Progetto Machine Learning – Febbraio 2020

MALATTIE CARDIOVASCOLARI: ANALISI E PREDIZIONE DEL RISCHIO DI SVILUPPO

TEAM 69[1]: Giulia Mura (MAT: 860910), Caterina Pisani (MAT: 853058), Filippo Maria Casula (MAT: 860894) , Jacopo Roberto Nicosia (MAT: 861812), Alessandro Pontini (MAT: 852793)

SOMMARIO

L'utilizzo delle intelligenze artificiali sta diventando determinante in numerosi settori a causa dell'esponenziale crescita nella produzione di dati che si sta verificando nella società contemporanea. Per poter tenere il passo con questa mole di dati senza precedenti, l'implementazione di tecnologie come le AI si sta rendendo necessaria in numerosi settori, tra cui quello sanitario.
Infatti sono già innumerevoli le trasformazioni che è stato possibile portare nel sistema sanitario grazie ad esse: dall'analisi e organizzazione di dati clinici a diagnosi precoci e soluzioni migliori e mirate per i pazienti.
Un perfetto esempio pratico è fornito da progetti come Watson Health di IBM, ovvero prototipi di soluzioni basate su tecnologie di machine learning per affiancare l'attività dei medici professionisti e fornire ulteriore supporto nel processo di diagnosi.
L'orizzonte delle possibilità offerte da algoritmi predittivi in quest'ambito è costantemente in crescita e il suo peso sarà probabilmente sempre maggiore nel futuro del settore.
L'obiettivo di questo progetto è quello di esplorare le possibili soluzioni per la costruzione di uno strumento predittivo adeguato ad integrarsi se non a sostituire le classiche "carte del rischio" di sviluppo di malattie cardiovascolari.
In particolare, ci si è chiesti se fosse possibile predire l'insorgere di una malattia cardiovascolare sulla base di dati riportanti diversi fattori di rischio riguardo a 70mila pazienti oggetto di analisi.

INDICE

Introduzione 1
Presentazione del Dataset 2
Data Cleaning 3
Analisi Esplorativa 4
Modelli di classificazione 5
Analisi e Risultati 5
Performance Evaluation 5
Test d'ipotesi e Intervalli di confidenza 9
Analisi finale con Misure di Performance 9
Conclusioni 11
Riferimenti 11

INTRODUZIONE

Il rischio cardiovascolare è una misura della probabilità di subire un danno a carico del distretto cardiocircolatorio, ovvero a cuore e vasi sanguigni.

La teoria medica si serve dei cosiddetti "fattori di rischio" nel calcolo della probabilità di sviluppo di una malattia cardiovascolare, ovvero caratteristiche che predispongono all'insorgenza della malattia.

Questi sono riclassificabili in due macrocategorie:

1. Fattori di rischio non modificabili quali:
   - Sesso
   - Età

[1] Università degli Studi di Milano Bicocca, CdLM Data Science

  - Familiarità (parenti diretti che hanno avuto patologie di questo tipo)
  - Patologie congenite (ad esempio malformazioni cardiocircolatorie)
2. Fattori di rischio modificabili, e quindi correggibili, quali:
  - Terapie farmacologiche
  - Abitudini e stili di vita (ad esempio abitudine al fumo, assunzione di alcol, attività sportiva)
  - Ipertensione (condizione caratterizzata dall'elevata pressione del sangue nelle arterie)
  - Ipercolesterolemia (eccesso di colesterolo nel sangue)
  - Obesità

È possibile calcolare una stima di questo rischio ricorrendo alla "carta del rischio cardiovascolare": una tabella che consente di stimare quale sia la probabilità di incorrere in un primo evento cardiovascolare maggiore (infarto o ictus), fatale o non, nei 10 anni successivi, utilizzando come input questi fattori.

In particolare, la carta del rischio è valida se i fattori vengono misurati seguendo la metodologia standardizzata ed è utilizzabile su donne e uomini di età compresa tra 40 e 90 anni che non hanno avuto precedenti cardiovascolari. Inoltre, non può essere applicata per valori estremi dei fattori di rischio.

La domanda cui si è cercato di rispondere nello svolgimento del progetto è se fosse possibile, sulla base di alcuni fattori di rischio forniti dal dataset a disposizione, costruire un modello predittivo riguardo all'insorgere della malattia cardiovascolare. E se si, quale fosse il più adatto/preciso.

PRESENTAZIONE DEL DATASET

Il Dataset da cui si è attinto per lo sviluppo del progetto consiste di 70000 record che riportano i dati di pazienti, raccolti in fase di analisi medica.

Gli attributi presentano alcuni dei fattori di rischio sopra citati (11 totali) e rappresentano gli input per l'attributo target in forma binaria, ossia presenza o assenza di malattie cardiovascolari.

I fattori di rischio non sono suddivisi tra fattori modificabili e non modificabili (differenziazione affrontata nell'introduzione), ma vengono invece distinti in altre 3 categorie:

- Informazioni oggettive (ad esempio età, peso, altezza)
- Informazioni soggettive (fornite a discrezione paziente ad esempio fumo, alcol)
- Risultati di esami medici

Andando ad analizzare meglio gli input del dataset troviamo:

1. Età: informazione oggettiva, espressa in giorni
2. Altezza: informazione oggettiva, espressa in cm
3. Peso: informazione oggettiva, espressa in kg
4. Genere: informazione oggettiva
5. Pressione sanguigna sistolica: risultato di esami medici, misurata in mmHg
6. Pressione sanguigna diastolica: risultato di esami medici, misurata in mmHg
7. Colesterolo: risultato di esami medici, distinto tra normale, sopra il normale e molto sopra il normale (1: normale, 2: sopra il normale, 3: molto sopra il normale)
8. Glucosio: risultato di esami medici, distinto tra normale, sopra il normale e molto sopra il normale (1: normale, 2: sopra il normale, 3: molto sopra il normale)

9. Fumo: informazione soggettiva
10. Assunzione di alcol: informazione soggettiva
11. Attività fisica: informazione soggettiva

Il dodicesimo ed ultimo attributo è "cardio", la variabile dipendente binaria che, come già detto, fornisce informazioni sulla presenza o meno della malattia cardiovascolare nel paziente in oggetto.

Data Cleaning

Il dataset presentava un discreto numero di anomalie di varia natura, quali il formato dei dati inadeguato ai fini di analisi e outliers di vario tipo.

Nello specifico, prendendo in oggetto le variabili numeriche:

- L'attributo "età" dei pazienti in origine era espressa in giorni ed è quindi stata modificata in termini di anni.

- Gli attributi "altezza" e "peso" presentavano diversi outlier. In particolare si è deciso di eliminare tutti i record che presentavano valori "impossibili", sia presi singolarmente che in relazione alle altre variabili associate. Sono quindi stati conservati solo i record di individui con altezza compresa tra 130-210 cm e peso tra i 30-180 kg.

- Le rilevazioni sulla pressione sistolica e diastolica (massima e minima) presentavano outlier che non sono stati mantenuti.
  In particolare, si è deciso di conservare solo le rilevazioni comprese negli intervalli di 70-190 mmHg per quanto riguarda la pressione sistolica e 40-110 mmHg per la diastolica. Infatti, si ricorda che indicativamente sono considerati normali i valori tra 120-130 mmHg per la Sistolica e 70-80 mmHg per la diastolica (FIGURA 1).
  Inoltre, le rilevazioni sulla pressione diastolica erano espresse in un ordine di grandezza errato, imputabile ad un probabile errore di trascrizione, ed è stato quindi necessario dividerle nell'ordine di 10^2.

- Sia per le variabili ordinali, descrittive dei livelli di colesterolo e glucosio nel sangue, che per le binarie non è stato necessario alcun tipo di intervento.

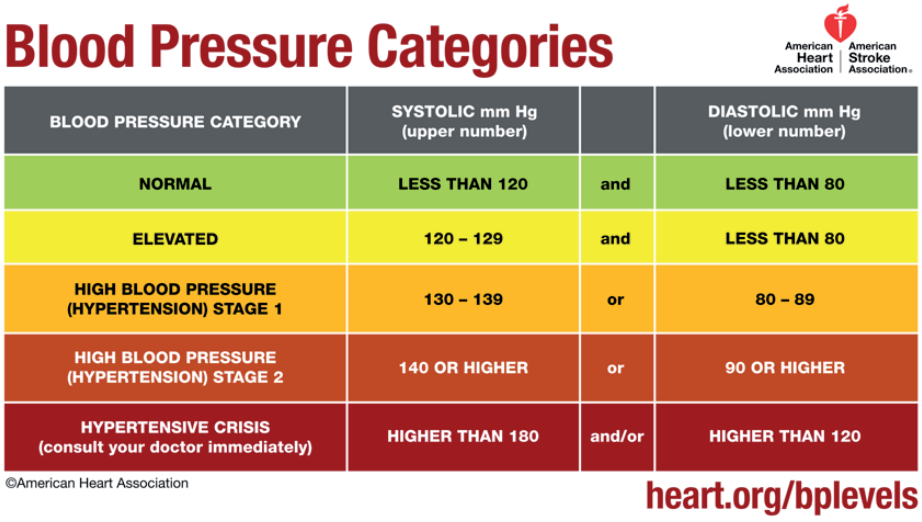

FIGURA 1: Categorie di pressione sanguigna secondo l'American Heart Association

Alla fine della fase di cleaning rimangono 68.246 delle 70.000 rilevazioni originariamente a disposizione, perdendo quindi solo il 2,5% dei dati inizialmente a disposizione.

Analisi Esplorativa

All'inizio dello studio è stato eseguito un breve processo di analisi descrittiva mediante l'utilizzo del software R, con l'obiettivo di determinare le macro-caratteristiche del dataset di riferimento e evidenziare i tratti di più immediata visualizzazione.

Il risultato di questo procedimento ha portato alle seguenti conclusioni: la variabile dipendente ("cardio") è equamente distribuita, toccando quasi il perfetto 50% (49,75% "Yes", 50,25% "No"). Questo è importante ai fini della modellazione successiva, in quanto la teoria della materia prevede approcci specifici per il trattamento di distribuzioni sbilanciate della variabile dipendente.

Si è verificato inoltre che parametri come "active", "smoke" e "alcol", a discapito di quelle che potrebbero essere le attese determinate dal comune buon senso, non risultano correlate alla distribuzione della variabile output per quanto riguarda il dataset a disposizione.

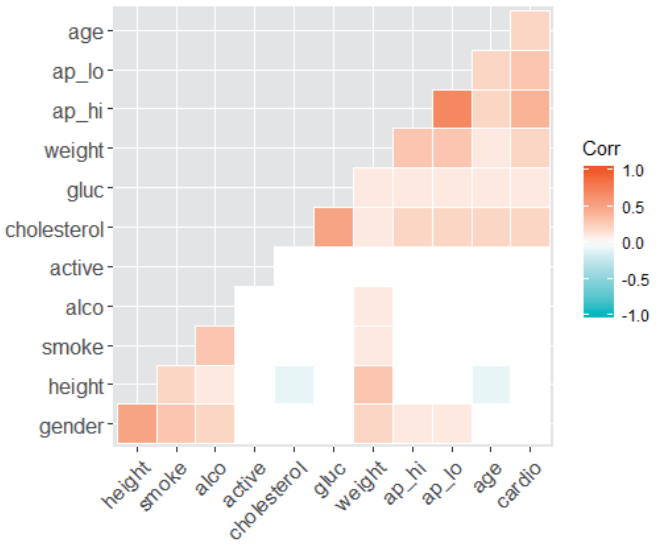

FIGURA 2: Matrice di correlazione tra fattori input e variabile target

Sembra invece più evidente la correlazione con tratti oggettivi quali età e peso, suggerendo una frequenza maggiore nell'insorgenza del tratto su soggetti più anziani e contraddistinti da peso superiore alla media.

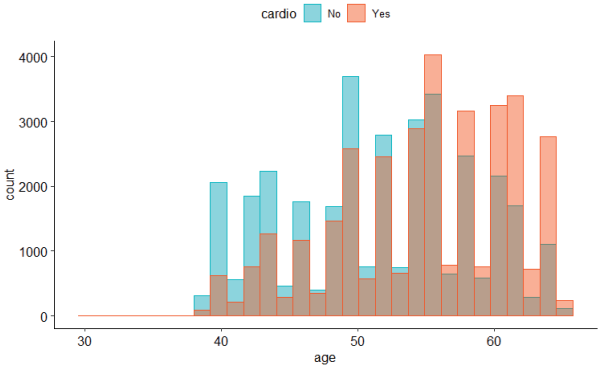

FIGURA 3: Manifestazione malattia in funzione dell'età

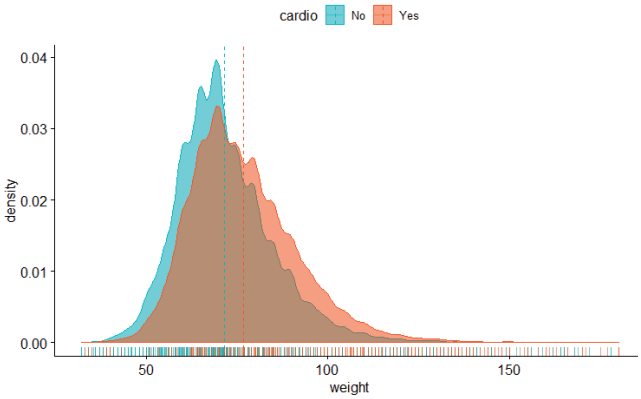

FIGURA 4: Manifestazione malattia in funzione del peso

È interessante notare che la maggior parte degli attributi che non hanno influenza sullo sviluppo della malattia cardiovascolare, in particolare alcol, smoke e active, corrispondono proprio con le informazioni classificate come "soggettive", ovvero quelle fornite dai pazienti. Questi risultati, cosi fortemente in contrasto con studi e ricerche ampiamente condivise in ambito medico, fanno ipotizzare che non tutti i pazienti abbiano fornito infor-

mazioni veritiere in merito a questi temi, viziando così i dati.

Concludendo, i risultati ottenuti dall'analisi hanno condotto ad una feature selection che eliminasse gli attributi irrilevanti ai fini della nostra analisi, i quali sono risultati essere: alcol, smoke, active, gender e height.

MODELLI DI CLASSIFICAZIONE

Per lo svolgimento del progetto si è usufruito di diversi modelli di classificazione al fine di individuare quale fosse il più adatto a predire questo genere di fenomeno, posto il dataset a disposizione.

Di seguito vengono elencati i classificatori di cui si è fatto uso, suddivisi in base alla macrocategoria del modello (di cui si riporta una breve descrizione):

- Modelli euristici: molto usati in quanto basati su schemi semplici e intuitivi.
  - Decision Tree
  - J48
  - Random Forest
- Modelli di regressione: basati sulla regressione logistica, nei quali la variabile da prevedere è dicotomica, ovvero descrittiva della presenza/assenza di una caratteristica (nel nostro caso la malattia cardiovascolare).
  - Logistic
  - Simple Logistic
- Modelli di separazione: partizionano in regioni distinte lo spazio degli attributi consentendo di separare le osservazioni appartenenti a classi differenti.
  - Multi layer perceptron (con 2, 4, 6, 8, 10 neuroni e un solo hidden layer)
  - Support Vector Machine (SPegasos e SVM – Poly)
- Modelli probabilistici: sfruttano il teorema di Bayes, fornendo un approccio di inferenza probabilistico.
  - Bayes Network (BayesNet con configurazione TANB e BNC)
  - NBTree
  - Naïve Bayes

Le performance dei modelli di classificazione sopra riportati sono state comparate tra loro tramite l'implementazione dei metodi di Holdout, Holdout Iterated, K-Folds Cross Validation simple e K-folds Cross Validation stratified.

In particolare, la comparazione è stata svolta in termini di accuracy. Secondo questa logica, si è constatato che il metodo più performante risultava essere il K-folds Cross Validation stratified e, sulla base dei suoi risultati, è stata fatta una selezione dei 6 modelli migliori, uno per ogni "macrofamiglia".

Solo sui modelli selezionati si è calcolato l'intervallo di confidenza per verificarne la validità statistica.

Infine sono stati valutati i risultati finali mediante altri criteri, in particolare Precision, Recall, F-measures, ROC e AUC.

Questi processi verranno descritti nei seguenti paragrafi.

ANALISI E RISULTATI

Performance Evaluation

Il primo step dell'analisi è stato quello di effettuare un campionamento casuale del dataset pari al 50% su cui svolgere il lavoro successivo, decisione presa allo scopo di ridurre l'onerosità computazionale richiesta alla piattaforma Knime e realizzabile grazie all'elevato numero di osservazioni a disposizione.

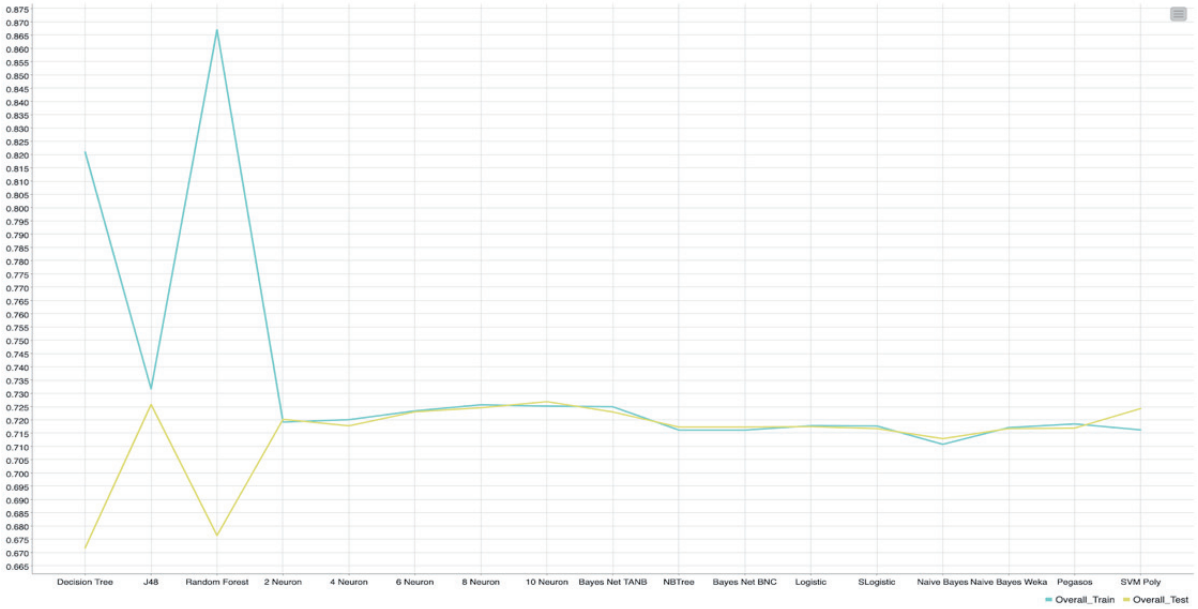

FIGURA 4: Metodo Holdout Simple

Il primo metodo di performance evaluation sviluppato è stato quello di **Holdout Simple.**

Il metodo di Holdout è l'approccio di validation più frequente, oltre che il più semplicistico. Consiste nel partizionamento del dataset in modo tale da ottenere un training set e un test set. Tipicamente si utilizzano ⅔ per alimentare l'inducer, e il rimanente ⅓ per testare i risultati.

I risultati ottenuti con questo metodo di partizionamento sono riportati in FIGURA 4.

In questo grafico, come nei successivi, la linea azzurra rappresenta il test sul training set, mentre quella gialla sul test set. Questo genere di visualizzazione è particolarmente utile in quanto permette di visualizzare contemporaneamente sia il livello di Accuracy riscontrato per ogni modello, che valutare eventuali problemi di overfitting o underfitting.

L'overfitting è un fenomeno che si riscontra quando un modello predittivo si "adatta troppo" ai dati forniti durante la fase di training. Si riconosce quando l'accuracy sul training set è significativamente più elevata rispetto a quella nel test set.

In questo caso possiamo riscontrare che i modelli Random Forest e Decision Tree presentano tali caratteristiche, presentando non solo accuracy ben al di sopra della media degli altri modelli sul training set, ma anche più bassi per il test set.

Riguardo agli altri modelli, hanno restituito risultati abbastanza omogenei.

Per quanto concerne l'analisi dei metodi utilizzati di seguito, è importante specificare che i risultati in termini di Accuracy, non essendo più stime puntuali bensì intervalli, sono stati "condensati" in medie, così da rendere omogeneo e più facilmente visualizzabile il confronto tra i metodi.

Il metodo presentato di seguito è quello di **Iterated Holdout**. Questo consiste nella ripetizione del metodo Holdout un R numero di volte, ovvero nell'iterazione del processo di training-testing. Tale approccio permette di li-

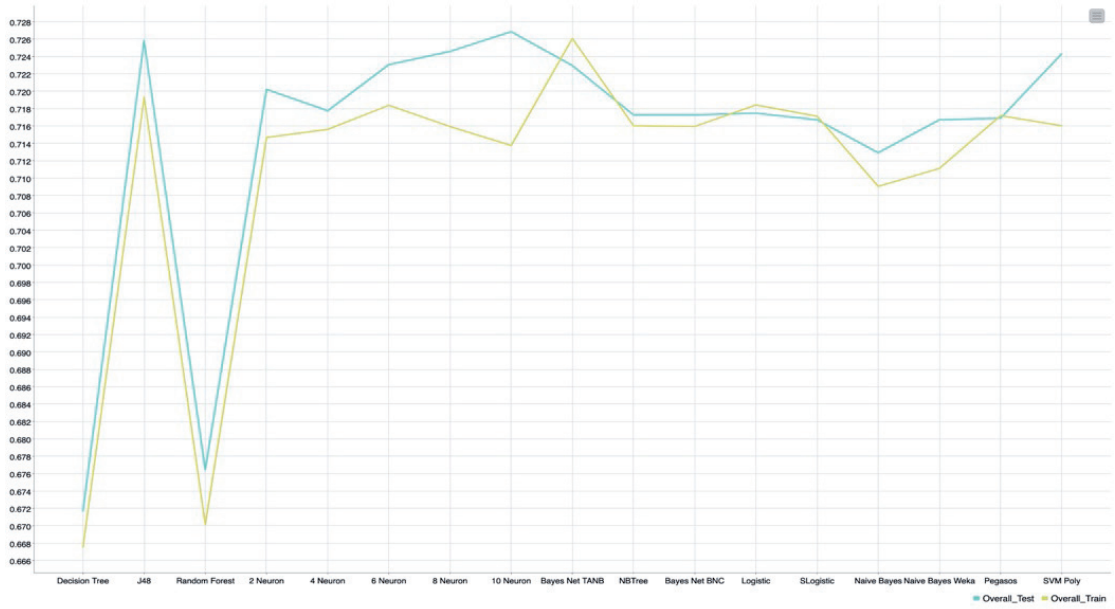

FIGURA 5: Metodo Holdout Iterated

mitare i possibili bias causabili all'inducer dal semplice partizionamento iniziale.

Il miglioramento rispetto all'Holdout Simple è evidenziato dal grafico riportato in FIGURA 5.

Il miglioramento è riscontrabile non tanto in termini del valore dell'accuracy, che rimane quasi invariata per buona parte dei modelli, quanto più in un netto aumento dell'omogeneità complessiva della distanza tra l'accuracy di train e di test.

Successivamente è stato esplorato il processo di **Cross Validation**, metodo nel quale il dataset viene suddiviso in modo casuale in k sottogruppi di pari dimensioni. Di questi ne viene trattenuto uno che viene utilizzato come test-set, mentre quelli rimanenti alimentano l'inducer nella fase di training. Il procedimento di training-testing viene ripetuto fino a quando tutti i sottogruppi hanno svolto la funzione di training set una volta.

Quest'ultimo metodo e il precedente sembrano restituire risultati molto simili tra loro sia in termini di Accuracy che di omogeneità tra test e train dei modelli e non sono particolarmente significativi per la prosecuzione dell'analisi. Per tale motivo non viene riportata la rappresentazione grafica del metodo di Cross Validation.

I migliori risultati si sono infatti ottenuti mediante il metodo di **Cross Validation Stratified**, variante del metodo di Cross Validation che consiste nel suddividere il dataset iniziale (la parte di dataset adibita al ruolo di train) in k sottogruppi, i quali hanno come condizione quella di mantenere costante al loro interno la proporzione originale del dataset.

I risultati di questo metodo si caratterizzano non tanto in termini di maggiore accuracy, che complessivamente risulta piuttosto simile a quella già rilevata con altri metodi, quanto nel presentarsi con una minore variabilità in funzione del fatto che venga calcolata sul training set o sul test set.

In particolare sono stati restituiti risultati complessivamente superiori nei test sul 1 set rispetto agli altri modelli. Questo riduce le probabilità che il livello di accuracy registrato sia causato da Overfitting.

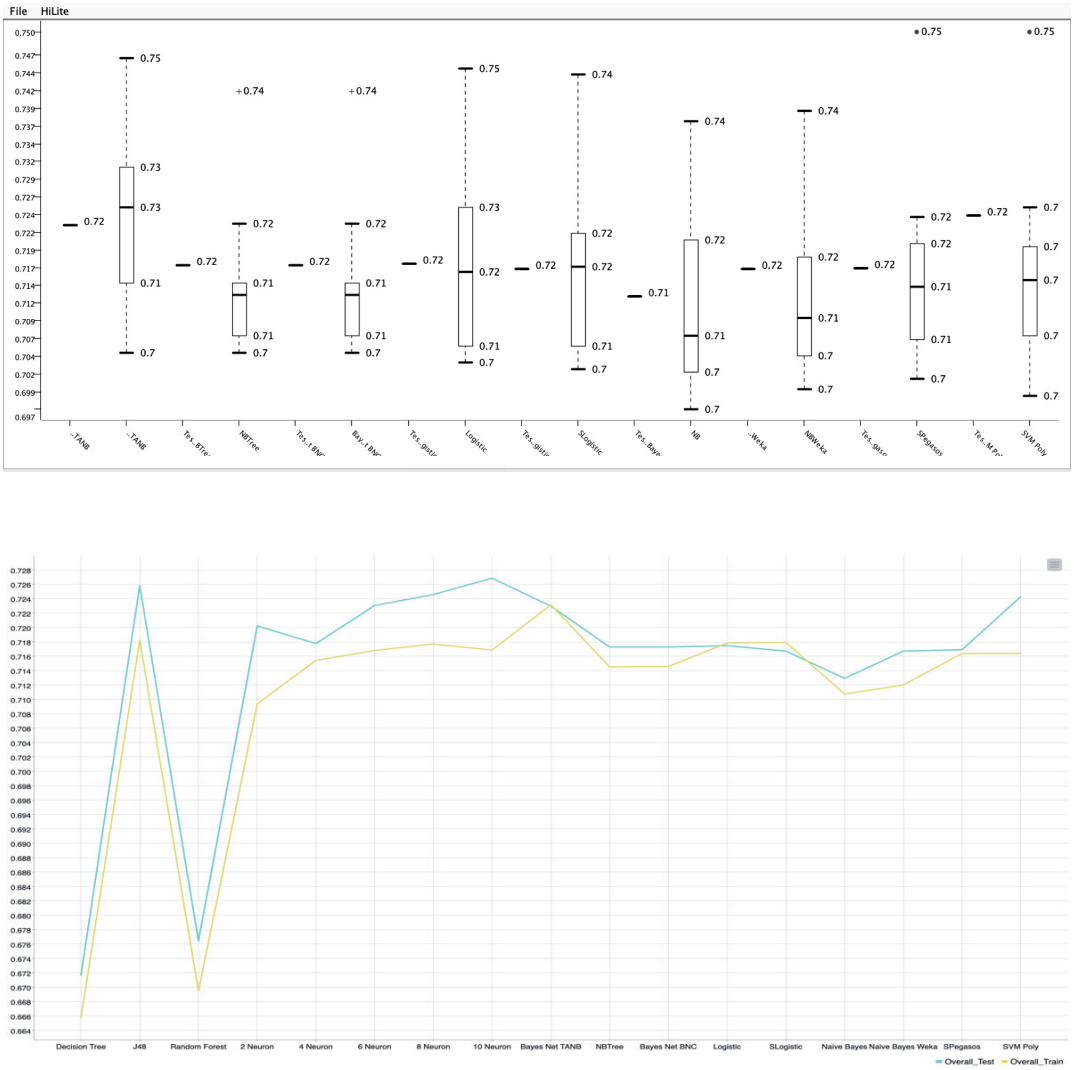

FIGURA 6: Metodo Cross Validation Stratified

Per completezza, viene mostrato anche il box-plot con i valori ottenuti da questo metodo (FIGURA 6).

Date queste considerazioni, si è deciso di utilizzare quest'ultimo procedimento di validation per la prima selezione dei modelli predittivi.

Si è infatti deciso di conservare un solo modello per ogni categoria, selezionandolo in base al più alto livello di accuracy tra quelli restituiti dal Cross Validation Stratified. Di seguito vengono elencati i modelli scelti e la categoria di appartenenza assegnata nel workflow:

1. Support Vector Machine: SVM Poly
2. Logistic: Simple Logistic
3. Naive Bayes: Naive Bayes Weka
4. MLP: 8 Neurons
5. Euristic: J48
6. Bayesian: Bayes Net TANB

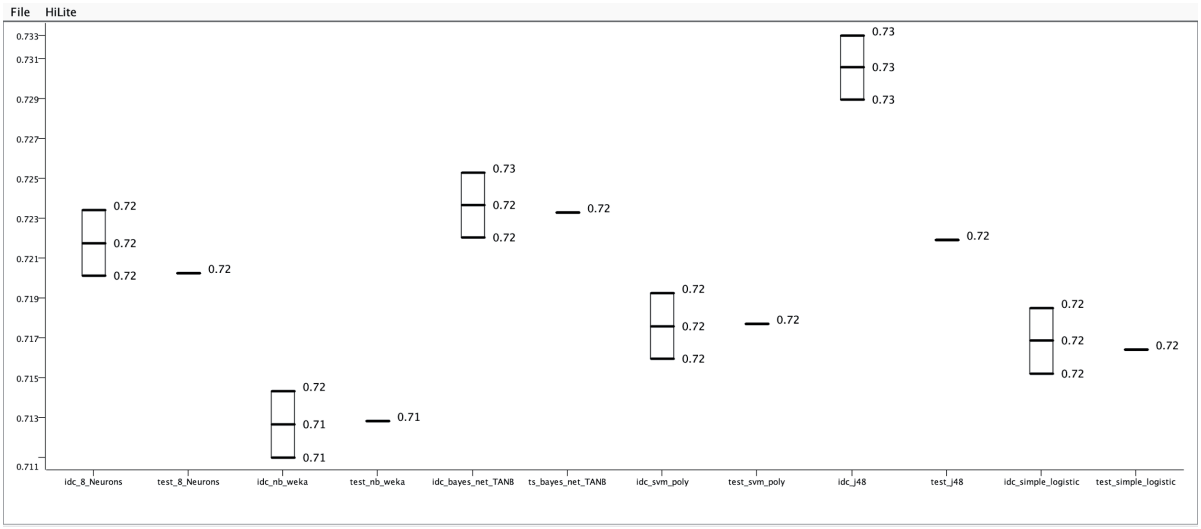

FIGURA 7: Intervalli di confidenza

Test d'ipotesi e Intervalli di confidenza

In questa fase del lavoro, i risultati ottenuti dai modelli in termini di accuracy non possono ancora essere considerati statisticamente significativi. Per questo motivo, lo step qui descritto consiste nel calcolare gli intervalli di confidenza su tali valori, per testare se le accuracy possano essere "accettate".

In particolare, si è deciso di sfruttare un livello di confidenza pari al 95%.

Il boxplot riportato (FIGURA 7) dimostra che per tutti i modelli tranne uno è possibile accettare e quindi considerare come "vero" il valore dell'accuracy.

Infatti, l'unico modello la cui Accuracy non ricade nell'intervallo di confidenza è il J48, che per questo verrà escluso dalle prossime fasi di analisi.

Analisi finale con Misure di Performance

A questo punto, dei 5 modelli rimanenti sono state calcolate altre misure di performance con l'obiettivo di selezionarne uno solo.

Data la natura del carattere che si è cercato di predire (insorgenza di malattie cardiovascolari) si è ritenuto opportuno dare maggiore rilievo alla capacità del modello di identificare correttamente i "positivi" e, parallelamente, si sono considerati fortemente indesiderabili i "falsi negativi".

Per questo motivo, si è data più importanza alle metriche di performance di Recall e Precision.

La **Recall** dà informazione della porzione di valori positivi correttamente predetti.

Più il valore di Recall è alto, minori saranno i falsi negativi.

$$\text{Recall} = \frac{TP}{TP + FN}$$

La **Precision** invece descrive la frazione di record effettivamente positivi, tra quelli classificati come tali dal modello.

In questo caso, più il valore di Precision è alto, minore è il numero di falsi positivi.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Si è ritenuto che fosse più rilevante la Recall rispetto alla Precision, in quanto l'errore più grave che si vuole evitare in questa anali-

si è quello di predire erroneamente la negatività alla malattia cardiovascolare. Come già spiegato, questa informazione è fornita proprio dalla Recall.

Meno importanti ai fini di questo progetto, ma comunque meritevoli di essere citate, sono altre misure di performance come la Specifity e la F1-measure.

La prima indica la proporzione di record davvero negativi che sono identificati come tali dal modello (i "veri negativi").

La F1-measure invece calcola la media armonica tra Recall e Precision. Un valore alto sta ad indicare che sia Recall che Precision sono ragionevolmente alte

Nella tabella vengono mostrati i valori delle performance per ogni modello portato avanti in questa fase di analisi:

CLASSIFICATORE	RECALL	PRECISION	SPECIFICITY	F-MEASURE
Bayes Net TANB	0.6563	0.7358	0.7844	0.6938
MLP 8 Neurons	0.6233	0.7443	0.8041	0.6784
Naive Bayes Weka	0.5957	0.7501	0.8184	0.6641
Simple Logistic	0.6346	0.7383	0.7942	0.6825
SVM Poly	0.5950	0.7594	0.8274	0.6672

TABELLA 1: Altre misure di performance

Come si può notare i valori più elevati di Recall sono imputabili a Bayes Net TANB e Simple Logistic, i quali però presentano anche i valori minori di Precision.

Tuttavia, per le ragioni appena evidenziate, si ritengono "più promettenti" questi due modelli rispetto agli altri.

Nella fase finale di valutazione, sono state calcolate e visualizzate le **ROC curve** (Receiver Operating Characteristic) dei metodi rimanenti.

La ROC curve riporta il True Positive Rate (frazione di veri positivi) sull'asse delle ordinate, espresso come percentuale del numero totale di record positivi; e il False Positive Rate (frazione di falsi negativi) sull'asse delle ascisse, espresso come percentuale dei record negativi.

Dunque, la curva mostra come varia la frazione di veri positivi al variare di quella di falsi positivi.

Da questo indicatore viene ricavato anche il valore dell'**AUC** (Area Under the Curve), ovvero l'area sottesa alla curva.

Vi sono due casi limite:

1. Quando la ROC corrisponde ad una retta di 45° che taglia il grafico all'origine e l'AUC è pari a 0,5, ci si ritrova nel caso di classificatore casuale in cui non si ha nessun "beneficio".

2. Quando la ROC corrisponde ad un segmento che dall'origine arriva al punto (0,1) e da quello si congiunge al punto (1,1), con AUC pari a 1, si è nel caso di classificatore perfetto.

In generale, l'AUC descrive il grado di efficacia con cui il modello assegna i valori alla corretta classe di appartenenza e assume valori da 0 a 1.

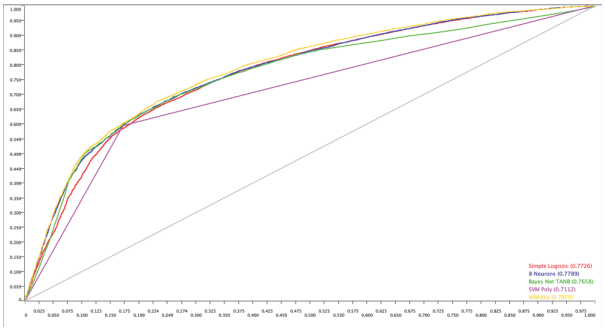

FIGURA 8: ROC curve

Osservando i risultati ottenuti nel grafico, si può affermare che nel complesso gli stimatori forniscono risultati abbastanza omogenei, ad eccezione del modello SVM. Si nota infatti che le ROC degli altri modelli tendono a sovrapporsi.

CLASSIFICATORE	AUC
Bayes Net TANB	0.766
MLP 8 Neurons	0.779
Naive Bayes Weka	0.787
Simple Logistic	0.773
SVM Poly	0.711

Riguardo all'AUC, i cui valori sono riportati in tabella, è possibile invece riscontrare che i risultati migliori corrispondono ai modelli che avevano mostrato un livello di Specificity più alto (infatti FPR= 1- Specificity).

Come già detto, in funzione dei criteri finora utilizzati, i risultati migliori fino a questo punto sono stati restituiti dai modelli Simple Logistic e Bayes Net TANB. Dunque per poter operare un'ulteriore selezione tra questi, si è deciso di guardare ai valori ottenuti dall'AUC.

Così facendo, il modello che risulta essere più adeguato e preciso per l'analisi attuata è il Simple Logistic.

CONCLUSIONI

In conclusione si può affermare che i risultati ottenuti sono stati soddisfacenti perché buona parte dei modelli ha performato bene, rendendo così il processo di selezione del migliore una questione di "decimali".

Grazie a quest'analisi si è potuto verificare non solo che è possibile utilizzare le tecnologie di Machine Learning per la predizione di malattie cardiovascolari, ma anche che queste restituiscono risultati sufficientemente accurati da essere effettivamente utilizzabili come strumento di integrazione in una completa diagnosi medica.

Si è inoltre riscontrato che molti modelli performano in maniera adeguata, rendendo quindi il processo di selezione del migliore non tanto una constatazione immediata quanto un controllo dei dettagli più rilevanti in relazione all'oggetto di analisi.

Volendo approfondire e migliorare ulteriormente l'analisi, si sarebbe potuto:

- Integrare o arricchire i dati con un numero maggiore di feature per allineare le variabili a disposizione del modello predittivo a quelli della "carta del rischio".
- Approfondire il tema dell'analisi dei costi in funzione della gravità della malattia cardiovascolare.

RIFERIMENTI

https://www.kaggle.com/sulianova/cardiovascular-disease-dataset
https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings
http://www.cuore.iss.it/valutazione/carte