UNIVERSITÀ
DI PARMA

Dipartimento di Ingegneria e Architettura

Corso di Laurea in Ingegneria informatica, elettronica e delle telecomunicazioni

---

Data augmentation per la predizione di interventi chirurgici basata su parametri anamnestici

Data augmentation for predicting surgical interventions based on anamnestic parameters

Relatore:
Chiar.mo Prof. Michele Tomaiuolo

Tesi di Laurea di:
Giovanni Annaloro

Correlatori:
Dott. Ing. Mattia Pellegrino

ANNO ACCADEMICO 2023/2024

Grazie ai recenti progressi nel campo del *Machine Learning*, l'intelligenza artificiale sta trovando sempre più applicazioni nel settore della scienza e della tecnologia medica. Uno degli utilizzi più promettenti riguarda il supporto alle decisioni nel processo di diagnosi, scelta del trattamento e pianificazione delle cure. Tuttavia, l'implementazione di modelli di Machine Learning prestazionali richiede l'accesso a una grande quantità di dati, il che può essere problematico, specialmente nel contesto medico, dove questi sono meno facilmente disponibili per via della severa normativa sulla privacy e della bassa propensione del pubblico a condividerli.
Una possibile strategia per affrontare la carenza di dataset adeguatamente ampi e variegati consiste nel generare nuovi dati sintetici a partire da quelli di cui si è già in possesso, così da poterli utilizzare per aumentare ampiezza e varietà del dataset originario. Questa tecnica, che prende il nome di *data augmentation*, permette di aumentare sensibilmente le prestazioni e le capacità di generalizzazione dei modelli di Machine Learning.
L'obiettivo del lavoro di tesi è aumentare le prestazioni di un predittore di interventi chirurgici usando tecniche di data augmentation allo stato dell'arte. Il predittore in questione consiste in un classificatore capace di predire se un paziente debba essere sottoposto a uno tra i seguenti interventi: *interventi sull'apparato digerente*, *interventi sul sistema endocrino* o *interventi sul sistema cardiovascolare*.
La classificazione avviene sulla base dei suoi parametri anamnestici, questi sono contenuti in un dataset tabulare fornito dall'Azienda ospedaliero-universitaria di Parma. La classificazione è stata effettuata utilizzando tre diversi algoritmi di Machine Learning: *AdaBoost*, *RandomForest* ed una *rete neurale*. Per ogni algoritmo sono stati selezionati gli iperparametri migliori utilizzando l'algoritmo di Hyperparameter-tuning *Gridsearch*. Una volta selezionati gli iperparametri migliori, si è proceduto alla selezione del migliore modello per ogni algoritmo. In particolare, per ognuno degli algoritmi sono stati addestrati 50 modelli, ognuno dei quali è stato poi testato sul validation set. Infine, per ogni algoritmo è stato selezionato il modello che massimizzava la *Macro F1Score*, ossia la media aritmetica della F1Score di ogni classe. I migliori modelli ottenuti sono poi stati valutati sul test set.
Successivamente sono stati generati tre dataset aumentati. Per la generazione dei dati sintetici è stato utilizzato l'algoritmo di sovracampionamento *Smotenc (Synthetic Minority Over-sampling Technique for Nominal and Continuous.)* e due architetture di deep learning generative: *Tvae (Triplet-Based Variational Autoencoder)* e *Nflow (Neural spline flow)* , di cui sono state utilizzate le implementazioni presenti nella libreria *Synthcity*. Per cercare di generare il miglior dataset aumentato possibile è stata eseguita una procedura di selezione. Per quanto riguarda la generazione tramite Smotenc, sono stati creati 10 dataset differenti, per ogni dataset sono stati allenati 10 modelli di AdaBoost, infine si è selezionato il dataset per cui la media delle F1Score macro calcolata sul validation set dei 10 modelli è risultata maggiore. Per quanto riguarda la generazione con Tvae e Nflow, essendo questi dei modelli generativi che necessitano di essere addestrati, si è proceduto prima ad addestrare due differenti modelli per ogni algoritmo utilizzando il miglior dataset aumentato con Smotenc. Successivamente per ogni modello sono stati

generati 5 dataset differenti, per ognuno di essi si sono addestrati 10 modelli di AdaBoost e si è poi selezionato il dataset per cui la media delle F1Score macro calcolate sul validation set è risultata maggiore. Una volta ottenuti i migliori dataset, questi sono stati utilizzati per addestrare i modelli dei 3 algoritmi di classificazione utilizzati. Per ogni algoritmo è stato poi selezionato il miglior modello attraverso lo stesso processo di selezione utilizzato nel caso del dataset non aumentato. I migliori modelli addestrati sui dataset aumentati sono stati infine confrontati con i migliori modelli addestrati sul dataset non aumentato.
Il miglior modello addestrato sul dataset non aumentato è risultato essere un modello di AdaBoost con una F1Score macro calcolata sul test set pari a 0.45. Il modello migliore addestrato sui dataset aumentati è invece risultato essere un modello di AdaBoost, addestrato sul miglior dataset ottenuto con Smotenc e avente una F1Score macro calcolata sul test set pari a 0.65. Si è quindi riusciti ad ottenere un aumento del 44% della F1Score Macro, l'aumento di prestazioni risulta tra l'altro evidente confrontando le matrici di confusione dei due modelli migliori:

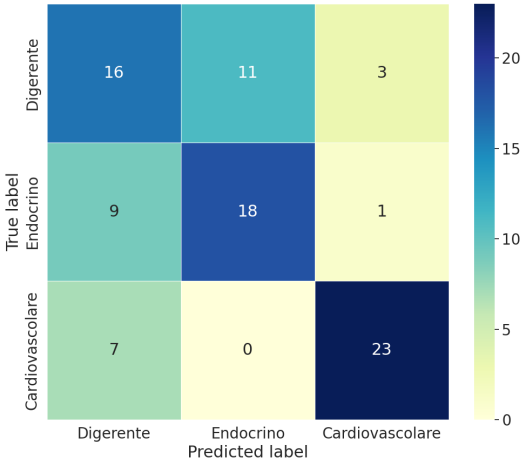

Figura 1: Matrice di confusione ottenuta addestrando AdaBoost sul miglior dataset aumentato con Smotenc

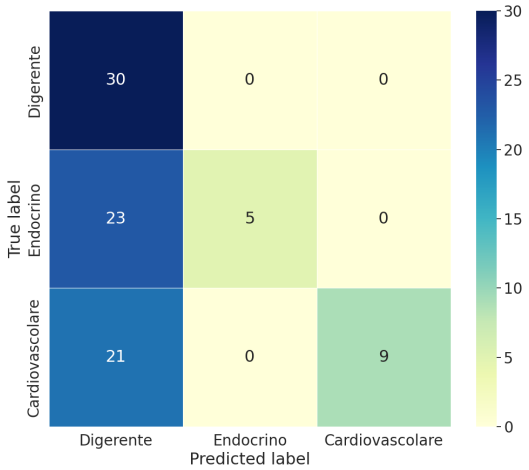

Figura 2: Matrice di confusione ottenuta addestrando AdaBoost sul dataset originale

Questo risultato suggerisce che l'uso della data augmentation possa migliorare significativamente l'accuratezza e la robustezza di un modello, dimostrando il potenziale di questa tecnica per superare le limitazioni imposte dalla scarsità di dati disponibili nel settore medico.
Un risultato notevole ottenuto dagli esperimenti effettuati è che la tecnica che ha permesso di ottenere l'aumento di prestazioni più significativo, ossia Smotenc, è basata sul sovracampionamento. Questo fatto, che trova riscontro in letteratura almeno per quanto riguarda la classificazione binaria per dataset tabulari, suggerisce la possibilità di proseguire il lavoro di tesi fin qui svolto andando a esplorare ulteriori approcci generativi per la data augmentation e verificando se attraverso questi sia possibile ottenere risultati migliori rispetto a quelli ottenibili con algoritmi di sovracampionamento come Smotenc.