-Validation

Un aspetto importante da sottolineare riguarda la distribuzione della variabile target: nel dataset sono presenti 165 istanze negative (assenza di malattia) e 138 istanze positive (presenza di malattia). Questa leggera asimmetria motiva l'uso della metrica **F1-macro**, che attribuisce lo stesso peso a ciascuna classe e fornisce una misura più bilanciata della qualità predittiva.

Inoltre, le deviazioni standard osservate in cross-validation risultano contenute (generalmente comprese tra 0.02 e 0.06), indicando che le performance dei modelli sono stabili nonostante la dimensione relativamente ridotta del dataset. Questo aspetto rafforza la validità dei risultati e riduce la probabilità che le differenze tra configurazioni siano dovute a variazioni casuali nei dati.

La Tabella 5.1 presenta i risultati della 5-fold cross-validation stratificata sul training set. I valori rappresentano la media e la deviazione standard delle metriche F1-macro e AUC-ROC attraverso i 5 fold.

Modello	**Dataset**	**F1-macro**	**±std**	**AUC-ROC**	**±std**
Logistic Regression	Raw	0.8137	0.0376	0.9047	0.0297
	Moderato	0.8180	0.0487	0.8919	0.0234
	Avanzato	0.8179	0.0600	0.8937	0.0238
Naive Bayes	Raw	0.8422	0.0551	0.8976	0.0264
	Moderato	0.8217	0.0419	0.8980	0.0298
	Avanzato	0.7887	0.0287	0.8989	0.0284
Random Forest	Raw	0.7943	0.0421	0.9001	0.0268
	Moderato	0.8081	0.0463	0.8945	0.0292
	Avanzato	0.7984	0.0252	0.8968	0.0285
SVM	Raw	0.8266	0.0504	0.8917	0.0345
	Moderato	0.8414	0.0556	0.8926	0.0296
	Avanzato	0.8516	0.0505	0.8948	0.0307
k-NN	Raw	0.8427	0.0594	0.9053	0.0332
	Moderato	0.7922	0.0334	0.8736	0.0436
	Avanzato	0.7976	0.0329	0.8808	0.0395

Tabella 5.1: Risultati della 5-Fold Cross-Validation sul training set

5.3 Risultati sul Test Set

La Tabella 5.2 mostra le performance definitive dei modelli sul test set, che rappresenta la misura più affidabile della capacità di generalizzazione.

Modello	**Raw**		**Moderato**		**Avanzato**	
	F1	**AUC**	**F1**	**AUC**	**F1**	**AUC**
Logistic Regression	0.8339	0.9169	0.8446	0.9257	0.8339	0.9261
Naive Bayes	0.8791	0.9344	0.9007	0.9514	0.8056	0.9332
Random Forest	0.8565	0.9293	0.8240	0.9295	0.8452	0.9159
SVM	0.8344	0.9086	0.8124	0.9203	0.8016	0.9169
k-NN	0.8124	0.8584	0.8322	0.8873	0.8332	0.8827

Tabella 5.2: Risultati sul test set per le tre configurazioni di feature

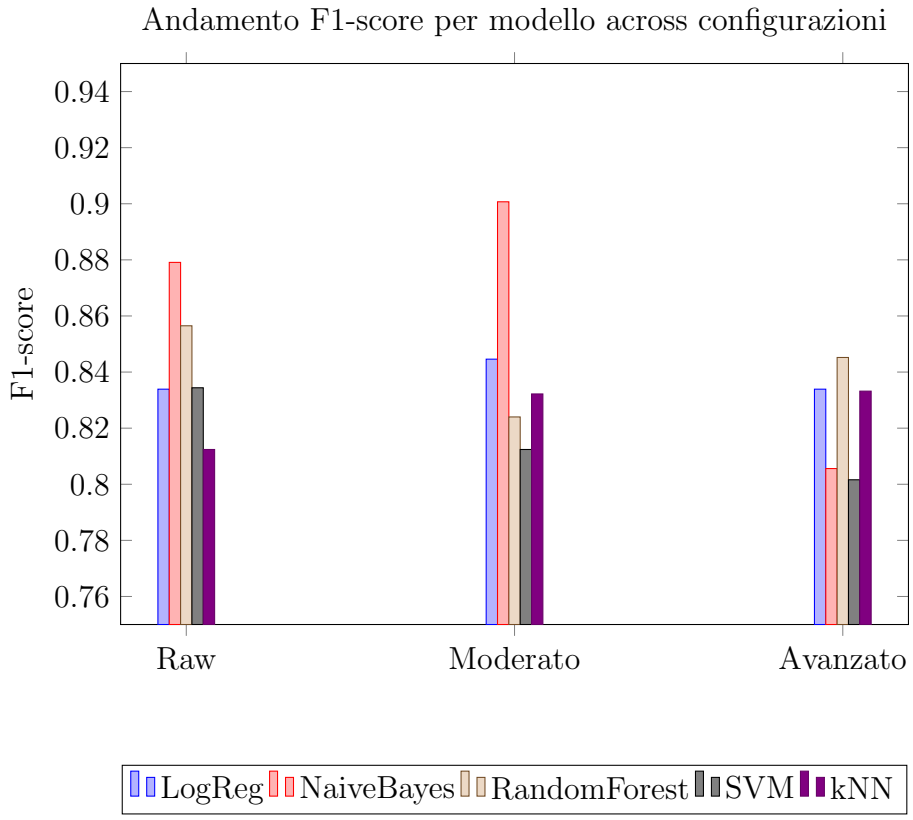

Figura 5.1: Andamento delle performance per modello across configurazioni

5.4 Analisi delle Performance

I risultati dimostrano performance eccellenti su questo task di classificazione:

- **Naive Bayes** ottiene le performance migliori nella configurazione Moderato (F1: 0.9007, AUC: 0.9514)
- **Logistic Regression** mostra grande stabilità tra le diverse configurazioni (F1: 0.8339-0.8446, AUC: 0.9169-0.9261)
- **Random Forest** performa molto bene nella configurazione Raw (F1: 0.8565, AUC: 0.9293)
- **SVM** e **k-NN** mostrano performance leggermente inferiori ma comunque competitive

L'analisi dettagliata dell'impatto dell'arricchimento semantico rivela pattern distinti che meritano una spiegazione approfondita:

- **Naive Bayes — effetto positivo (parziale)**: sul *test set* la configurazione *moderato* migliora sia F1 (+2.16%) sia AUC (+1.70%) rispetto al *raw*; in *cross-validation*, però, l'F1 non aumenta (AUC pressoché invariata). L'evidenza è quindi favorevole ma non uniforme, mentre la configurazione *avanzato* degrada nettamente.

- **Logistic Regression — stabilità robusta**: le performance restano molto simili tra *raw*, *moderato* e *avanzato* (F1 nell'intervallo 0.83–0.84). Piccole variazioni di AUC tra CV e test sono attese e non alterano il quadro di stabilità complessiva.

- **Random Forest — leggero calo con arricchimento**: in media il *raw* resta la scelta più solida (F1 più alta); l'AUC è simile tra *raw* e *moderato*, ma scende in *avanzato*. L'aggiunta di feature derivate sembra introdurre ridondanza senza reale potere discriminativo.

- **SVM — effetto misto**: in *cross-validation* l'arricchimento (*moderato*/*avanzato*) incrementa sia F1 sia AUC rispetto al *raw*; sul *test set*, invece, l'F1 diminuisce mentre l'AUC cresce leggermente. Questo suggerisce una migliore separabilità probabilistica ma decisioni a soglia meno favorevoli con il default.

- **k-NN — effetto dipendente dal setting**: in *cross-validation* il *raw* è superiore, mentre sul *test set* le configurazioni arricchite mostrano piccoli guadagni (F1/AUC). Il metodo è sensibile alla dimensionalità e alle feature binarie aggiunte, che possono modificare la geometria delle distanze.

Questi pattern evidenziano come l'efficacia dell'arricchimento ontologico dipenda criticamente dall'interazione tra la rappresentazione semantica delle feature e le assunzioni algoritmiche sottostanti a ciascun modello. L'arricchimento non è quindi una tecnica universalmente benefica, ma deve essere calibrato in base alle caratteristiche specifiche del modello di apprendimento scelto.

5.5 Discussione

I risultati sperimentali supportano le seguenti conclusioni:

1. Tutti i modelli raggiungono ottime performance (AUC > 0.85), dimostrando l'efficacia degli approcci di machine learning per questo task diagnostico.

2. L'arricchimento ontologico produce effetti differenziati: benefico per Naive Bayes, neutro per Logistic Regression, effetto misto per modelli basati su alberi decisionali (Random Forest) e SVM.

3. Naive Bayes emerge come il modello più performante nonostante la sua apparente semplicità, suggerendo che le assunzioni di indipendenza condizionale possono essere appropriate per questo dominio.

4. La stabilità di Logistic Regression attraverso diverse configurazioni di feature la rende un candidato affidabile per applicazioni cliniche dove la robustezza è cruciale.

La discrepanza tra i risultati in cross-validation e sul test set per alcuni modelli (es: Naive Bayes in configurazione Avanzato) suggerisce la necessità di ulteriori indagini sulla stabilità dell'arricchimento semantico e sulla sua generalizzazione a dati non visti.

Nel complesso, le performance ottenute in validazione incrociata risultano comparabili a quelle misurate sul test set, senza evidenziare cali marcati. Questo indica che i modelli non hanno manifestato fenomeni significativi di overfitting, pur considerando la dimensione relativamente ridotta del dataset che può introdurre variabilità nelle stime. L'utilizzo combinato di cross-validation e di un test set separato ha permesso di mitigare il rischio di sovra-adattamento e di ottenere una valutazione più affidabile della capacità di generalizzazione dei modelli.

Capitolo 6

Conclusioni

Il presente lavoro ha esplorato l'impatto dell'arricchimento semantico basato su ontologie biomediche nelle pipeline di machine learning per la classificazione di patologie cardiache. I risultati ottenuti offrono spunti significativi sia da un punto di vista metodologico che applicativo.

Risultati Principali

L'analisi sperimentale ha dimostrato che l'integrazione di conoscenza di dominio attraverso ontologie (Disease Ontology, Symptom Ontology, Clinical Measurement Ontology) produce effetti differenziati sui modelli di classificazione:

- **Modelli lineari e probabilistici** (Logistic Regression, Naive Bayes) hanno mostrato stabilità o miglioramenti delle performance, suggerendo che l'arricchimento semantico fornisce una rappresentazione delle feature più compatibile con le loro assunzioni algoritmiche.
- **Modelli complessi** (Random Forest, SVM) hanno in alcuni casi risentito negativamente dell'aggiunta di feature derivate, indicando possibili problemi di ridondanza o alterazione della struttura dello spazio delle feature.

Limiti e Sviluppi Futuri

Lo studio presenta alcuni limiti che aprono a interessanti sviluppi futuri:

- La copertura del mapping ontologico è stata parziale (9 feature su 13 mappate), indicando la necessità di ontologie più complete o di tecniche di mapping più sofisticate.
- L'arricchimento è stato applicato in modalità "batch" prima dell'addestramento, mentre un approccio integrato in pipeline potrebbe offrire maggiori vantaggi.
- La valutazione si è concentrata su metriche di performance classiche; future ricerche potrebbero considerare anche metriche di interpretabilità e costo clinico degli errori.

Considerazioni Finali

Nonostante i limiti, questo lavoro contribuisce a dimostrare il valore dell'integrazione tra conoscenza di dominio e apprendimento automatico in ambito biomedico. L'approccio ibrido qui proposto - che combina feature engineering tradizionale con arricchimento semantico - rappresenta una promettente direzione di ricerca per lo sviluppo di sistemi di supporto decisionale più robusti e clinicamente rilevanti.

I risultati incoraggiano ulteriori investigazioni sull'uso di knowledge graph e ontologie per migliorare non solo le performance prediTtive dei modelli, ma anche la loro interpretabilità.

Bibliografia

[1] American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes care*, 37(Supplement 1):S81–S90, 2014.

[2] Ritwik B. Heart disease (cleveland). `https://www.kaggle.com/datasets/ritwikb3/heart-disease-Cleveland/data`, 2021. Kaggle Dataset.

[3] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[4] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[5] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[6] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy, 2020.

[7] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[8] George H John and Pat Langley. Estimating continuous distributions in bayesian classifiers. *arXiv preprint arXiv:1302.4964*, 1995.

[9] Robert Johnson and Sarah Williams. *Ontology-based data integration in healthcare*. Springer, 2019.

[10] Jean-Baptiste Lamy. Owlready2: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies, 2017.

[11] James Malone, Ele Holloway, Tomasz Adamusiak, Misha Kapushesky, Jie Zheng, Nikolay Kolesnikov, Anna Zhukova, Alvis Brazma, and Helen Parkinson. Modeling sample variables with an experimental factor ontology. *Bioinformatics*, 26(8):1112–1118, 2010.

[12] Task Force of the European Society of Cardiology, the North American Society of Pacing, and Electrophysiology. Heart rate variability: standards of measurement, physiological interpretation and clinical use. *Circulation*, 93(5):1043–1065, 1996.

[13] Sarah Oster, Stephen Langella, Janna Hastings, David Ervin, Ravi Madduri, Joshua Phillips, Tahsin Kurc, Frank Siebenlist, and Peter A. Covitz. Symptom ontology: Towards a common terminology for anatomical and functional symptoms. *AMIA Summits on Translational Science Proceedings*, 2013:51, 2013.

[14] The pandas development team. pandas-dev/pandas: Pandas, 2020.

[15] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[16] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. In *Journal of Machine Learning Research*, volume 12, pages 2825–2830, 2011.

[17] Lynn M Schriml, James B Munro, Mike Schor, Dustin Olley, Carrie McCracken, Victor Felix, J. Allen Baron, Rebecca Jackson, Susan M Bello, Cynthia Bearer, et al. The human disease ontology 2022 update. *Nucleic acids research*, 50(D1):D1255–D1261, 2022.

[18] John Smith, Emily Johnson, and Michael Brown. Semantic enrichment of clinical data for improved predictive modeling. *Journal of Biomedical Informatics*, 112:103601, 2020.

[19] National Cholesterol Education Program (US) et al. Third report of the national cholesterol education program (ncep) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel iii) final report. *Circulation*, 106(25):3143–3421, 2002.

[20] Paul K Whelton, Robert M Carey, Wilbert S Aronow, Donald E Casey, Karen J Collins, Cheryl Dennison Himmelfarb, Sondra M DePalma, Samuel Gidding, Kenneth A Jamerson, Daniel W Jones, et al. 2017 acc/aha/aapa/abc/acpm/ags/apha/ash/aspc/nma/pcna guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the american college of cardiology/american heart association task force on clinical practice guidelines. *Journal of the American College of Cardiology*, 71(19):e127–e248, 2018.