ALMA MATER STUDIORUM – UNIVERSITÀ DI BOLOGNA
CAMPUS DI CESENA

Dipartimento di Informatica - Scienza e Ingegneria
Corso di Laurea in Ingegneria e Scienze Informatiche

APPLICAZIONE DI TECNICHE DI INTELLIGENZA ARTIFICIALE PER L'ANALISI DI DATI ACQUISITI DURANTE IL TRAVAGLIO DI PARTO

Elaborato in
ARCHITETTURA DEGLI ELABORATORI

Relatore
Prof. DAVIDE MALTONI

Co-relatore
Dott.ssa SARA MONTAGNA

Presentata da
GIADA BOCCALI

Prima Sessione di Laurea
Anno Accademico 2018 – 2019

PAROLE CHIAVE

Intelligenza Artificiale

Machine Learning

Progetto DIANA

Parto cesareo

Analgesia

A Christian,
a cui potrei dedicare libri interi.

Indice

Introduzione

Tra le innovazioni tecnologiche più emergenti degli ultimi anni spicca sicuramente l'Intelligenza Artificiale. Con le sue molteplici sfaccettature, questa disciplina sta entrando a far parte, più o meno evidentemente, della quotidianità di sempre più persone.
L'attenzione è rivolta soprattutto ad uno dei rami che sta trovando più applicazione: il Machine Learning. La ragione è legata alla versatilità e alla fruibilità di questo approccio, il quale può essere utilizzato per risolvere le problematiche più disparate. Col passare del tempo sempre più settori vi si stanno interessando e stanno cercando di capire come poter trarre benefici da questi strumenti. Tra questi, il settore sanitario sta investendo sempre più risorse per quanto riguarda la ricerca e lo sviluppo con l'utilizzo di strumenti informatici avanzati.

In questa tesi si descriveranno i concetti più importanti dei principali algoritmi e tecnologie del ML. Verrà inoltre fornita una panoramica dell'impatto che queste tecnologie stanno avendo, ad oggi, nella ricerca biomedicale e nella pratica clinica.
Sarà illustrato nel dettaglio l'approccio da seguire quando ci si interfaccia con un problema di Machine Learning, descrivendo i passaggi necessari per rendere completa una ricerca in questo ambito. Si spiegheranno inoltre i principali problemi che è possibile risolvere e verranno introdotti gli algoritmi principali più in uso al giorno d'oggi.

Verrà poi preso in esame un problema di studio reale, a cui verranno applicati concretamente le tecnologie e in particolare gli algoritmi di ML. Lo studio fa parte del progetto DIANA, nato dalla collaborazione tra l'Ospedale M. Bufalini AUSL Romagna e il dipartimento di Ingegneria e Scienze Informatiche di Cesena.
Il progetto è stato ideato per analizzare, attraverso algoritmi di ML, i dati relativi ai parti in cui si è fatto uso di analgesia spino-peridurale, al fine di identificare i fattori che maggiormente influiscono sulla necessità di ricorrere a taglio cesareo urgente.

I parti cesarei infatti costituiscono una percentuale molto elevata in Italia sui parti totali, numero che si sta cercando di ridurre cercando di capire le cause che influenzino su questi numeri. Secondo alcune statistiche ci sarebbe un'incidenza maggiore di questo tipo di parto quando si fa ricorso all'analgesia come terapia per ridurre il dolore del parto. La domanda che è sorta da questo dato è se sia l'analgesia in sé ad aumentare il rischio o se invece sono altri i fattori che, combinati con l'analgesia o relativi a metodologie utilizzate in essa, portano con più facilità ad un parto cesareo non programmato.
L'obiettivo sarà quello di sfruttare gli strumenti messi a disposizione dal Machine Learning per trovare eventuali correlazioni tra le caratteristiche dei parti e il loro esito e capire quali cause legate all'analgesia portino un aumento dei tagli cesarei.

Capitolo 1

AI e applicazioni nel biomedicale

Negli ultimi anni si sente parlare sempre più frequentemente di Intelligenza Artificiale. Il trasporto è in continuo aumento soprattutto per quanto riguarda il Machine Learning, driving force di questo settore, grazie alle innovazioni e al progresso tecnologico che sta già portando nella vita di tutti.

Tra i settori che maggiormente stanno venendo coinvolti da quest'ultima disciplina c'è il mondo della sanità, il quale sta partecipando a questa espansione tecnologica in modo attivo, sotto forma di incentivi, investimenti e dando vita a numerosissimi progetti. Il Machine Learning infatti si sta rivelando capace di trovare miglioramenti a problematiche da sempre collegate alle dinamiche ospedaliere e sanitarie, permettendo miglioramenti con soluzioni che prima risultavano troppo complesse da trovare.

Ma per capire come queste tecniche siano in grado di fare ciò e come stiano impattando il mondo e in particolare il settore medico, è necessario capire cosa siano e come funzionino.
In questo lavoro di tesi parleremo di come queste nuove tecnologie stiamo entrando nella vita di tutti, analizzando i benefici e gli svantaggi che queste portano, concentrandoci in particolare sul settore medico.

1.1 Intelligenza Artificiale

L'intelligenza artificiale, o AI, è una branca dell'informatica che ha l'obiettivo di imitare i modi di ragionare, apprendere e scegliere tipici dell'essere umano. L'idea è quindi quella di apprendere da situazioni conosciute per generalizzarne le regole.

Si è cominciato a parlare di questa disciplina già a metà degli anni 50 del ventesimo secolo. Nel 1955 nasceva infatti Logic Theorist, il primo programma che, grazie ad alcune operazioni logiche, era in grado di dimostrare teoremi matematici astratti.
Tuttavia negli anni successivi la progressione di questa disciplina è stata lenta, si sono presentati numerosi problemi come la difficoltà di riprodurre il ragionamento umano, la crescente complessità dei problemi da risolvere, la scarsa disponibilità di dati e di potenza di calcolo degli elaboratori.
È soltanto nell'ultimo ventennio che l'AI ha iniziato ad affermarsi. Il merito si deve alla potenza crescente dei processori, ai dispositivi di memorizzazione meno costosi e più capienti e ai progressi nel settore della robotica che hanno fatto tornare ampiamente l'interesse per questa disciplina.
La svolta si è vista nel 1996 con Deep Blue, il super computer dell'IBM che fece scalpore quando vinse a scacchi contro il campione del mondo Garry Kasparov. Per la prima volta nella storia una macchina era riuscita ad imparare la logica di un gioco complesso come gli scacchi e a concludere una partita senza ulteriori interventi umani durante lo svolgimento di essa, conseguendo inoltre una vittoria.

Al giorno d'oggi l'Intelligenza Artificiale è un campo molto vasto. I settori di studio in cui viene utilizzata ricoprono molte tipologie di problematiche. Le principali sono:

- la **rappresentazione della conoscenza e del ragionamento** che punta a riprodurre il modo di ragionare del cervello umano tramite la definizione di simbolismi e linguaggi al fine di creare macchine in grado di eseguire ragionamenti automatici

- la **pianificazione** e il **coordinamento** che trattano lo sviluppo di sistemi che, dato un dominio applicativo, hanno l'obiettivo di prevedere risultati futuri e prendere decisioni per raggiungere tali obiettivi e massimizzarne i benefici

- la **robotica**, in particolare per gli studi legati al movimento di parti meccaniche per la creazione di una sequenza di azioni che svolgano un determinato compito o che sappiano reagire ad un determinato fenomeno, alla manipolazione o spostamento di oggetti, alla localizzazione o alla costruzione delle mappe

- il **Natural Language Processing** (processamento del linguaggio naturale) che riguarda l'interpretazione automatica del linguaggio naturale, sia in forma scritta che parlata

- la **visione artificiale**, l'insieme dei processi per l'elaborazione e l'analisi di immagini, la pattern recognition e l'analisi automatica della scena, utilizzando tecnologie 2D e 3D
- il **Machine Learning** raggruppa tutta quella serie di studi che puntano a creare programmi in grado di imparare da soli a migliorare le proprie prestazioni sulla base dell'esperienza accumulata

Negli ultimi anni gli investimenti sono cresciuti in maniera esponenziale. Nel 2018 l'UE si è mossa in questo senso, pianificando come obiettivo il raggiungimento di 20 miliardi di euro di investimenti in ambito AI entro il 2020, e di 20 miliardi all'anno per gli anni successivi [1].

1.2 Machine Learning

L' **apprendimento automatico**, traduzione impropria di Machine Learning (ML), è la branca dell'AI che negli ultimi anni sta riscuotendo il maggior successo, infatti la maggior parte dei finanziamenti diretti all'Intelligenza Artificiale (stimati tra i 20 e i 30 miliardi di dollari solo per le aziende del settore tecnologico nel 2016, di cui il 60% destimati al ML [2]) sono stati diretti a questi studi.

La forza di questa disciplina riguarda la capacità di sfruttare delle risorse reperibile ovunque ma da cui non si era in grado di trarre ricchezza. Queste risorse sono i dati, e in particolare i Big Data.
Big Data è il termine utilizzato per indicare una raccolta di dati enorme che contiene al suo interno un valore informativo. Il sempre crescente numero di dispositivi che raccolgono informazioni sulle persone e sull'ambiente permettono di avere dati di ogni genere e costruire banche dati che rendono possibile l'applicazione di queste tecniche dove prima non si poteva a causa della scarsità di informazioni a disposizione.

Alla base del Machine Learning vi è il concetto di applicare regole matematico-computazionali per apprendere direttamente dai dati. I mondi matematici, statistici e probabilistici vengono uniti a quello informatico con lo scopo di trovare collegamenti, spesso impercettibili all'occhio umano, attraverso lo studio di un'enorme mole di dati di uno stesso dominio. L'idea è di poter studiare un campione di dati per comprendere quali caratteristiche legano le variabili indipendenti a quelle dipendenti del problema e in quale misura.

In ogni problema di Machine Learning viene costruita una funzione in cui le caratteristiche (*feature*) dei dati che si conoscono e che non dipendono dalla

modellazione del problema, ossia le variabili indipendenti fornite in input, vengono messe in relazione alla variabile da trovare, la quale dipende dalle altre ed è attesa in output.

A seconda delle variabili che si conoscono, i modelli possono essere divisi in:

1. apprendimento supervisionato
2. apprendimento non supervisionato
3. apprendimento con rinforzo

Nel primo caso oltre ad avere a disposizione gli input con le caratteristiche descrittive dei dati, sono presenti anche gli output, detti classi. Questo significa che si conosce l'output previsto per ogni istanza dei dati che abbiamo a disposizione. Si dice allora che i dati sono etichettati rispetto alle classi di output.

Nel secondo caso invece il modello verrà costruito per mettere in relazione le *feature* e raggrupparle in classi di cui non si conosce a priori la natura. Questo secondo tipo di modelli è solitamente più complicato da gestire.

L'apprendimento con rinforzo invece sfrutta un meccanismo di ricompense, in cui si lascia libero l'algoritmo di effettuare azioni non supervisionate e lo si premia ogni volta che raggiunge il risultato desiderato o lo si ammonisce in caso contrario. L'obiettivo per l'algoritmo è quindi quello di trovare le sequenze giuste d azioni che permettono di massimizzare il premio.

Gli algoritmi vengono suddivisi ulteriormente in base al tipo di problema da risolvere, il quale è influenzato anche dalla conoscenza o meno delle variabili di output. Questa divisione comprende i seguenti tipi di algoritmi:

- **Predizione**, in cui l'obiettivo è di generalizzare le regole dei dati di cui si conosce l'output atteso per prevedere l'output di nuove istanze di cui invece non si conoscono i risultati. Questo problema si divide a sua volta in:
  - **Classificazione** con la quale il modello assegna ai nuovi input la classe più probabile tra quelle dei campioni dai quali si sono generalizzate le regole.
  - **Regressione** che si occupa dei problemi in cui gli output da prevedere sono valori continui.
- **Riduzione delle dimensionalità** che punta a ridurre la dimensione del problema individuando le *feature* principali che meglio lo descrivano e che causino meno perdita di informazione.

- **Clustering** nel quale si vogliono trovare raggruppamenti tra i dati senza conoscere a priori le classi attese. Tipicamente questo tipo di problema non è supervisionato.

- **Representation Learning** che riguarda l'estrazione automatica di *feature* in problemi in cui i dati non hanno una struttura evidente. L'Obiettivo è la rappresentazione dei "raw data", i dati grezzi che non hanno un formato preciso di archiviazione, in un formato utili all'analisi. Questo tipo di modelli non verrà trattato in questa tesi.

1.3 Reti Neurali e Deep Learning

Le Reti Neurali e il Deep Learning sono diramazioni del Machine Learning, utilizzate principalmente per risolvere problemi più complessi di quanto fanno gli algoritmi di Machine Learning tradizionali.

Queste tecnologie sono state create prendendo ispirazione dai neuroni degli esseri umani e cercano infatti di riprodurne il funzionamento.
Sono sistemi che sono in grado di modificare la propria struttura basandosi sulle informazioni in loro possesso e su quelle ricavate internamente.
Alla base c'è il concetto di neurone artificiale, il quale può ricevere diversi input e fornire un solo output.

La Rete Neurale, traduzione italiana del termine inglese Neural Network (NN), è formata da gruppi di questi neuroni artificiali ed essere organizzata su più livelli. Il primo è il livello degli input, a cui segue un livello nascosto (hidden layer) e uno di output. Ad ogni livello corrispondono uno o più neuroni e comunicano tra loro, infatti l'output di un livello costituisce l'input di quello dopo.

Quando i livelli nascosti nella NN sono più di uno e sono organizzati gerarchicamente, si parla di Deep Learning (o Deep Neural Network).
La peculiarità dell'organizzazione gerarchica consiste nel rendere condivisibili e riusabili le informazioni o selezionare specifiche *feature*.

L'obiettivo di questi algoritmi è di mappare nella maniera più precisa gli input con gli output attesi.

1.4 L'impatto nel settore medico

Il settore medico americano è stato il primo ad interessarsi all'AI, quando negli anni settanta sono cominciati gli studi sull'applicazione di questa nella biomedica. Il coinvolgimento di sempre più studiosi, universitari e non, anche al di fuori dei confini americani, ha fatto nascere il desiderio di interazione tra le varie ricerche fatte a livello mondiale. Per rispondere a questa esigenza, è stato organizzato nel 1985 il primo meeting internazionale sull'AI presso l'Università di Pavia. L'evento ho riscontrato molto successo, il ché ha permesso la nascita della Società per l'Intelligenza Artificiale per la Medicina (Artificial Intelligence In Medicine o, abbreviato, AIME), nell'anno successivo. L'AIME organizza da allora conferenze mondiali ogni due anni per rendere noti, a tutta la comunità medica, gli studi derivati dall'applicazione dell'AI in ambito sanitario, al fine di facilitare lo scambio di idee sulle possibili innovazioni e applicazioni, per far conoscere pubblicazioni mediche innovative e per dare luce a nuovi progetti.

1.4.1 Applicazioni e Benefici

Sebbene questo tipo di studi sia ancora giovane e abbia espresso per ora solo in minima parte il suo potenziale, esso ha già portato numerosi benefici nel settore della sanità.
L'Intelligenza Artificiale sta venendo adottata per una vasta gamma di problematiche.

Le applicazioni e gli studi, per quanto riguarda il Machine Learning e il Deep Learning, vanno dalle scansioni mediche fino ad arrivare alle diapositive patologiche, dalle lesioni cutanee alle immagine immagini retiniche, poi ancora al funzionamento e ai disturbi della mente, al monitoraggio dei segni vitali, alla lettura degli elettro-cardiogrammi, alla sicurezza degli ospedali, alle interazioni tra i medici.

In quasi tutte le ricerche effettuate i risultati hanno evidenziato che le prestazioni delle macchine eguagliavano o addirittura superavano le capacità diagnostiche dei medici [3]. Tuttavia i numerosi software sono scissi tra di loro e ognuno è specializzato su determinati argomenti, per cui risulta ancora impossibile effettuare diagnosi partendo dalla cartella clinica generica del paziente, senza soffermarsi sul problema specifico.

Tuttavia un'applicazione che aiuta i medici davanti alla generalità dell'argomento è l'insieme delle tecniche del Clinical Decision Support System, un sistema composto da software e applicativi di varia natura che aiutano i dottori a consultare la letteratura medica a disposizione in maniera più veloce, accurata e inerente alle tematiche di interesse, oltre a rimanere costantemente

aggiornati sulle innovazioni.

Le prospettive future indicano che col passare del tempo e con la progressione di strumenti informatizzati, il sistema sanitario inizierà ad appoggiarsi sempre più frequentemente alla tecnologia, limitando sempre di più la necessità dell'intervento umano, al fine di migliorare i risultati delle prestazioni sanitarie e delle diagnosi cliniche.

Alcuni esempi Per quanto riguarda le applicazioni legate al funzionamento della mente, è stato sviluppato è Keepon, un robot interattivo che riesce ad aiutare i bambini autistici a comunicare, interagire e giocare in maniera più naturale rispetto a come avviene con le persone, con cui questi bambini faticano ad approcciarsi [4].

L'interesse per la ricerca è così alto che è stata creata un'intera area di Google, Google Health, che si occupa di sanità. Uno dei gruppi di ricerca ha sviluppato negli ultimi anni un algoritmo, chiamato Lymph Node Assistant (LYNA), che è in grado di diagnosticare correttamente i tipi di cancro al seno con una precisione del 99% [5]. Questo strumento oltre a facilitare la decisione della terapia corretta da prescrivere al paziente, permette anche di trovare con largo anticipo metastasi ancora troppo piccole per essere riconoscibili dall'occhio umano.

Molto interessante è anche Da Vinci, realizzato dell'azienda americana Intuitive Surgical [6]. Da Vinci è un robot chirurgico che aiuta a diminuire l'invasività degli interventi. Non è completamente autonomo ma grazie alla ridotta dimensione e dell'estrema precisione dei sui bracci meccanici diventa un'estensione del corpo del chirurgo, che è così in grado di effettuare movimenti più puliti. Questa macchina permette di intervenire chirurgicamente riducendo il numero e la dimensione delle cicatrici, permettendo di operare passando da piccoli tagli anche per operare zone del corpo più interne, di cui normalmente non si avrebbe molta visibilità e per cui sarebbe necessario effettuare interventi più invasivi.

I vantaggi che si stanno riscontrando non riguardano solo le diagnosi e gli interventi, bensì anche ciò che fa da contorno alla sanità, come l'organizzazione interna del personale e dei pazienti, la programmazione delle visite, i costi del personale e delle cure mediche.

Proprio su questo lavora l'azienda LeanTaaS Inc. che ha creato un software in grado di ottimizzare l'organizzazione all'interno degli ospedali, per migliorare l'utilizzo delle sale operatorie e dei letti ospedalieri e ridurre i tempi di attesa dei pazienti per fare visite mediche. Grazie al Machine Learning questi software sono in grado di studiare i dati storici degli interventi, degli appun-

tamenti e dell'utilizzo delle sale per trovare l'allocazione ideale delle risorse a disposizione degli ospedali e portare benefici alla struttura ospedaliera.
In termini di costi, nel 2017 l'UCHealth University of Colorado Hospital (UCH) annunciava un aumento delle entrate di 10 milioni di dollari con l'introduzione di questo software, oltre ad un aumento della produttività dei medici a parità di ore lavorative [7].

Si stima che utilizzando strumenti di questo genere le ripercussione sull'economia per i paesi con più alto reddito si avrebbero risparmi compresi tra lo 0,5 e l'1% del PIL, aumentando la produttività degli infermieri del 40-50% e diminuendo drasticamente i tempi di attesa dei pazienti e i costi della sanità [2].

1.4.2 Problemi

A causa dell'ancora recente sviluppo di questa tecnologia non si conoscono tutti gli effetti che potrà avere all'interno della società. Se da una parte c'è molta curiosità verso di essa e si inizia ad utilizzarla in sempre più aspetti della vita, come ad esempio installando un assistente vocale o utilizzando applicazioni sugli smartphone che sfruttano il ML per migliorare l'user experience, dall'altra c'è scetticismo, causato dalla paura di una nuova tecnologia complessa da comprendere ai non esperti e alimentato dalla presenza di film e serie TV apocalittici in cui le macchine prendono il controllo del pianeta e dell'uomo.

Sostituire l'uomo C'è la paura che l'automatizzazione di molti processi possa rendere l'uomo superfluo e sostituirle l'uomo.

Queste paure possono essere comprensibili, infatti l'intensificazione della tecnologia ha sicuramente già avuto un impatto su molti tipi di lavori e in molti sta per farlo. Si pensi ad esempio alle auto a guida autonoma che, se rendessero piede, potrebbero sostituire completamente i taxisti.

Tuttavia anche questo aspetto ha una duplica faccia della medaglia, infatti bisogna anche considerare che l'automatizzazione di processi permette anche all'uomo di non avere l'onere di effettuare quei lavori più faticosi, ripetitivi e pericolosi.

Nel settore medico non si è in una fase in cui la problematica della sostituzione dei robot ai medici è già presente e non lo sarà ancora per molto tempo.

La differenza sostanziale tra questo settore e altri è legato ala tipologia di lavoro che si svolge. Le scelte mediche non sono lavori ripetitivi in cui le decisioni vengono basate sui soli dati ma coinvolgono molteplici aspetti, come

le esigenze del paziente, le questioni etiche, morali, le condizioni economiche, i fattori emotivi.

Ci si chiede quindi fino a che punto questi aspetti interpersonali potranno essere recepiti da un algoritmo o da un robot e se sarà possibile immaginare un futuro in cui i dottori saranno sostituiti dai computer. Per il momento questo tipo di tecnologie devono essere viste solamente come mezzo di sostegno alle decisioni e non come sostituti.

Etica Parlare di macchine che sostituiscono gli esseri umani fa emergere una questione etica: cosa succederebbe in caso di un errore della macchine o dell'algoritmo? Come verrebbe valutato un errore di questo genere e a chi verrebbe attribuita la colpa? Sebbene la probabilità che questi sbaglino sia di molto inferiore rispetto all'essere umano, in casi rari o mai visti prima saprebbero come comportarsi e come reagire?

Sempre parlando di automobili a guida autonoma e in particolare di quelle progettate dall'azienda Uber, è famoso l'incidente in Arizona nel quale, proprio a causa di una di queste auto, ha perso la vita una signora. Gli algoritmo di riconoscimento delle immagini non hanno rilevato la signora che stava attraversando la strada spingendo la sua bicicletta e l'automobile di conseguenza non ha frenato. Al posto di guida era presente un impiegato di Uber che doveva intervenire in caso di possibili errori ma che in quel momento era distratto [9].

In questo caso lo stato della contea di Yavapai, Sheila Polk, non ha ritenuto perseguibile penalmente l'azienda per l'incidente perché a bordo era presente un autista umano. Tuttavia reagendo a questa notizia ci si chiede a chi debba essere imputata la colpa di incidenti in cui sono coinvolti sistemi autonomi, soprattutto quando questi non saranno più assistiti da personale umano.

Per le decisioni mediche queste domande diventano ancora più difficili da rispondere, questo perché non sempre gli effetti di una diagnosi sono immediati, quindi imputabili alla specifica diagnosi o non si può avere la certezza di cosa sarebbe potuto succedere con una diagnosi diversa.

Per tutti questi fattori e perché, come ci insegna la storia, la maggior parte delle nuove tecnologie possono essere usate con intenti sia positivi che negativi, al giorno d'oggi non c'è ancora una completa fiducia da parte delle persone verso queste tecnologie e le questioni etiche e morali rimangono aperte [2].

Capitolo 2

Principi e Algoritmi del Machine Learning

Le tecniche di Machine Learning sono strumenti molto utili per studiare insiemi di dati di grandi dimensioni e apparentemente distribuiti in modo casuale.
Il merito è dovuto alla capacità degli elaboratori di applicare concetti matematici complessi su moli di dati molto più grandi e su molte più variabili rispetto a come fanno le tecniche tradizionali, il che spesso permette di far emergere le correlazioni meno evidenti che esistono tra i vari record.

Quando si manifesta la volontà di utilizzare il ML va innanzitutto conosciuta la natura del problema preso in esame e compresa la natura dei dati sui quali si interverrà. Questo passaggio va fatto indispensabilmente prima di applicare qualunque tipo di algoritmo, altrimenti qualsiasi tipo di interpretazione risulterebbe impossibile o potrebbe portare a risultati non corretti. Vanno inoltre conosciuti i principi alla base degli algoritmi che si applicano, in modo da poter valutare quale si presti meglio agli interrogativi presi in esame.

Formalizzando quanto detto, si può dire che i passaggi necessari da seguire per poter rendere significativo uno studio di ML sono: modellazione del problema, pre-processamento e trasformazione dei dati, scelta e applicazione degli algoritmi, valutazione dei risultai.

2.1 Analisi del problema

Un aspetto cruciale da non sottovalutare è l'analisi iniziale, durante la quale si modella il problema e si individuano gli aspetti da prendere in esame durante le fasi successiven le quali dovranno essere sviluppate per adattarsi a ciò che

viene valutato.

Per effettuare un'analisi ottimale è necessaria l'interazione tra le diverse figure che prendono parte al progetto, in quanto il background e le conoscenze di ognuno di essi sono differenti e solitamente non sono completi rispetto alla questione da trattare.
Di solito chi genera i dati e li ha a disposizione non ha la capacità di interpretarli, per cui commissiona le analisi a terzi, specializzati nell'utilizzo degli strumenti e delle tecniche necessarie allo studio, ma esterni al settore a cui appartengono i dati. I soggetti che prendono in carico i dati devono poter capire il significato di ogni *feature* che è coinvolta nel problema e i valori che queste possono assumere, per poter creare in un secondo momento un modello efficace. L'interscambio di informazioni permette quindi alle figure terze di avere un quadro chiaro della situazione e di poter procedere con le analisi.

2.2 Pre-processamento

Una volta conclusa la valutazione del problema, si passa alla trasformazione dei dati.
Questa trasformazione comincia dagli archivi messi a disposizione da chi commissiona lo studio. Si parla di trasformazione in quanto nella maggior parte dei casi i dati non sono direttamente computabili a causa della struttura con cui sono memorizzati, non pensata per le esigenze dagli algoritmi. È più facile infatti capire come poter studiare i dati già presenti piuttosto che modellare un sistema dedicato alla raccolta delle informazioni necessarie.

La raccolta dei dati sufficienti può richiedere molto tempo ed essere costoso, per cui non è possibile in molti casi realizzare sistemi appositi per raccogliere i dati da studiare in modo prospettico. Un lavoro di questo genere può essere fatto se i dati già disponibili non sono sufficienti per produrre dati validi, se le tempistiche lo permettono e se esiste ancora la possibilità di reperire le informazioni.

Al giorno d'oggi è comune di quasi tutte le aziende l'utilizzo di strumenti informatici per memorizzare le informazioni legate all'attività che esse svolgono, il che rende facile reperire i dati.

Per estrarre e sfruttare i raw data, ossia i dati presenti negli archivi nella loro forma grezza, è necessario effettuare un lavoro di pre-processamento.

Il pre-processamento dei dati consiste inizialmente nel creare una struttura sui dati, nel caso questa non fosse presente, e modellare i dati in maniera opportuna, al fine di creare il dataset su cui si andrà a lavorare. Il dataset (letteralmente "collezione di dati") è la struttura di base del Machine Learning, sulla quale verranno applicati tutti gli algoritmi.

I dati a disposizione si dicono **strutturati** nel caso in cui siano organizzati in schemi e tabelle rigide. Questa modalità nella maggior parte dei casi è mediamente facile da trattare. Si dice invece che sono **non strutturati** quando, al contrario, non esistono regole rigide e sistematiche sulla forma, come nel caso di contenuti testuali o multimediali. I dati non strutturati sono più complicati da trattare ed esistono intere discipline che studiano come trattarli per estrapolare da essi un pattern. Questa ultime non verranno approfondite in questo testo.

Una volta strutturate le informazioni, dalla collezione deve essere eliminato il "rumore", cioè quelle informazioni non rilevanti che possono causare imprecisioni nelle valutazioni e nei calcoli degli algoritmi.

È in questa fase che si correggono anche gli eventuali errori di inserimento dei dati. L'errore umano crea rumore e i valori non corretti possono portare a risultati non veritieri. Se per esempio una persona è facile interpretare parole come "Si", "sì" e "si" nello stesso modo, per un compilatore la differenza di accenti, maiuscole e minuscole basta per trattare i valori in maniera distinta.

Sebbene una piccola percentuale di errore nella maggior parte dei database sia considerata intrinseca nella raccolta dei dati e quindi accettabile, questa deve essere ridotta il più possibile.

I dati devono essere portati ad una forma normalizzata che non deve contenere nessun livello di ambiguità per quanto riguarda le *feature* considerate e la loro interpretazione deve essere univoca. Ogni variabile deve inoltre rientrare in un dominio di valori ammissibili, che per ogni caratteristica deve poter essere ricavabile dalle informazioni fornite durante l'analisi da chi conosce il significato specifico dei dati.

Una volta eseguita la pulizia da errori e inconsistenze, è quasi sempre possibile estrarre l'insieme di dati strutturati (pattern) su sui poter lavorare nel resto dei passaggi.

In alcuni casi tuttavia i dati devono essere trasformati ulteriormente. A seconda degli strumenti e degli algoritmi che si applicano possono essere infatti espressi dei vincoli ulteriori sulla forma dei dati.

I più comuni sono:

- avere nell'intero dataset solo valori numerici. Per rispondere a questa esigenza si può intervenire sui dati categorici: per quelli non ordinati è possibile mappare ogni valore in una nuova *feature* booleana, in cui il valore sarà 1 per le righe in cui il valore compariva, 0 altrimenti; per le colonne in cui importa l'ordine si possono mappare i valori con numeri crescenti.

- non avere dati mancanti. Questi devono essere gestiti o eliminando le righe con valori nulli (rischioso perché causa perdita di informazioni ed è possibile solo nel caso in cui queste costituiscano una minima parte dei dati) o trasformare i valori mancanti nel loro valore più probabile, o in un valore neutro (come la media dell'intervallo) o in un valore chiaramente non appartenente al dominio.

- uniformare i valori per renderli confrontabili. Spesso i domini diversi causano difficoltà nel momento di applicare calcoli trasversali alle *feature*. Esiste per questo lo Scaling, un metodo che permette di normalizzare le variabili indipendenti in uno stesso intervallo.

- avere dati bilanciati. Questo problema esiste nei dataset in cui le istanze per ogni classe sono presenti in proporzione diversa, creando uno sbilanciamento dei record. Può essere necessario attribuire alle classi un peso diverso nei calcoli o intervenire sul dataset per aggiungere record nelle classi in inferiorità numerica.

Dopo questi passaggi è possibile passare alla fase di decisione degli algoritmi migliori per il problema.

2.3 Algoritmi di Machine Learning

Gli algoritmi di ML vengono applicati partendo dalla funzione che modella il problema. A seconda che si conoscano, oltre le variabili di input, anche quelle di output attese, si parla di apprendimento supervisionato o non supervisionato. Nel primo caso per ogni oggetto del dataset si conosce già l'output atteso; si dice quindi che il dataset è etichettato e lo scopo del modello sarà di trovare le regole che massimizzino la corretta etichettatura di nuove istanze. Nel secondo caso invece gli output sono ignoti e tra gli oggetti dovranno essere cercate regole associative che accomunino i pattern, in modo da per poter fare ragionamenti e previsioni sui nuovi input, senza porre vincoli sulle classi attese. Nel caso di apprendimento con approccio supervisionato si possono avere problemi

di classificazione, regressione e riduzione delle dimensionalità. L'approccio non supervisionato invece può essere ricondotto a problemi di clustering, riduzione delle dimensionaltà e representation learning.

2.3.1 Classificazione

Nei problemi di classificazione si vuole determinare a quale classe appartenga un record, sulla base delle regole di classificazione ottenute dal dataset precedentemente studiato.
Gli algoritmi più utilizzati per questo tipo di problema sono: classificatore Bayesiano, Regressione Logistica, SVM, Nearest Neighbor, Decision Tree.

Classificatore Bayesiano Questo classificatore per determinare il modello migliore si basa su teoremi probabilistici, in particolare sul Teorema di Bayes, il quale valuta la probabilità che una causa scateni un evento con la seguente formula:

$$P(w_i \mid X) = \frac{P(X \mid w_i)\, P(w_i)}{P(X)}$$

Dove:

- X è il valore in input
- w_i è la classe di output
- $P(X)$ è la densità assoluta del pattern X, ossia la probabilità che il prossimo elemento da predire sia uguale a X
- $P(w_i)$ è la probabilità marginale della classe rispetto al totale delle classi
- $P(w_i \mid X)$ è la probabilità condizionata della classe dato il pattern, cioè la misura in cui si ritiene che la classe predetta sia w_i, dando per certo il valore di X.
- $P(X \mid w_i)$ è la probabilità condizionata di X, nota la classe w_i

La formula permette di calcolare la probabilità che il nuovo input appartenga alla classe w_i invertendo il problema di partenza, infatti come mostra l'equazione essa può essere derivata se si conoscono le singole probabilità di w_i e X e la ricorrenza di X rispetto alla classe w_i che si sta considerando.
Per utilizzare questo tipo di modelli è necessario avere quindi informazioni sulle probabilità degli eventi, non sempre note ma spesso derivabili dalla distribuzione dei dati.

Regressione Logistica La Regressione Logistica viene utilizzata per modelli con output dicotomici (due possibili valori) per individuare un piano di separazione tra i dati.
È un tipo di classificatore lineare il cui scopo è trovare l'iperpiano migliore che divida i dati, per poi mappare gli input in un valore continuo appartenente all'intervallo [0,1]. Questo valore indica la probabilità che il pattern in input appartenga ad una delle due classi. Se il risultato della mappatura dell'istanza è < 0.5, allora apparterrà alla prima classe, e in caso contrario alla seconda.

La funzione che descrive questo classificatore è detta sigmoide ed è data da

$$\sigma(X) = \frac{1}{1 + e^{(b+w \cdot X)}}$$

in cui $(b + w \cdot X)$ descrive l'iperpiano che divide le classi.

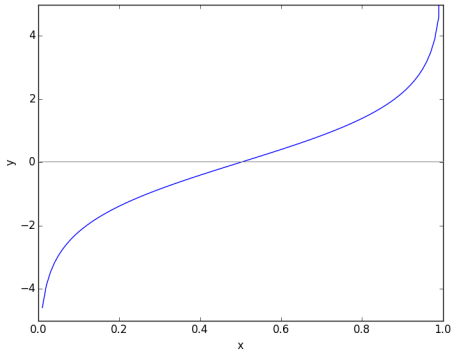


Figura 2.1: Sigmoide

SVM L'SVM, acronimo di Support Vector Machine, è un modello che sfrutta lo stesso principio di separazione dei dati della regressione logistica, quindi tramite la descrizione di un iperpiano.
La differenza sostanziale proposta da questo algoritmo consiste nel ricercare l'iperpiano che massimizzi la distanza tra le due classi, sfruttando i support vectors. I support vectors sono quelle istanze che si trovano sul bordo della soglia di separazione delle classi, la cui distanza, se massimizzata, permette di trovare il piano che possa prestarsi meglio alla previsione dei nuovi input.
La retta che costituisce il piano intermedio tra le due classi sarà l'insieme di punti descritti da:

$$(b + w \cdot X) = 0$$

mentre assumerà valore 1 e -1 sui support vectors delle due classi da prevedere.

Nearest Neighbor Il Nearest Neighbor applica l'idea secondo cui se due oggetti in input sono molto vicini nello spazio e quindi si assomigliano, allora si può dedurre che appartengano alla stessa classe.

$$P(w_i \mid X) \approx P(w_i \mid X')$$

Una tecnica più precisa e affidabile rispetto al considerare solamente il vicino più simile, è quella di valutare k vicini, con k definibile arbitrariamente. Col primo metodo infatti il più piccolo errore sul dataset causerebbe la predizione scorretta di tutti i nuovi input che si trovino vicini all'istanza errata.
Il k-Nearest Neighbor considera non uno bensì i k record più vicini, ed etichetta il pattern in base alla classe più ricorrente tra i vicini.

Decision Tree Il Decision Tree è un algoritmo che permette di estrarre un albero di classificazione dai dati. Un albero è una struttura gerarchica costituita da nodi, che contengono le informazioni, e da archi, che costituiscono i collegamenti gerarchici tra i nodi. Inoltre i nodi possono avere un solo arco entrante e zero o più archi uscenti. I nodi senza nessun arco uscente sono detti foglie.

In un Decision Tree i nodi interni contengono le regole di classificazione, le quali sono scelte in base ad una singola *feature* alla volta. Ogni nodo contiene una condizione discriminante mentre ogni foglia contiene l'etichetta corrispondente alla classificazione.

Esistono diversi tipi di classificatori per sviluppare alberi binari:

- ID3, che crea alberi a più vie (ossia in cui ogni nodo non foglia ha due o più archi uscenti) utilizzando un criterio greedy su ogni nodo per selezionare la *feature* categorica che divida in maniera migliore le classi (*split*).

- C4.5, il quale aggiunge la possibilità di lavorare su dati numerici oltre che categorici e traduce i valori continui in intervalli di valori. Le regole di divisione migliori per separare gli output sono ordinate in base all'accuratezza di divisione degli intervalli.

- C5.0, versione successiva a C4.5 che ne ottimizzata la gestione della memoria e la definizione delle regole di split. Non è un algoritmo open source.

- CART è simile a C4.5 ma costruisce alberi binari nei quali è possibile gestire i valori continui.

2.3.2 Regressione

La regressione permette di formulare un problema per prevedere variabili continue. Partendo dalla funzione che descrive il modello, la regressione mette in relazione gli input indipendenti x con gli output dipendenti y e ricerca i coefficienti di x che minimizzino l'errore su y.

A seconda del grado della funzione e del numero di variabili indipendenti, i modelli possono essere divisi in: Regressione Lineare Semplice, Regressione Lineare Multipla, Regressione Non Lineare.

Regressione Lineare Semplice Si usa se la variabile indipendente che prende parte al problema è una sola. In questo caso ognuno degli n record è descritto dalla relazione

$$y_i = \alpha + \beta \cdot x_i + \epsilon_i$$

In cui:

- ϵ è l'errore intrinseco alla funzione
- α e β sono rispettivamente il termine noto e il coefficiente della retta, ossia incognite da trovare

Le incognite in questo caso possono essere semplicemente trovate cercando il minimo errore della retta ai minimi quadrati, in cui è minima quindi:

$$\sum_{i=1}^{n} \epsilon^2$$

.

Regressione Lineare Multipla Se le variabili coinvolte sono molteplici, allora ogni x può essere rappresentata da un vettore, rendendo possibile la creazione di una matrice X in cui ogni riga corrisponde ad un oggetto del pattern e ogni colonna ad una *feature* diversa. Anche y e β, che per ogni variabile sono degli scalari, complessivamente vengono rappresentati come vettori. Questo permette la modellazione del problema in forma matriciale e il calcolo del minimo.

Regressione Non Lineare Nel caso in cui la funzione sia di grado superiore, allora non sarà più descritto da una retta ma da una curva o da un iperpiano. Questo permette di risolvere problemi in cui la dipendenza tra x e y non è lineare.

Il grado più elevato del problema rende necessario l'utilizzo di tecniche più avanzate per minimizzare l'errore. Le più comuni sono la discesa sul gradiente e l'algoritmo GaussNewton.

2.3.3 Clustering

Il Clustering viene utilizzato per trovare raggruppamenti tra i dati in modo non supervisionato, cioè in problemi in cui le classi non sono note a priori.

A causa delle numerosissime partizioni possibili su un pattern, è indispensabile definire criteri di ottimalità per ridurre i possibili guppi formati dalle classi. Il criterio della minimizzazione dei centroidi ad esempio è ottimo per strutture che assumono forme circolare. Tuttavia esso si comporta male con figure allungate o innestate, che al contrario vengono descritte in maniera migliore minimizzando la distanza intra-classe.

Clustering gerarchico Gli algoritmi a struttura gerarchica vengono utilizzati per creare un ordine di gerarchia tra i raggruppamenti, partendo da dati divisi in modo unitario fino ad avere raggruppamenti che includono molti valori comuni.

L'approccio che viene utilizzato è detto *bottom-up*. Si parte da gruppi unitari separati e si aggregano ad ogni iterazione quelli con distanza minore.

Il risultato dell'aggregazione può essere rappresentato tramite un dendrogramma, grafico utile per visualizzare le strutture di gruppo.

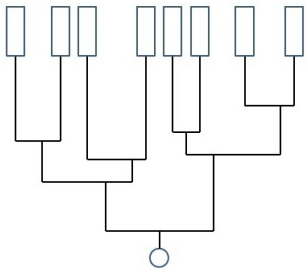


Figura 2.2: Dendogramma

Il modo in cui viene calcolata la distanza può far variare molto il risultato. Alcuni dei metodi più utilizzati per calcolare quali cluster raggruppare ad ogni *step* sono:

- Single link: considera i cluster che hanno i due pattern più vicini, quindi con distanza minima
- Average link: sceglie i cluster in cui la distanza media tra tutti i pattern è minima
- Complete link: al contrario del Single link, cerca la distanza massima tra due pattern

Clustering di partizione Questo insieme di algoritmi punta a minimizzare la distanza rispetto ai centroidi per ottenere k partizioni, con k arbitrario.

L'algoritmo più semplice di questa famiglia di algoritmi è il k-Means. Il k-Means parte da una soluzione base casuale e iterativamente la migliora. A causa della sua semplicità, il k-Means può creare problemi se si incorre in minimi locali, i quali non permettono di trovare una soluzione ottima del problema.

Anche il Fuzzy k-Means sfrutta lo stesso principio di miglioramento iterativo del k-Means, ma permette inoltre di specificare la probabilità che ha ogni input di appartenere ai vari cluster, in modo da avere soluzioni potenzialmente più robuste.

Clustering di densità I cluster basati sulla densità vengono ricavati connettendo regioni di dati in cui la densità è alta.

Il DBSCAN è un algoritmo che permette di calcolare queste densità. Presi due punti, l'algoritmo li considera appartenenti allo stesso cluster se la distanza tra loro è minore di una certa soglia e se tra i due punti sono circondati da sufficienti altri punti.

2.3.4 Riduzione delle dimensionalità

La riduzione delle dimensionalità ha come scopo la diminuzione della complessità del problema iniziale, attraverso la riduzione dello spazio del problema ma con l'obiettivo di preservare il più possibile il contenuto informativo dei dati.
In un problema il cui dominio ha dimensione m, con m uguale al numero di *feature* iniziali, si vuole ottenere uno spazio con dimensione n, con $n < m$, mantenendo il più alto valore informativo delle dimensioni restanti.

Questi tipi di modelli spesso vengono utilizzati come pre-processamento nei problemi ad elevata dimensionalità, quindi prima di applicare altri algoritmi. Questo viene fatto per avere problemi meno complessi, in cui le *feature* prese in esame sono soltanto le più influenti e per permette di eliminare le informazioni marginali che potrebbero portare a modelli troppo complessi o poco generalizzabili.

I modelli principali utilizzati in questo contesto sono: PCA, LDA e t-SNE.

PCA La Principal Component Analysis (PCA) è una trasformazione lineare non supervisionata che ha lo scopo di mantenere più informazioni sui dati a prescindere dalla classe di appartenenza di essi.

LDA Al contratio della PCA, la Linear Discriminant Analysis (LDA) è una tecnica lineare supervisionata in cui si vogliono preservare maggiormente le informazioni correlate alla classe di appartenenza dei pattern.

t-SNE L'algoritmo t-distributed Stochastic Neighbor Embedding (t-SNE) utilizza un modello non lineare e non supervisionato. La non linearità permette di trovare separazioni migliori in casi complessi.

Questo tipo di approccio è utilizzato principalmente per la rappresentazione in 2D e 3D dei dati.

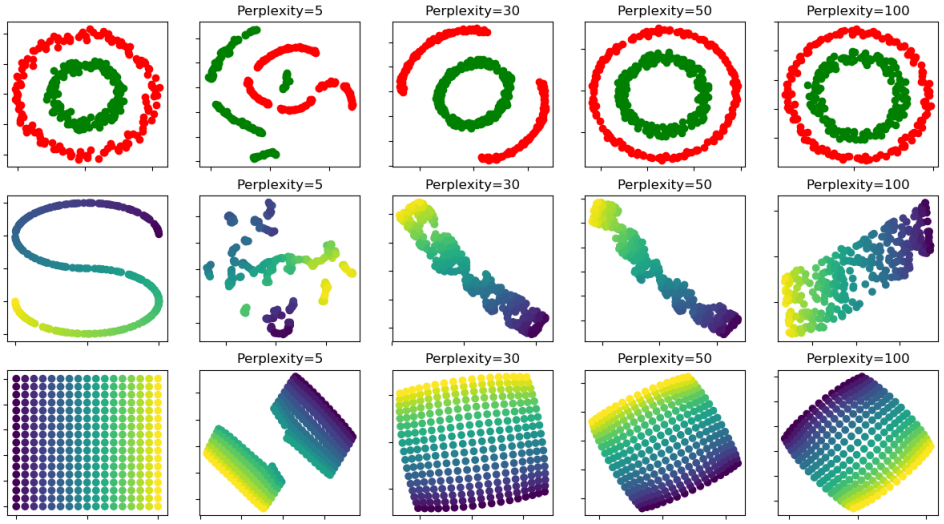


Figura 2.3: Alcuni esempi di grafici realizzati con t-SNE

2.4 Addestramento dell'algoritmo

Quando si addestra un algoritmo, ossia gli si fanno studiare i dati, è possibile impostare delle variabili dette iperparametri. Esse sono variabili che possono essere cambiate per migliorare le prestazioni e la qualità dell'algoritmo.

Per valutare quale algoritmo sia il migliore rispetto a quelli utilizzati o in relazione a iperparametri diversi, è indispensabile poter valutare le prestazioni dei modelli sviluppati su nuovi dati per capire quanto questo sia performante sui dati. Per poterlo fare è necessario testare l'algoritmo sui dati.
Non sempre è possibile controllare la correttezza dei modelli su dati effettivamente nuovi. Per ovviare alla necessità di effettuare test in tempo reale non dovendo però attendere nuovi dati, nel Machine Learning si utilizzano

delle tecniche che permettono di dividere i dati a disposizione in sottoinsiemi, ognuno dei quali è utilizzato con uno scopo diverso:

- Training set, che è l'insieme di dati sul quale si effettuerà l'addestramento dell'algoritmo, ossia si cercano i valori dei parametri che meglio possano descrivere i dati in funzione del problema da risolvere.

- Validation set, sul quale si correggeranno i valori degli iperparametri. Anche su questo set di dati l'algoritmo verrà addestrato.

- Test set, il quale servirà per testare la capacità dell'algoritmo di generalizzare in modo corretto le regole estrapolate dal Training set.

La trasformazione dei sottoinsiemi precede l'applicazione dell'algoritmo e bisogna considerare che la selezione di un sottoinsieme dei dati deve essere rappresentativo dell'insieme di partenza.

Esistono due modalità principali per dividere i Training/Validation set dal Test set: Holdout e k-fold Cross Validation.

Nella Holdout la divisione tra i due set avviene dividendoli semplicemente, solitamente 70% dei dati sono riservati per l'addestramento e il 30% per il test. Questo può causare problemi nel caso in cui i set non siano entrambi rappresentativi dei dati iniziali.

La k-fold Cross Validation risolve questo problema dividendo i set in k sottoinsiemi e iterando l'addestramento dell'algoritmo k volte. In ogni iterazione viene scelto un sottoinsieme diverso per il test e le prestazioni finali dell'algoritmo vengono calcolate come media dei vari addestramenti.

Per dividere ulteriormente il Training set dal Validation set è possibile applicare nuovamente le tecniche sopra elencate, stavolta solo sull'insieme di dati da addestrare.

2.5 Analisi delle prestazioni

Una volta che l'algoritmo è stato addestrato sul Training e sul Validation set, è possibile valutare le sue prestazioni sul Test set. Le prestazioni misurano la capacità dell'algoritmo di effettuare una previsione corretta sul dominio in esame e variano a seconda del modello scelto.

Di seguito si elencheranno solo le metriche degli algoritmi supervisionati, per i quali è possibile conoscere il valore reale atteso.

Classificazione

Nel caso della classificazione è possibile creare metriche di misurazione che sfruttano i valori corretti ricavati dai dati forniti e quelli predetti dal modello.

Prendendo come esempio un problema con due classi, possiamo definire le relazioni tra i valori attesi e quelli predetti nel seguente modo:

- TP (True Positive), le istanze della prima classe predette correttamente
- FP (False Positive), le istanze della seconda classe attribuite alla prima
- TN (True Negative), le istanze della seconda classe predette correttamente
- FN (False Negative), le istanze della prima classe attribuite alla seconda

Questo concetto può essere generalizzato, in un problema con n classi avremo una matrice risultante di dimensione $n \times n$.

La matrice risultante dalla combinazione dei risultati delle classi è detta **Matrice di Confusione**.

		Classe Attesa	
		Classe 1	Classe 2
Classe	Classe 1	TP	FP
Predetta	Classe 2	FN	TN

Tabella 2.1: Confusion Matrix

I numeri presenti nella Matrice di Confusione corrispondono al numero di record classificati, secondo le relazioni descritte sopra. Ne nel caso ottimo la matrice è diagonale.

Le metriche derivate dalla matrice e utilizzate comunemente nei problemi di classificazione sono riportate nella tabella sottostante.

Accuracy	
$\frac{TP+TN}{TP+TN+FP+FN}$	Percentuale di righe predette correttamente. Viene utilizzata come metrica principale ma non è affidabile in caso di dataset molto sbilanciati.

Error rate	
$\frac{FP+FN}{TP+TN+FP+FN}$	Percentuale di righe predette erroneamente. Speculare rispetto all'accuratezza.

Precision	
$\frac{TP}{TP+FP}$	Percentuale di record classificati come positivi rispetto a tutti quelli positivi.

Recall	
$\frac{TP}{TP+FN}$	Percentuale di record realmente positivi rispetto a quelli classificati come positivi.

F1-mmeasure	
$\frac{2 \cdot TP}{2 \cdot TP+FP+FN}$	Media armonica tra Precision e Recall.

Cost based evaluetion Esistono problemi in cui TP, TN, FP e FN non hanno lo stesso peso e può essere necessario preferire le soluzioni con un ridotto numero di uno piuttosto che di un altro.

La Cost Based Evaluation permette di attribuire pesi diversi ai risultati della matrice di confusione, aumentando o diminuendo il costo di classificazione di ogni record. Si crea una matrice di costi, calcolata aggiungendo le penalità alle classi. Il costo totale del classificatore sarà dato dalla somma dei costi.

$$costo = \alpha_1 \cdot TP + \alpha_2 \cdot FP + \alpha_3 \cdot TN + \alpha_4 \cdot FN$$

Questo permette di trovare l'algoritmo migliore secondo i nuovi criteri, che corrisponderà al modello con costo minore.

ROC Le curve ROC (Receiver Operating Characteristic) permette di mette in relazione la sensitività con la specificità o la precisione. Questi valori corrispondono a:

sensitivity = true positive rate = $\frac{TP}{TP+FN}$

false positive rate = 1− sensibility = $1 - \frac{FP}{FP+TN}$

precision = $\frac{TP}{TP+FP}$

Graficamente si può dire che, confrontando più curve ROC, quella più vicina a vertice in alto a sinistra è migliore.

Nel caso visivamente risulti difficile mettere a confronto più curve, è necessario calcolare l'area sotto la curva, detta AUC (Area Under Curve).

Il vantaggio di questa curva è che non è influenzata dallo sbilanciamento del dataset.

Regressione

I metodi più utilizzati per valutare algoritmi di regressione sono RMSE e R^2.

Il RMSE (Root Mean Square Error, o radice dell'errore quadratico medio) permette di valutare la bontà del modello considerando l'errore sui dati predetti, che per questi tipi di algoritmi sono valori continui. La sua formula è:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}-y_i)^2}{n}}$$

La metrica R^2 invece viene utilizzata per mettere in relazione i dati con la variabilità del modello. Questo valore è compreso tra 0 e 1, dove un valore che si avvicina all'1 indica che il modello descrive in maniera ottimale i dati, mentre più è vicino allo 0 più il modello non è in relazione coi dati. La formula per il calcolo di R^2 è:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}$$

Errori di classificazione

Addestrando un insieme di dati e valutandone le prestazioni si può incorrere in diversi errori:

- Training Error: è l'errore dell'algoritmo rispetto ai dati sui quali si è addestrato.
- Generalization error: sono legati alla scarsa capacità dell'algoritmo di fare previsioni di dati su cui non è stato addestrato.
- Underfitting: Il modello è troppo semplice e non garantisce prestazioni buone né sul Training che sul Test set

- Overfitting: Il modello è troppo complesso e specializzato rispetto ai dati dell'addestramento, il che causa difficoltà nel prevedere dati nuovi.

Nel caso si presentino questi errori è bene addestrare nuovamente l'algoritmo imponendo vincoli e criteri diversi.

2.6 Tecnologie

Sono molteplici i *tool* disponibili per l'analisi dei dati e l'utilizzo di algoritmi di Machine Learning.

Questi si differenziano per l'astrazione, alcuni infatti sono più di alto livello di altri. Spesso dalla maggior astrazione derivano anche più limiti del software, per cui non tutti gli ambienti di lavoro supportano gli stessi algoritmi.

Molti di questi framework richiedono la conoscenza del linguaggio di un programmazione. Tuttavia, a causa del crescente interessamento da parte di figure non esperte di programmazione o matematica, sono sempre di più i software semi-automatizzati sviluppati per rendere più facile e veloce l'utilizzo dei modelli di base, anche senza bisogno di conoscere il funzionamento esatto e dettagliato degli algoritmi.

Weka Weka è un software basato su Java e, sebbene sia possibile utilizzarlo all'interno di programmi scritti in Java, l'utilizzo più comune di questo è basato sull'interfaccia grafica.

Lo strumento più interessante per l'analisi dei dati è l'Explorer, il quale permette di caricare i dati direttamente da un dataset, avere informazioni sui domini degli attributi (es. valore minimo, massimo, medio, numero di valori distinti, valori non presenti) e visualizzarne graficamente la distribuzione, sia rispetto alla singola *feature* che mettendone due a confronto. Inoltre è possibile effettuare operazioni preliminari di preparazione ed eseguire algoritmi di classificazione, clustering, feature selection e determinazione di regole associative.

Scikit-learn Scikit-learn è una libreria Open Source di Python, un linguaggio di programmazione general purpose di alto livello, diventato famoso per essere stato tra i primi linguaggi ad integrare la possibilità di lavorare con complessi algoritmi di ML.

La libreria implementa la maggior parte degli algoritmi di Machine Learning tradizionale. È abbastanza facile da utilizzare ma è limitata per gli aspetti più avanzati, come le Reti Neurali e il Deep Learning. Viene utilizzata principalmente in combinazione con altre librerie che la rendono molto efficace per

tutti gli aspetti di data mining. Per la gestione dei dati si utilizzano di base NumPy e Pandas e per la visualizzazione grafica Matpltlib.

Tensorflow Rispetto a Scikit-learn, Tensorflow è più di basso livello, il che la rende più difficile da utilizzare ma le permette di non avere limiti sulla gestione degli algoritmi più complessi.

Grazie all'idea di non nascondere i passaggi dell'algoritmo tramite astrazione ma permettendo all'utente di scrivere i vari passaggi da effettuare e poi combinarli con semplici operatori, questa libreria è ottima per addestrare complesse Reti Neurali e per la modellazione degli algoritmi avanzati utilizzati nel Deep Learning.

Sempre in ambito di Deep Learning, Tensorflow implementa la differenziazione automatica, un insieme di tecniche indispensabili nelle Reti Neurali per il calcolo automatico delle derivate. Questo meccanismo viene utilizzato nella backpropagation (o propagazione dell'errore), uno delle modalità utilizzate per ottimizzare i modelli durante l'addestramento delle Reti Neurali.

È disponibile una versione meno pesante della libreria, Tensorflow Lite, che permette l'utilizzo del Machine Learning sui dispositivi mobile.

La gestione delle risorse è resa performante grazie alla possibilità di spostare calcoli complessi sulla GPU.

PyTorch Anche questo strumento è progettato per lo sviluppo di algoritmi in Pyhton. Permette la progettazione sia di modelli tradizionali che di Deep Learning, tuttavia ha più limitazioni rispetto a Tensowflow.

Questa libreria sfrutta in modo ottimale la potenza dell'elaboratore, appoggiandosi sia alla GPU che alla CPU per addestrare gli algoritmi.

Recentemente il suo utilizzo è stato esteso al mondo mobile, sia Android che IOs.

La versione attuale di PyThorch comprende al suo interno Caffe2, una libreria che è stata inglobata a questa. La ragione è legata a Facebook, uno dei maggiori utilizzatori di entrambe le librerie, che nel 2018, grazie anche all'aiuto di Microsoft, ha abbandonato la progettazione parallela tra queste per inglobarle in un'unica piattaforma.

Capitolo 3

Progetto DIANA

I risultati evidenti che il Machine Learning sta portando nel settore della sanità hanno fatto nascere l'idea di integrare queste tecniche anche nella realtà medica della città di Cesena, in particolare presso l'Ospedale M. Bufalini Cesena AUSL Romagna, il quale sta già lavorando a 360 gradi per diventare un Ospedale 4.0.

L'Ospedale 4.0 è un nuovo modo di vedere il contesto ospedaliero, ripensandolo e riorganizzandolo per aumentare l'efficienza e l'efficacia dei servizi sanitari. Lo scopo è di migliorare le diagnosi, la gestione interna, l'organizzazione e la comunicazione di pazienti e medici, utilizzano innovazioni tecnologie e conoscenze informatiche.

Gli ambiti di innovazione nell'Ospedale 4.0 sono svariati, come la già citata Intelligenza Artificiale, la robotica e l'Internet Of Things (letteralmente "Internet delle cose", che consiste nell'integrazione della tecnologia negli oggetti di uso comune e permettere loro di connettersi con altri oggetti e dispositivi).

L'idea di sfruttare il Machine Learning è partita da una problematica sollevata dai Medici Anestesisti Rianimatori (U.O.C Anestesia e Rianimazione) e dai Medici Ginecologi e Ostetrici (U.O.C Ginecologia e Ostetricia) dell'ospedale in questione.

3.1 Motivazioni

Durante una gravidanza è indispensabile valutare correttamente i fattori che definiscono il profilo di rischio di un parto, col fine di impostare in modo ottimale il piano assistenziale da offrire alla madre durante la gestazione, il travaglio, il parto e il post-parto e per ridurre la probabilità di complicazioni inattese.

La scelta di questo percorso deve essere valutata da diverse figure mediche all'interno dell'ospedale quali: ginecologo, ostetrica e anestesista.
La presenza dell'anestesista non è una costante, infatti il suo intervento è necessario solo nel caso in cui venga fatta richiesta di parto-analgesia, ossia l'utilizzo di terapie per il contenimento del dolore durante il parto. Sebbene non sia tipico di tutti i parti, negli ultimi anni è sempre più frequentemente l'utilizzo di queste tecniche per diminuire il dolore.

In Emilia-Romagna il suo impiego è in crescita, nel 2017 ad esempio è stata usata per il 21,7% dei parti, in aumento dall'anno precedente [10].

La problematica non riguarda tanto l'aumento di richieste di parto-analgesia ma la relazione che, sempre secondo i rapporti CeDAP, c'è tra essa e il ricorso al parto cesareo.
Nel rapporto si può leggere infatti: "Escludendo i parti senza travaglio (quindi gli interventi elettivi e urgenti fuori travaglio), la frequenza di taglio cesareo risulta maggiore nei travagli con epidurale (16,1%) rispetto a quelli senza (8,5%). La differenza rimane anche considerando i soli parti a inizio spontaneo (12,4% con epidurale, 6,1% senza epidurale). Anche la frequenza di parto vaginale operativo è maggiore in caso di epidurale (9,4% vs.4,7%). Analizzando i dati delle donne con travaglio e gravidanza con feto singolo, l'analgesia epidurale è associata a un maggior rischio di parto operativo vaginale e di taglio cesareo, anche aggiustando per possibili confondenti."

3.2 Letteratura medica

La necessità di capire se vi sia un nesso tra le due cose è legato al richiamo che c'è stato nel 2018 da parte dell'OMS sulla percentuale troppo alta di parti cesarei in Italia, circa 10% in più rispetto alla media europea.

Il taglio cesareo in sé non è un problema se è effettuato in modo elettivo e per motivi di urgenza fuori dal travaglio, ma spesso viene utilizzato, sempre secondo l'OMS, anche quando si è in una situazione in cui il cesareo potrebbe essere evitato se si lasciasse abbastanza tempo alla gestante per raggiungere la dilatazione necessaria per avere un parto naturale. Il cesareo andrebbe evitato se non necessario perché è un intervento chirurgico e come ogni intervento aumenta i rischi di complicanze sia durante che dopo il parto, oltre che aumentare notevolmente i costi per gli ospedali[11].

Sebbene in Italia il trend di parti cesarei sia negativo (-7,3% di parti con taglio cesareo tra il 2011 e il 2016), la media nazionale è ancora una delle più alte in Europa 3.1.

Come mostrano i grafici pubblicati dall'Osservatorio Nazionale sulla Salute nelle Regioni Italiane, nel 2016 l'Italia era il terzo paese per numero di parti cesarei [12].

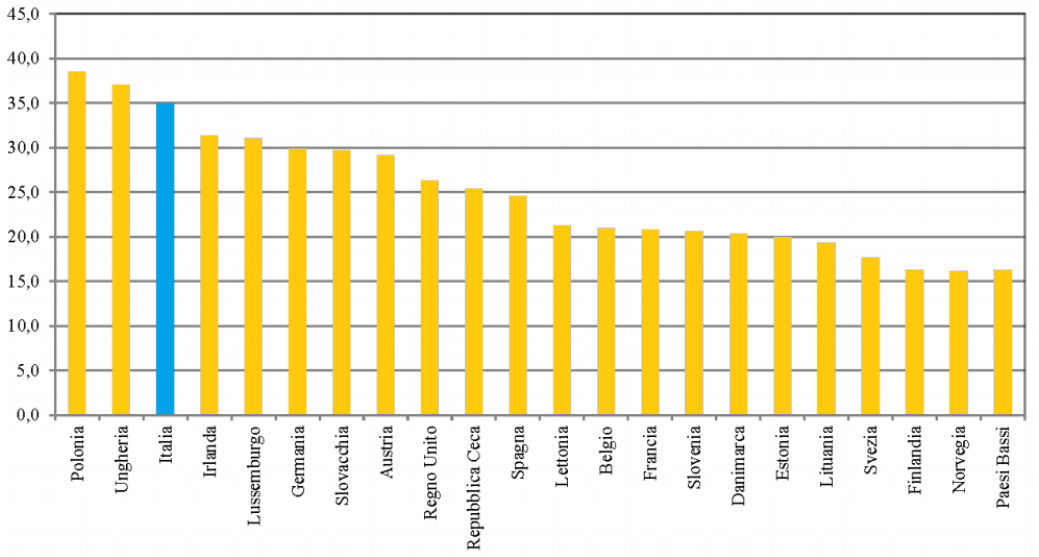


Figura 3.1: Percentuale di parti cesarei per alcuni paesi Europei nel 2016 [12]

Il dibattito sul fatto che l'analgesia possa determinare un aumento di incidenza di tagli cesarei per le pazienti che effettuano l'analgesia è ancora aperto, infatti non vi sono revisioni mediche recenti in merito che evidenzino una correlazione causa-effetto tra le due cose.

3.3 Obiettivi

L'utilizzo di tecniche tradizionali per lo studio dei dati non ha permesso di produrre risultati esaustivi che eliminassero ogni perplessità sui motivi che causano l'espletamento di un parto eutocico con complicanze materne.

È nata quindi una collaborazione tra l'Ospedale e l'Università Alma Mater Studiorum - Università di Bologna, Department of Computer Science and Engineering, unendo quindi le conoscenze informatiche a quelle mediche, con l'intento di utilizzare tecniche più avanzate per analizzare i dati.

Questa collaborazione ha dato vita presso l'Ospedale M. Bufalini di Cesena al progetto DIANA, nato dalla necessità di capire se vi sia effettivamente una correlazione tra analgesia e cesareo, e con lo scopo di affrontare lo studio dei

dati di analgesia in maniera prospettica tramite tecniche diMachine Learning.

Entrando più nel dettaglio, l'obiettivo del progetto è quello di ricercare dai dati in possesso quali fattori possano influenzare in modo negativo l'outcome materno naturalmente atteso, costringendo quindi al ricorso del parto cesareo e se esistano dei fattori tipici soltanto dell'analgesia che possano spiegare l'incidenza maggiore.

3.4 Studi futuri

L'intenzione nel medio-lungo periodo è di integrare il Progetto DIANA con il progetto CERERE, acronimo di di "maChinE leaRning in travaglio di parto con e senza analgEsia spino-periduRalE: outcome materno".

Questo secondo progetto è nato per rispondere alle necessità dell'ospedale di raccogliere facilmente i dati delle pazienti ricoverate per il travaglio di parto. Attualmente infatti i dati non vengono raccolti subito in modo digitale. La procedura prevede che le informazioni vengano riportate in un primo momento sulla cartella clinica cartacea e successivamente su quella elettronica, dallo stesso medico o da uno diverso. Questa modalità è problematica perché aumenta i rischi di errori nella raccolta dei dati e raddoppia il lavoro svolto dai dottori.
Il progetto vuole rendere possibile evitare questo duplice passaggio, rendendo la raccolta i dati più veloce, sicura e accurata dalle metodologie utilizzate attualmente e snellire e facilitare il lavoro dei dottori.

L'unione dei due progetti sarà significativa una volta che, raccolti abbastanza nuovi dati tramite le carelle elettroniche, si potrà essere in grado di applicare gli strumenti prospettici anche sui nuovi dati delle pazienti che effettuano l'analgesia spianale, peridurale o spino-peridurale.

Con il progetto CERERE si vuole inoltre estendere la ricerca degli indicatori correlati all'aumento del rischio di parto cesareo. I soggetti di questo studio saranno tutte le pazienti, cioè sia quelle che utilizzeranno metodi di contenimento del dolore, sia quelle che non ne faranno ricorso, al fine di trovare più fattori e correlazioni.

Capitolo 4

Caso di studio

Per lo studio sono stati presi in analisi i dati raccolti negli anni 2016, 2017 e 2018 dall'Ospedale M. Bufalini di Cesena. Queste informazioni comprendono 1130 pazienti gravide affluite presso l'Ospedale per espletamento del parto con travaglio condotto in analgesia e hanno reso possibile la realizzazione di una collezione di 1773 parti.

Sono stati esclusi i casi di morte endouterina fetale e di analgesia per travaglio di parto non condotti mediante analgesia epidurale.

Per analizzare il contesto ospedaliero, protagonista dello studio prospettico, è stato essenziale il colloquio con i membri del reparto di analgesia, grazie al quale è stato possibile far emergere le dinamiche riguardanti le procedure eseguite internamente per assistere le nascite e per la compilazione delle cartelle cliniche.

Assistenza al parto Le visite vengono effettuate dai medici nella stanza della paziente.

Differentemente dai parti senza antidolorifico, per effettuare la parto-analgesia è indispensabile il ruolo dell'anestesista, il quale, per rendere il parto meno doloroso alla partoriente, utilizza una o più tecniche a disposizione. I farmaci per l'anestesia vengono somministrati periodicamente fino al momento del parto. Ogni somministrazione è detta "bolo" e ognuna contiene una certa concentrazione di farmaco.

A causa della divisione dei reparti di analgesia e ostetricia su piani diversi dell'ospedale, i membri dell'equipe si devono scambiare informazioni sulle visite da effettuare tramite telefono.

Cartella clinica Durante la visita i dati della paziente sono stati raccolti in forma cartacea e inseriti nel database solamente in un secondo momento. Per

riportare i dati in forma elettronica è stato utilizzato, per l'intera durata della raccolta dei campioni, un applicativo che non utilizza vincoli ferrei su molti valori ma permette comunque la raccolta di dati in un formato strutturato.

A causa della trascrizione in due passaggi, spesso fatta da medici diversi, e della ridotta presenza di restrizioni sull'inserimento dei valori nel database, i dati sono intrinsecamente soggetti ad un alto rischio di errore umano.

4.1 Analisi

Partendo dalla struttura dei dati forniti dall'ospedale, è stato possibile fare un elenco dei parametri coinvolti nello studio e analizzarne i domini.

Il file contenente i dati, infatti, era già strutturato, il che ha reso possibile controllare i domini degli attributi e visualizzarne i valori tramite gli strumenti di **Weka** e **Scikit-learn**.
Il risultato dell'analisi iniziale delle caratteristiche, è riassunto nella tabella seguente.

Colonne nel database

Nome	Descrizione	Dominio
Nazionalità		0 Italiana, 1 Straniera
PesoAttuale		Kg
Altezza		cm
Età		anni
Primipara/Pluripara	Primipara se primo figlio Pluripara altrimenti	1. Primipara 2. Pluripara
Età gestazionale		Numero settimane
Gravidanza gemellare		0. No, 1. Si
PMA	Procreazione Medicalmente Assistita	0. No, 1. Si
Precesarizzata		0. No, 1. Si
Travaglio		1. Spontaneo 2. Indotto

Colonne nel database

Nome	Descrizione	Dominio
TIPO induzione	Se travaglio indotto	0. Nessuna, 1. Prostaglandine, 2. CRB, 3. Ossitocina, 4. Amniorexi
prostaglandine_ind	Farmaco	0. No, 1. Si
CRB_ind	Cervical Ripening Baloon	0. No, 1. Si
ossitocina_ind	Farmaco	0. No, 1. Si
amniorexi_ind	Farmaco	0. No, 1. Si
Inizio analgesia a travaglio avviato		0. No, 1. Si
Ossitocina	Farmaco, al di fuori dell'induzione	0. No, 1. Si
RotturaMembrane	Rottura delle membrane amniocoriali (Rottura delle acque)	1. Spontanea 2. Amniorexi
Tecnica	Per l'anestesia	1. Combinata 2. Peridurale
Puntura Durale Accidentale		0. No, 1. Si
Dilatazione Cervicale		cm
VAS	Dolore provato nel parto secondo la madre	[0,10]
durata_analgesia		minuti
Intervallo Primo-Secondo Bolo	Tempo tra i primi due boli, raggruppati per intervalli di tempo	1. 30-60 min 2. 61-90 min 3. 91-120 min 4. 121-180 min 5. $>$ 180 min
NumeroTotaleBoli	Iniezioni totali	> 1

Colonne nel database

Nome	Descrizione	Dominio
TipoDiParto		0. Vaginale 1. Ventosa 2. TC
MotivoTC		legenda a parte, non inserita nel documento
Intervallo ultimo bolo-TC	Tempo intercorso tra l'ultimo bolo e il taglio cesareo	minuti
TC CODICE ROSSO		0. No, 1. Si
Apgar1min	Vitalità neonato dopo un minuto di vita	[1,10]
Apgar5min	Vitalità neonato dopo 5 minuti di vita	[1,10]
Peso neonato		grammi
PPH	Emorragia post partum	0. $<$ 500 ml 1. 500-1000ml 2. 1000-2000 ml 3. $>$ 2000 ml
Esito Neonatale	UTIN se ricovero del neonato dopo il parto, Reparto altrimenti	1 Reparto, 2 UTIN
Patologie materne		1. Ipotiroidismo gravidico 2. Ipotiroidismo pregrav 3. Diabete gestazionale 4. Ipertensione/ Preclampsia
Ipotiroidismo	Pregravidico e gravidico insieme	0. No, 1. Si
Diabete		0. No, 1. Si
Ipertensione		0. No, 1. Si
Terapie materne		1. Levotiroxina 2. Insulina 3. Farmaci antipertensivi

Colonne nel database

Nome	Descrizione	Dominio
Farmaci Primo Bolo		1. Spinale 2. Spino-peridurale 3. Peridurale
FarmaciSecondoBolo	Percentuale di farmaco	1. $\leq$0.1% 2. $>$ 0.1%
Farmaci secondo bolo Volume	Quantità di farmaco	1. $\leq$ 10 ml 2. $>$ 10 ml

La variabile di **output** è **TipoDiParto**, la quale indica se un parto è di tipo vaginale naturale, vaginale ma con l'utilizzo della ventosa o è stato fatto un taglio cesareo. La classificazione più interessante per l'Ospedale era comunque tra i parti con e senza cesareo.

Analizzando le righe del database è stato possibile evidenziare che la distribuzione dei tipo di parto non era bilanciata, infatti sono stati eseguiti molti più **parti vaginali** (1352) rispetto ai **parti con ventosa** (112) e ai **parti cesarei** (309).

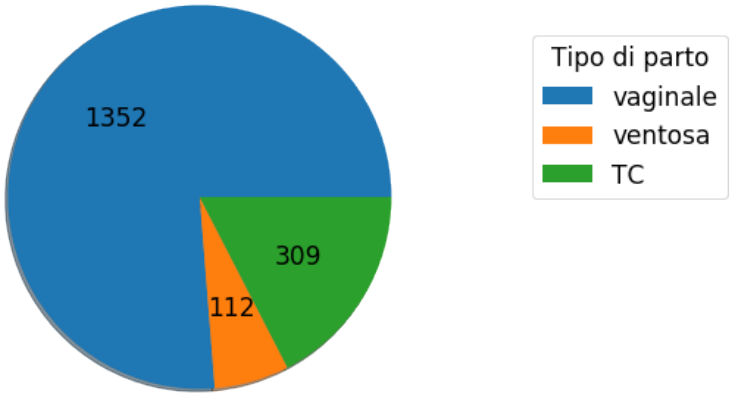


Figura 4.1: Distribuzione di tipi di parto

Tra i valori si è riscontrata anche la presenza di numerosi valori mancanti. Per una parte degli attributi il mancato inserimento è stato intenzionale, ad esempio nelle variabili legate ai boli, come "Intervallo Primo-Secondo Bolo", "FarmaciSecondoBolo" e "FarmaciSecondoBoloVolume", molti valori erano nulli a causa dell'assenza del secondo bolo, ossia della seconda iniezione di

farmaco analgesico. Per le altre variabili invece questa assenza è stata causata dalla mancata raccolta dell'informazione da parte dei medici o ad un errore nella registrazione del dato.

Non sono stati trovati numerosi dati con valori non coerenti, ma in quei pochi casi sono stati considerati come valori mancanti (come l'età gestazionale di 0 settimane o il peso del neonato di 0 grammi).

Un'ulteriore informazione fornita dai medici riguardava la presenza di variabili ottenute a posteriori e/o direttamente collegate al tipo di parto effettuato, le quali non dovevano essere considerate nell'analisi perché rilevanti solo per il tipo di parto a cui si riferivano. Queste variabili sono: "apgar1min", "apgar5min", "PPH", "MotivoTC", "intervallo ultimo bolo-TC", "TC CODICE ROSSO".

4.2 Pre-processamento

In seguito all'analisi, è stato possibile apportare le modifiche necessarie per permettere agli algoritmi di avere dati conformi alle proprie necessità. Per l'addestramento degli algoritmi si è scelto di utilizzare la libreria Scikit-learn, per cui il pre-processamento si è basato sulle necessità richieste di alcuni algoritmi di questa libreria.

Inizialmente è stato necessario modificare i tipi di variabili utilizzati nel dataset in quanto il tipo di dato avrebbe influenzato i calcoli effettuati dai modelli.
Il primo problema riscontrato riguardava i dati categorici multivalore memorizzati con valori numerici. In questo caso gli algoritmi, nel momento di effettuare operazioni sui dati, avrebbero trattato i dati come variabili continue rendendo impossibile la distinzione tra i valori.
Il secondo riguarda la necessità di gestire i valori nulli, a causa dell'impossibilità di molti degli algoritmi di non poter gestire valori mancanti.

Per gestire questi punti e le osservazioni emerse durante l'analisi, sono stati effettuati i seguenti passaggi:

- le variabili categoriche multivalore sono state divise, creando, in caso essa non fosse già presente, una colonna di tipo booleano per ogni valore categorico, in cui 1 indica la presenza del valore e 0 l'assenza. L'unica esclusione è stata per la variabile di output, TipoDiParto, che non è stata

modificata. In caso di due soli valori, si è lasciata una sola colonna e i valori sono stai diminuiti proporzionalmente per portarli a 0 e 1.

- sono state eliminate le *feature* indicate come non utili per lo studio: apgar1min, apgar5min, PPH, MotivoTC, intervallo ultimo bolo-TC, TC CODICE ROSSO.
- i missing sono stati trasformati nel valore più probabile o con un valore neutro. A parte in alcuni casi particolari, per le variabili continue si è usato il valore medio e per quelle categoriche la moda.
- a causa della difficoltà di gestire parametri di output multiclasse molto sbilanciati, per alcuni strumenti è stato necessario inglobare parti vaginali e parti con ventosa. Questi aspetti verranno spiegati nel dettaglio successivamente.

Modifiche principali

Nome	Precedente	Modifica	Dominio
Induzione	TIPO induzione	Da categorico a booleano	0. No, 1. Si
TerapieMaterne	Terapie materne	Da categorico a booleano	0. No, 1. Si
Levotiroxina	1. Levotiroxina	Da categoria a booleano	0. No, 1. Si
Insulina	2. Insulina	Da categoria a booleano	0. No, 1. Si
Alpa_metildopa	3. Farmaci antipertensivi	Da categoria a booleano	0. No, 1. Si
Farmaci PrimoBolo_1	1. Spinale	Da categoria a booleano	0. No, 1. Si
Farmaci PrimoBolo_2	2. Spino-peridurale	Da categoria a booleano	0. No, 1. Si
Farmaci PrimoBolo_3	3. Peridurale	Da categoria a booleano	0. No, 1. Si

4.3 Data Visualization

Prima di procedere con l'utilizzo di algoritmi che puntano alla predizione delle etichette, si è deciso di utilizzare strumenti di visualizzazione dei dati per capire la complessità del problema. Infatti, grazie alla rappresentazione su un grafico delle correlazioni dei dati, si possono avere informazioni sulla separazione di essi e sulla loro vicinanza nello spazio.

4.3.1 t-SNE

Si è deciso di utilizzare un approccio non supervisionato per rappresentare graficamente i dati in uno spazio bidimensionale e tridimensionale, in modo da analizzare se esiste un'evidente correlazioni tra i dati. L'approccio non supervisionato permette di trovare correlazioni senza specificare di ricercare quelle che meglio permettano di individuare il tipo di output da prevedere, a differenza di come invece viene fatto con le tecniche supervisionate.

Scaling I dati sono stati inizialmente scalati, per eliminare la differenza tra i domini delle variabili. Lo scaler utilizzato è stato lo standard scaler, una tecnica che sottrae ad ogni dato il valore medio di quella variabile e lo divide per il numero di tuple, permettendo di mappare i valori in base a come variano.

$$x_i^{'} = \frac{x_i - \bar{x}}{n}$$

Addestramento Sui dati scalati è stato poi addestrato il t-SNE, sul quale sono stati impostati diversi valori dell'iperparametro relativo al **numero di componenti** di interesse, inizialmente due per ottenere le due dimensioni x e y, successivamente tre per la rappresentazione sugli assi x, y e z.
Per ogni dimensione inoltre si sono testati diversi valori di **perplessità**, che è il numero di vicini che l'algoritmo tiene in considerazione per determinare i raggruppamenti dei dati.
Il dati sono stati visualizzati con colori diversi in base al tipo di parto: **verde** per i parti **vaginali**, **blu** per i parti con **ventosa** e **rosso** per i parti **cesarei** .

I risultati dopo questa prima operazione sono stati poco soddisfacenti. Anche moficando la perplessità, i tipi di parto diversi sono risultati sovrapposti e l'algoritmo ha generato un numero molto elevato di cluster, molto maggiore del numero di classi del nostro problema, senza evidenziare una forma particolare della distribuzione.

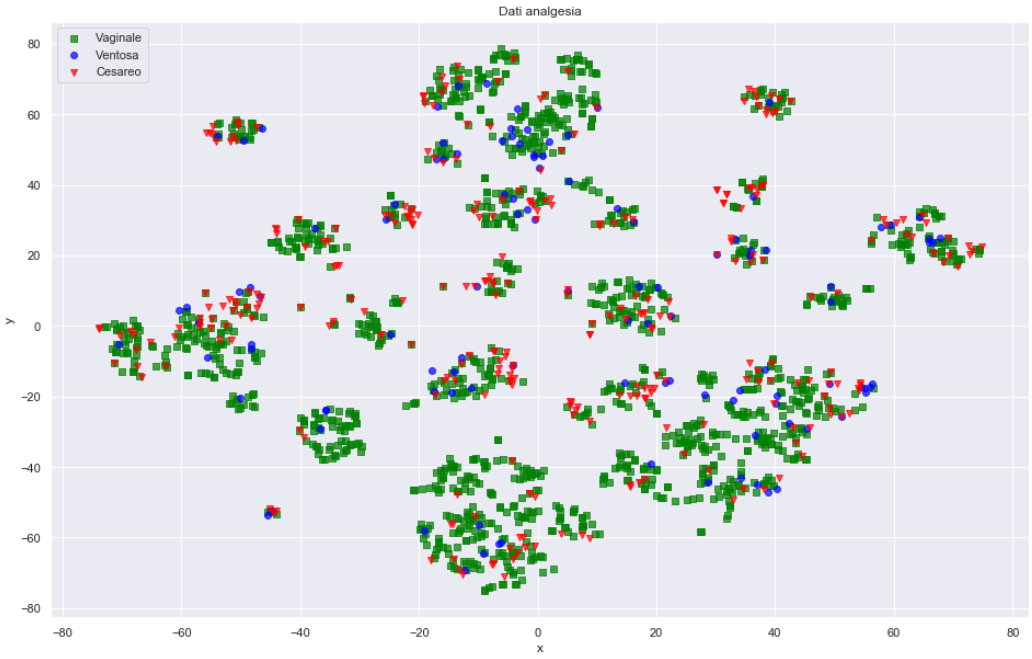


Figura 4.2: t-SNE in 2D con perplessità 15

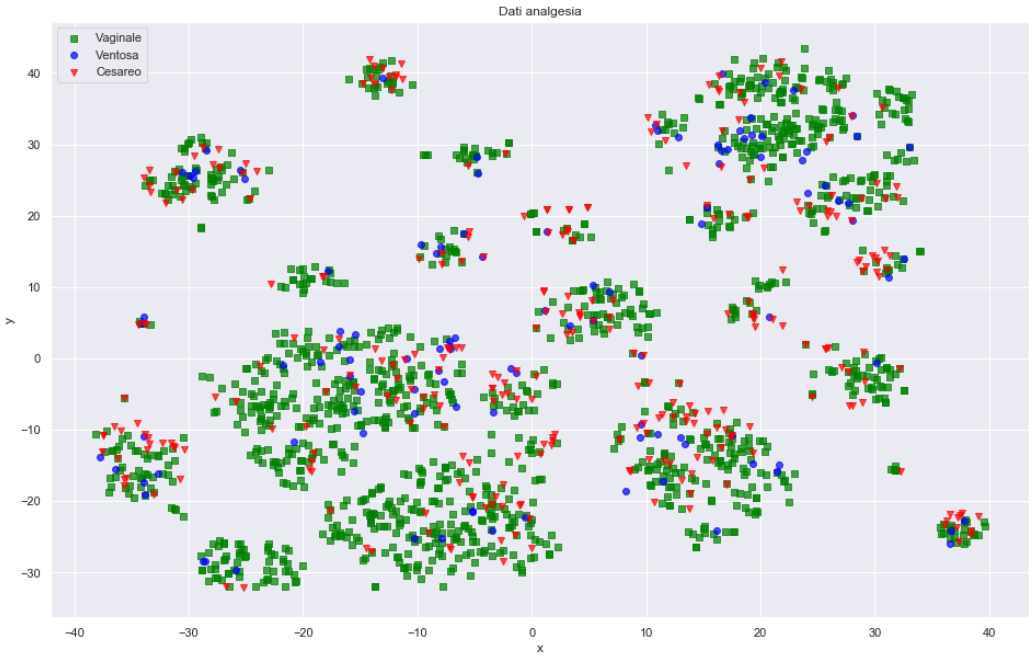
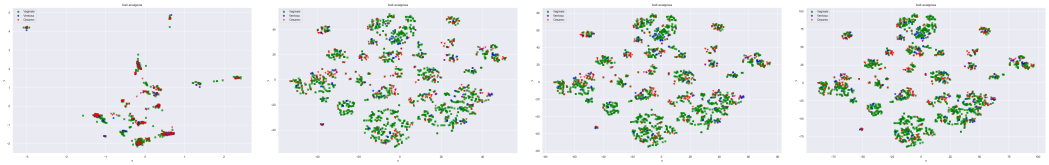


Figura 4.3: t-SNE in 2D con perplessità 50

Si è provato quindi a modificare, oltre alla perplessità, anche il **numero di iterazioni massime** effettuate dall'algoritmo per ottimizzare l'algoritmo, partendo da un minimo di 250. Gli addestramenti precedenti avevano un numero massimo di iterazioni impostato a 1000.

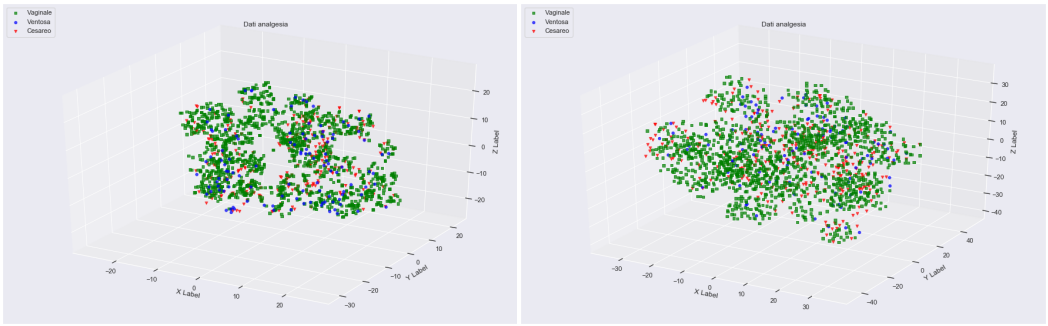
Anche questi risultati non hanno permesso di estrapolare informazioni visivamente utili.



(a) 250 iter. (b) 500 iter. (c) 1000 iter. (d) 2000 iter.

Figura 4.4: t-SNE in 2D, perplessità 15, variazioni del numero di iterazioni

Lo stesso è stato con i grafici tridimensionali.



(a) perplessità 15 (b) perplessità 50

Figura 4.5: t-SNE in 3D

Interpretazione del grafico I dati sono risultati sparsi e sovrapposti in molti punti. Questo mostra che il problema in questione non presenta evidenze dirette rispetto al Tipo Di Parto. Tuttavia questo risultato può essere legato all'influenza di alcune variabili non discriminanti, le quali aggiungono complessità al problema rendendo difficile la divisione dei dati.

Per permettere una visualizzazione dei dati maggiormente legata al tipo di parto, si è deciso di utilizzare un algoritmo di riduzione di dimensionalità supervisionato.

4.3.2 LDA

Linear Discriminant Analysis (LDA) è un algoritmo supervisionato che permette di ridurre le dimensionalità preservando maggiormente le informazioni relative alle classi di output. L'obiettivo è quello di massimizzare il più possibile la distanza tra le classi.

Grazie alla sua natura supervisionata questo tipo di algoritmo, oltre ad essere usato per ridurre le dimensionalità, può essere utilizzato anche come classificatore e possono essere applicate su di esso le metriche di valutazione dei classificatori.

Addestramento su tre classi

LDA è stato addestrato sui dati, scalati anche in questo caso con la tecnica dello Standard Scaler, e sono stati poi rappresentati graficamente per analizzarne la distribuzione, nello stesso modo di t-SNE.

Per questo algoritmo si è utilizzando l'80% dei dati come training set, che sono i punti rappresentati sul grafico, mentre sul restante 20% dei dati sono state calcolate le metriche per la valutazione delle prestazioni.

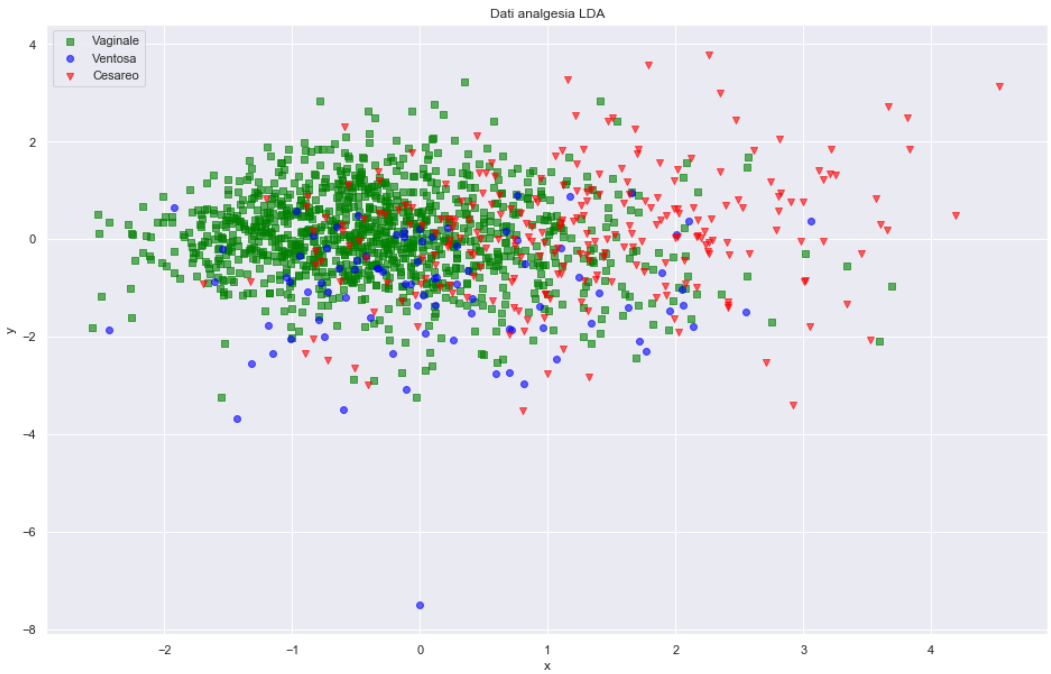


Figura 4.6: LDA su due dimensioni

Interpretazione del grafico Analizzando il primo grafico si è notato che nella parte sinistra del grafico, che corrisponde a valori piccoli sull'asse x, c'era una presenza maggiore di punti verdi, ossia di parti vaginali, mentre nella parte destra di punti rossi, corrispondenti ai parti cesarei. I parti operativi con ventosa non sembravano differenziarsi in maniera particolare.
Rispetto all'asse delle y invece non sembravano esserci evidenti differenze nella distribuzione dei punti.
Per confermare questa osservazione si sono rappresentate le coordinate x e y in due grafici diversi.

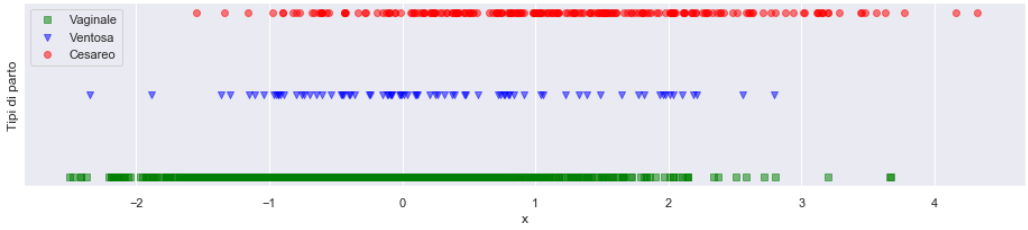


Figura 4.7: Asse x

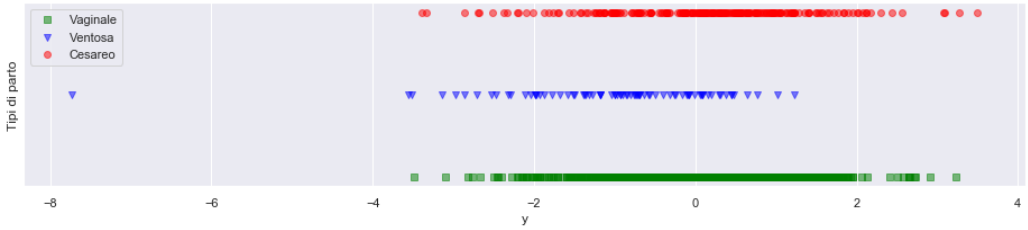


Figura 4.8: Asse y

Per capire quali attributi influenzassero maggiormente il risultato si sono considerati solamente i valori dell'asse delle x.
Questi valori corrispondono al valore ottimo dei coefficienti di ogni *feature*, calcolati durante l'addestramento dall'algoritmo, per ottenere la riduzione delle dimensionalità e la classificazione migliore.

Per l'estrazione dei valori si è analizzato il ridimensionamento dei pattern rispetto allo spazio esteso dai centroidi della classe.

Questi i coefficienti vengono utilizzati per trovare le coordinate rispetto agli n attributi secondo la seguente funzione:

$$x_i = \alpha \cdot a_{i1} + \beta \cdot a_{i2} + \cdots + \gamma \cdot a_{in}$$

Dove x_i è l'i-esimo punto, i coefficienti delle variabili indicano i pesi delle *feature* e a_i sono i valori scalati degli attributi per l'i-esimo record.

Sono stati quindi utilizzati i coefficienti della funzione per capire quali influenzino maggiormente il posizionamento dei punti sull'asse delle ascisse. Si è impostato 0.16 come valore soglia dei coefficienti in valore assoluto, sotto al quale le *feature* non sono state considerate perché troppo vicine alle 0.

Le più rilevanti sono state:

1. Durata_analgesia → 1.355
2. NumeroTotaleBoli → -0.883
3. InizioAnalgesiaATravaglioAvviato → -0.379
4. Precesarizzata → 0.300
5. Altezza → -0.274
6. GravidanzaGemellare → 0.227
7. Ipotiroidismo → -0.213
8. IntervalloPrimo_SecondoBolo → -0.201
9. Tecnica → 0.189
10. Età → 0.165

Le variabili non vanno considerate singolarmente per trovare i valori sulle x, tuttavia si nota che alcune di esse influenzano il risultato molto più di altre. L'ordine in cui sono state riportate è relatovo a quanto l'attributo influenza la x, mentre il segno indica se il valore è direttamente o indirettamente proporzionale alla x. Quelle con valori negativi indicano più influenza per il parto vaginale mentre quelli più positivi incidono di più su quello cesareo. Ad esempio se ad una durata dell'analgesia del parto (scalata) corrisponde un valore alto, la x aumenterà più facilmente in particolare di 1.355 ogni unità, portando la x verso destra. Se invece il numero di boli è molto alto, la x diminuirà.

Sebbene non possa essere fatta una distinzione precisa tra le classi, dal grafico della distribuzione della x possiamo vedere che per valori di x minori di -1 è più probabile che il parto sia naturale e per valori superiori a 2 invece l'incidenza predominante è di parti cesarei.

Valutazione dell'algoritmo L'accuratezza dell'algoritmo era del 78%, risultato non estremamente buono se si considera che idealmente un modello preciso dovrebbe avere un'accuratezza vicino al 100%. Questo valore più essere spiegato facilmente se si considera la sovrapposizione dei dati evidenziata dal grafico, la quale rende difficile trovare una retta lineare che possa dividere i dati.
Tuttavia questo dato però non è stato sufficiente per fare una valutazione di questo algoritmo in quanto i dati non sono bilanciati.
Per questo motivo sono stati considerati anche altri misuratori, a partire dalla matrice di confusione.

		Classe Attesa		
		Vaginali	Ventosa	Cesarei
Classe Predetta	Vaginali	249	3	18
	Ventosa	20	0	5
	Cesarei	40	1	19

Tabella 4.1: Matrice di Confusione di LDA su 2 dimensioni

Da questa è stata ricavata un'informazione importante sui parti con ventosa. Il basso numero di questi parti ha reso difficile il test su questa classe, infatti nel test set questi erano solamente 4. Il motivi non era legato ad un errore nel selezionare i set ma per la scarsa presenza di parti operativi con ventosa, infatti anche campionando diversamente i set si avevano circa lo stesso numero di record.
Si è notato inoltre che nessuno di essi è stato etichettato correttamente, mentre diversi parti vaginali sono stati identificati erroneamente come appartenenti a questa classe.

Le metriche calcolate sulla matrice di confusione hanno prodotto i seguenti risultati:

	Precision	Recall	F1-score
Vaginale	0.82	0.93	0.87
Ventosa	0.00	0.00	0.00
Cesareo	0.52	0.39	0.44
micro avg	0.78	0.78	0.78
macro avg	0.45	0.44	0.44
weighted avg	0.72	0.78	0.75

Tabella 4.2: Precisione, Recall e F1-score di LDA su 2 dimensioni

Gli ultimi tre valori nella tabella sono riferiti a modi diversi utilizzati per calcolare le medie delle metriche nei problemi multi-classe.
La "micro avg" viene calcolata considerando i TP di tutte le classi al numeratore e i TP insieme ai FP al denominatore; la "macro" calcola la media della metrica per ogni classe e fa la media su queste, non tenendo conto di un'eventuale sbilanciamento del dataset, infine "weighted avg" considera la media per ciascuna etichetta e trova la loro media ponderata rispetto al numero di TP.

Per i dataset sbilanciati è più significativa la metrica micro, infatti anche tra le metriche calcolate si può notare che la media macro abbia dei valori molto bassi. La causa risiede nei valori calcolati sulle previsioni dei parti con ventosa, che sono tutte uguali a 0 per l'assenza di TP nella classe. In questa metrica essi incidono per 1/3 sul risultato nonostante riguardino solamente 4 parti.

Il rapporto tra TPR e FPR, riferiti alla classe "Cesareo" è la seguente:

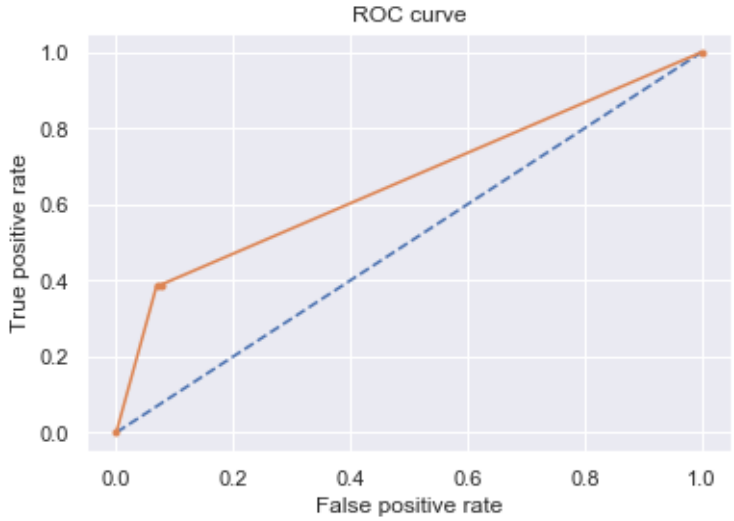


Figura 4.9: Curva ROC dei parti cesarei di LDA

La curva ROC evidenzia che i valori del dataset non hanno dato risultati molto buoni. L'AUC infatti è dello 0.66, molto lontana quindi dall'1, il valore di area ottimale. Si sono considerati i parti cesarei perché più significativi per il problema da analizzare.

I valori di queste metriche sono dovute alla troppa sovrapposizione dei valori.

Addestramento su due classi

Dato che l'interesse da parte dei medici era legato principalmente alla rilevazione dei parti cesarei e che nel grafico dell'LDA i parti con ventosa sono risultati scarsamente divisi e peggiorativi per l'analisi dei dati, si è deciso di inglobare parti vaginali e parti con ventosa per ridurre la complessità del problema e valutare se in questo modo si potesse avere un miglioramento delle previsioni sui parti con cesareo. L'unione dei due tipi di parto è stata identificata dalla classe **Parti Naturali**.

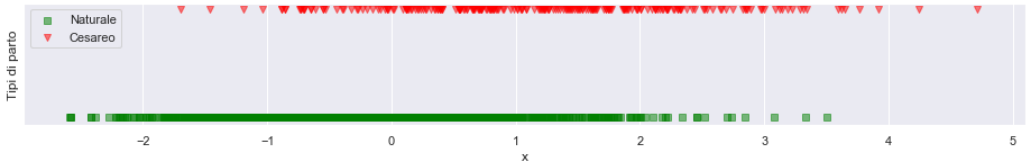


Figura 4.10: LDA su una dimensione

Il numero di dimensioni massime rappresentabile con LDA è pari al numero di classi meno uno, per cui nel grafico l'unica dimensione è l'asse x.
Come prima, i parti naturali e cesarei sono stati rappresentati ad altezze diverse per rendere più chiara la separazione dei valori e permettere di capire la distribuzione dei valori delle due classi senza avere sovrapposizioni.
Sono state estratte anche in questo caso le *feature* più rilevanti.

1. Durata_analgesia → 1.359
2. NumeroTotaleBoli → -0.990
3. InizioAnalgesiaATravaglioAvviato → -0.402
4. Altezza → -0.300
5. Precesarizzata → 0.281
6. GravidanzaGemellare → 0.248
7. IntervalloPrimo_SecondoBolo → -0.231
8. Travaglio → 0.193
9. Ipotiroidismo → -0.213

Possiamo notare che in entrambi i casi le *feature* che ricorrono, ad esclusione di "Travaglio", sono le stesse.

Valutazione dell'algoritmo Unendo le classi l'accuratezza dell'algoritmo è aumentata, arrivando all'86%.
Tuttavia anche in questo caso la matrice di confusione evidenzia che l'algoritmo circa in 1/3 dei casi non classifica in modo corretto i parti cesarei, che vengono attribuiti alla classe sbagliata.

		Classe Attesa	
		Naturale	Cesareo
Classe	Naturale	424	23
Predetta	Cesareo	54	31

Tabella 4.3: Confusion Matrix di LDA su 1 dimensione

	Precision	Recall	F1-score
Naturale	0.89	0.95	0.92
Cesareo	0.57	0.36	0.45
micro avg	0.86	0.86	0.86
macro avg	0.73	0.66	0.68
weighted avg	0.84	0.86	0.84

Tabella 4.4: Precisione, Recall e F1-score di LDA su 1 dimensione

La curva ROC e l'AUC sono rimaste praticamente invariate, infatti sebbene siano aumentate le tuple correttamente predette, sono aumentati anche i parti naturali classificati come cesarei.
La scelta di inglobare le classi di parti con ventosa e parti vaginali ha portato miglioramenti globali al modello, tuttavia questi non hanno inciso in maniera evidente sulla classe del cesareo.

4.4 Alberi decisionali

Per poter classificare i dati ma allo steso tempo avere una rappresentazione significativa del modello, si è deciso di utilizzare gli alberi decisionali. La peculiarità di questo approccio non riguarda solo la possibilità di rappresentare i nomi degli attributi che determinano la divisione in forma di albero, ma anche di poter trovare valori soglia che permettano di dividere le classi, aspetto non possibile con LDA.
Questo classificatore non è stato utilizzato con l'intento di prevedere le classi di nuovi record ma di determinare il percorso che porta alla classificazione in parti cesarei.

Definizione del modello

I criteri che sono stati presi considerazione per modellare l'albero decisionale sono stati:

- i criterio di "split", per trovare il nodo migliore che divida i dati ad ogni passo
- i criteri di "stop", che vanno scelti in modo da evitare l'overfitting o l'underfitting dell'algoritmo
- i criteri di valutazione del modello migliore

Criteri di Split L'albero di decisione considera un solo attributo per volta per dividere i record, per cui è importante conoscere quali criteri vengono utilizzati per sceglierlo in modo da comprendere il significato della classificazione. I criteri di split sono utilizzati dagli alberi decisionali per misurare l'impurità delle classi, ossia la concentrazione delle diverse classi nel nodo.
Per l'addestramento sono stati considerati due diversi misuratori di impurità: Gini ed Entropia.

L'impurità Gini calcola la probabilità che preso un elemento del dataset ed etichettato con la classe determinata dal nodo, la classificazione sia sbagliata. Questo valore è dato dalla formula:

$$Gini(i) = 1 - \sum_{j=1}^{k} [p(j|i)]^2$$

dove k è il numero di etichette di output e $p(j|i)$ è la probabilità che, dato l'i-esimo record, la classe risultante sia la j-esima.

L'Entropia invece è l'indicatore della disordine di un nodo. Questo valore rispetto ai dati è massimo, quindi uguale a 1, se i risultati hanno la stessa probabilità di accadimento. Valori bassi dell'entropia sono migliori per lo "split". La formula che la descrive è:

$$Entropy(i) = -\sum_{j=1}^{k} p(j|i) \log p(j|i)$$

Come sopra, $p(j|i)$ è la probabilità che per i-esimo record la classe di etichettatura sia la j-esima.

Criteri di Stop Per determinare quando smettere di dividere ulteriormente i nodi, i criteri che sono stati modificati sono:

- l'altezza massima dell'albero
- il numero minimo di tuple nel nodo per effettuare ulteriori divisioni
- il numero massimo di foglie nell'albero

Criteri di valutazione Per valutare l'ottimalità dell'albero il parametro utilizzato normalmente è l'accuratezza. Tuttavia è possibile utilizzare altri criteri nel caso in cui questo valore non fosse sufficientemente indicativo della bontà della classificazione. Infatti, come detto precedentemente, l'accuratezza non è un buon indicatore nel caso di dataset sbilanciati.

Addestramento

Per l'addestramento sull'albero si è utilizzato lo strumento Grid Search Cross Validation, il quale permette di testare iperparametri diversi sull'algoritmo e scegliere quello che produce risultati migliori in termini di ottimalità dell'algoritmo. L'implementazione utilizzata sfrutta la k-fold Cross Validation per scegliere i valori migliori, nel nostro caso con 3 fold.
I set di training e di test utilizzati sono stati gli stessi dell'addestramento su LDA, divisi con la tecnica Holdout 80/20.

Senza restrizioni Inizialmente l'albero è stato addestrato non ponendo vincoli sull'altezza o sul numero di nodi. Si sono utilizzati entrambi i criteri di "split".

L'accuratezza media di questi due addestramenti è stata del 76%, più bassa rispetto ai valori ottenuti precedentemente su due classi con LDA.

Tuttavia l'informazione più importante la si è ottenuta dal grafico è stata l'eccessiva complessità della soluzione, dovuta ad un problema di Overfitting. Inoltre l'F1-measure relativo al taglio cesareo era del 32%, con un'AUC di 0.59.

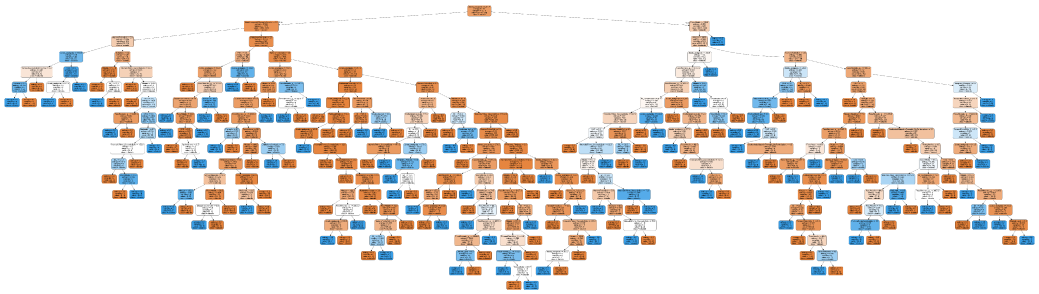


Figura 4.11: Decision Tree overfitted

Vincoli di dimensione Sono quindi stati impostati parametri per determinare quale altezza e quale numero di nodi fornisse una soluzione migliore per descrivere il problema, senza overfitting e che lo rendesse interpretabile. Questo passaggio di sfoltimento dell'albero è detto Pruning.

Sebbene questa soluzione sia risultata più facile da leggere, la semplificazione ha reso l'albero meno preciso. Anche se l'accuratezza del modello migliore sia risultata del 83%, l'F1-measure dei parti cesarei è il 30% e l'AUC è 0.58. Sono migliorate quindi le previsioni dei parti naturali ma peggiorate quelle dei parti cesarei.

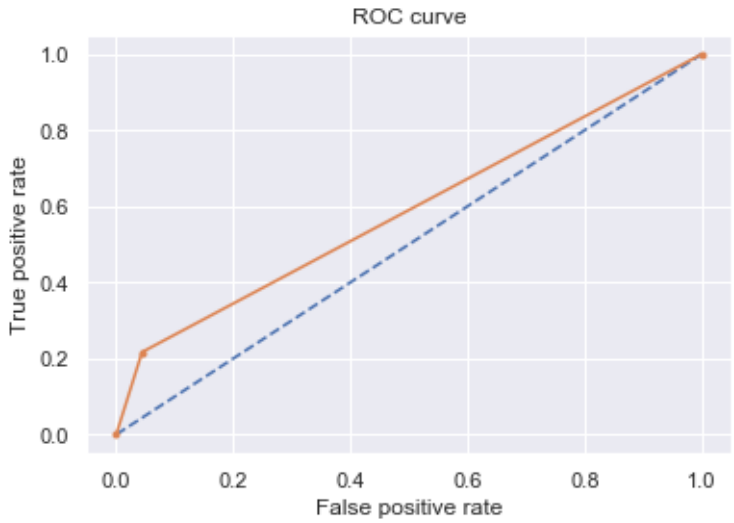


Figura 4.12: ROC del decision tree limitato in altezza

Si sono voluti apportare ulteriori miglioramenti prima di procedere all'interpretazione dell'albero e porre maggiore peso ai parti cesarei.

Metriche di valutazione Utilizzando l'accuratezza per determinare quale algoritmo sia migliore non abbiamo ottenuto delle prestazioni sufficientemente buone rispetto al problema in esame.
Si è deciso quindi di utilizzare il valore della F1-measure relativo al taglio cesareo.

Contrariamente a quanto ci si aspettava, dopo l'addestramento ottenuto cambiando la metrica, le prestazioni dell'algoritmo sono rimaste invariate. La metrica non ha quindi portato a un miglioramento delle prestazioni ma ha confermato la struttura dell'albero trovato precedentemente.

Analizzando l'immagine relativa al grafico è però stato notato che lo sbilanciamento dell'albero è evidente.
Nel grafico i nodi prevalentemente in arancio indicano la maggior presenza di tuple corrispondenti ai parti naturali, i nodi neutri sono equamente distribuiti mentre quelli in azzurro hanno maggior concentrazione di parti cesarei.

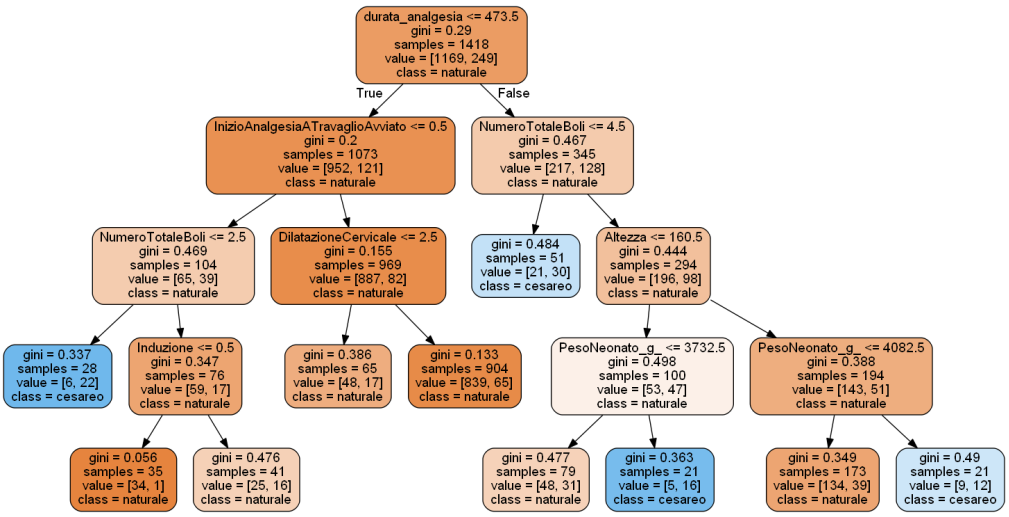


Figura 4.13: Decision Tree sbilanciato

Per cercare di migliorare le prestazioni si è deciso di modificare il bilanciamento dell'albero, al fine di porre più attenzione sui parti cesarei.

Bilanciamento dell'albero Facendo un proporzione è stato calcolato che i parti naturali sono 4.5 volte maggiori rispetto ai parti cesarei.
Per bilanciare l'albero si è sfruttato un parametro che rende possibile attribuire pesi diversi alle classi, aumentando nel nostro caso il peso dell'etichetta relativa ai parti cesarei per un valore di 4.5 volte.

Avendo utilizzato un insieme di dati bilanciati, per questo addestramento si è deciso di utilizzare sia l'accuratezza che l'F1-measure, oltre ai vari parametri di split e di stop.

I parametri scelti dagli addestramenti con le diverse metriche sono stati gli stessi, quindi l'albero ricavato è stato il medesimo.
Questa informazione indica che l'albero in questione è il migliore tra quelli ottenuti col bilanciamento, sia per quanto riguarda l'accuratezza che l'F1-measure relativa al taglio cesareo.

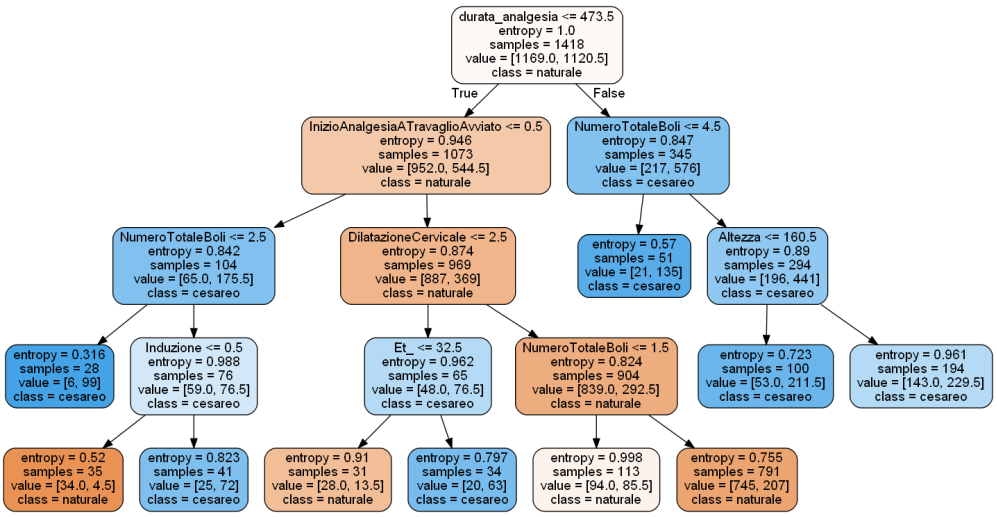


Figura 4.14: Decision Tree bilanciato

Questo albero aveva un'accuratezza del 69%, più bassa degli altri addestramenti ma più veritiera, infatti il valore più alto calcolato negli alberi sopra era dovuto allo sbilanciamento dei dati. Sono state quindi valutate le altre metriche.

	Precision	Recall	F1-score
Naturale	0.89	0.72	0.80
Cesareo	0.29	0.57	0.39
micro avg	0.70	0.70	0.70
macro avg	0.59	0.64	0.59
weighted avg	0.79	0.70	0.73

Tabella 4.5: Precisione, Recall e F1-score del Decision Tree

Il miglioramento dell'F1-measure è stato dello 0.09%, l'AUC di conseguenza è passata a 0.64.

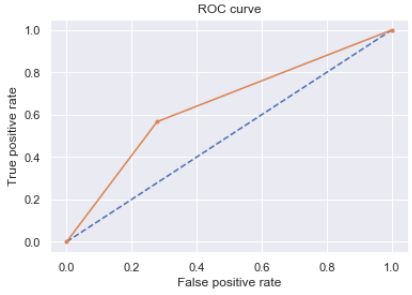


Figura 4.15: ROC del decision tree limitato in altezza e bilanciato

Avendo effettuato miglioramenti a tutti i parametri, si è potuta ritenere sufficiente la fase di addestramento per passare alla lettura e all'interpretazione dell'albero.

Interpretazione dei grafici

Per leggere un albero non si può considerare ogni nodo a sé stante. Partendo dalla radice bisogna seguire il percorso fino ad una foglia, in modo da avere tutte le informazioni necessarie per ricostruire la logica di divisione.
È possibile seguire anche il percorso inverso, quindi partendo dalle foglie per ripercorrere gli archi fino alla cima dell'albero.

Analizzando i percorsi dei due alberi, quello bilanciato e quello non bilanciato, è stato possibile estrarre i percorsi, molti dei quali sono risultati essere comuni ad entrambi gli alberi.

Leggendo l'albero partendo dalla radice fino alle foglie, sono emerse le seguenti osservazioni:

1. Se la durata dell'analgesia è inferiore a 7 ore e 53 minuti e l'analgesia è iniziata prima dell'inizio del travaglio, allora un numero di boli inferiore a 3 porta al cesareo.

2. Se la durata dell'analgesia è inferiore a 7 ore e 53 minuti e l'analgesia è iniziata prima dell'inizio del travaglio ma il numero di boli è superiore a 3, si ha meno probabilità di cesareo se non c'è stata induzione nel parto.

3. Se la durata dell'analgesia è inferiore a 7 ore e 53 minuti e l'analgesia è iniziata dopo il travaglio, è più probabile che il parto si concluda come parto naturale, soprattutto se la dilatazione cervicale è maggiore di 2.5cm.

4. Sui parti con analgesia di durata maggiore alle 7 ore e 53 minuti incide molto il numero di boli. Se i boli sono 4 o meno c'è più probabilità di subire un cesareo.

5. Sui parti con analgesia di durata maggiore alle 7 ore e 53 minuti e con almeno 5 boli, l'altezza della madre influenza la nascita. La soglia calcolata è 160cm, sotto i quali il rischio di cesario aumenta.

Estrazione delle *feature*

La libreria utilizzata per addestrare gli alberi decisionali mette a disposizione strumenti di estrazione delle *feature*, indicando quindi quali di esse abbiano impattato maggiormente sullo split dei nodi.

Dai modelli è stato possibile estrarre le seguenti:

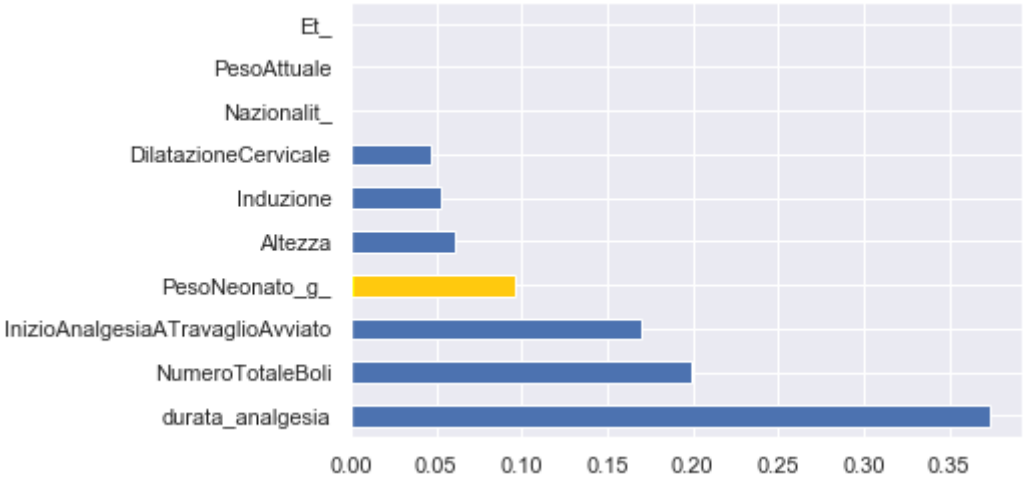


Figura 4.16: Importanza delle feature, albero non bilanciato

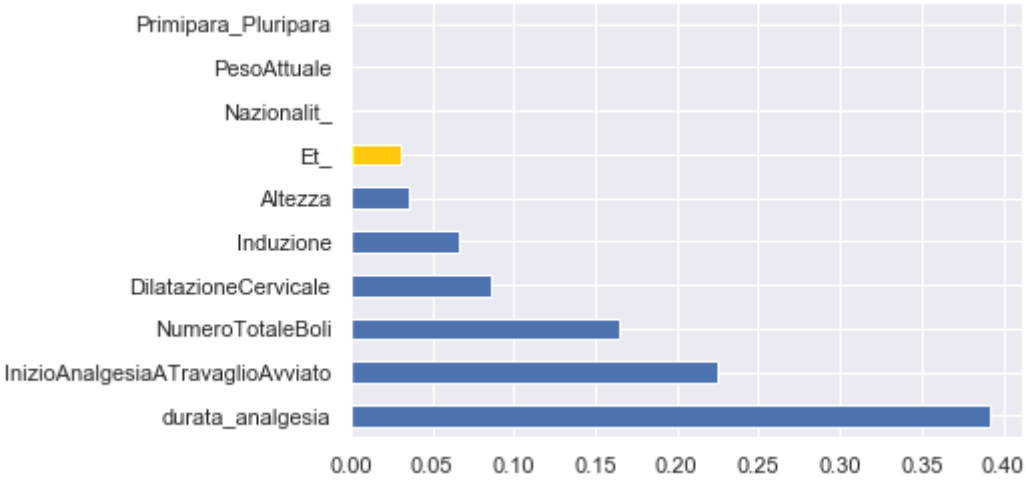


Figura 4.17: Importanza delle feature, albero bilanciato

Le barre blu nei grafici sono relative agli attributi che si ripetono in entrambe le estrazioni. Viceversa, le barre gialle non compaiono in entrambe.

Conclusioni

Il progetto preso in esame per il lavoro di tesi aveva l'obiettivo di estrarre informazioni da dati relativi a parti con utilizzo di analgesia spino-peridurale, tramite l'utilizzo di tecniche di Machine Learning, per capire quali fattori concorrono al ricorso del parto cesareo.
Per affrontare questo problema è stato prima necessario individuare i passaggi fondamentali da seguire quando ci si approccia ad un problema di Machine Learning. I passaggi evidenziati sono stati: l'analisi, il pre-processamento dei dati, la scelta e l'addestramento degli algoritmi da utilizzare e la valutazione dei modelli estratti.

Il caso di studio aveva come obiettivo l'analisi dei dati raccolti dal gruppo che si occupa di ostetricia nell'U.O. di Anestesia e Rianimazione dell'Ospedale M. Bufalini di Cesena nel corso degli anni 2016-2017-2018. Il fine dell'analisi è stato quello di identificare possibili fattori che concorrono all'aumento del rischio di TC.
Ogni record del dataset che ci è stato fornito corrisponde a una paziente e contiene un insieme di *feature* e la *label* della classe (Parto Vaginale, Operativo o Taglio Cesareo). La distribuzione dei dati forniti era di 1352 parti vaginali, 112 parti operativi e 309 tagli cesarei. Per poter interpretare nel modo corretto le *feature* di ciascun *record* è stato essenziale il ruolo dei medici, i quali, grazie alle informazioni fornite durante gli incontri, hanno reso possibile la comprensione dettagliata degli aspetti clinici.

Una volta in possesso di tutti i dettagli sul dominio in questione, è stato possibile adattare i dati alle tecnologie scelte, convertendoli in modo opportuno rispetto ai requisiti dell'algoritmo scelto.

La scelta degli algoritmi è fatta sulla base degli obiettivi individuati, che si è concentrata sull'estrazione delle caratteristiche più rilevanti. Al fine di spiegare quali fattori incidessero maggiormente sull'espletamento con ricorso al taglio cesareo non programmato, il problema è stato affrontato da punti di vista diversi.

Per prima cosa è stata effettuata una riduzione delle dimensionalità per esaminare graficamente la loro distribuzione nello spazio e capire se i dati

fossero facilmente correlabili o meno.
È stato utilizzato inizialmente il t-SNE, un approccio non supervisionato, in cui all'algoritmo non sono stati forniti i valori di output. I risultati non sono stati significativi. Dalla visualizzazione grafica sono state ottenute nubi sparse di dati e non organizzate rispetto alla classe di appartenenza, impedendo quindi di trovare una divisione netta tra i tipi di parto.

Si è testato quindi un algoritmo di riduzione delle dimensionalità supervisionato, in modo da preservare maggiormente le informazioni legate ai risultati attesi. I grafici otteniti con LDA sono stati più informativi, infatti si è individuata una dimensione parzialmente discriminante, rispetto alla quale i dati sono risultati maggiormente divisi.
Questa distinzione ha reso possibile estrapolare le *feature* più influenti. È stato individuato che valori alti della durata dell'analgesia aumentassero la probabilità di parto cesareo, e in maniera minore anche la precesarizzazione della paziente. Al contrario, la probabilità era maggiore per il parto vaginale al crescere del numero di boli e secondariamente anche in presenza di parti in cui l'analgesia era fatta a travaglio già avviato e per valori maggiori dell'altezza della madre.

Addestrando questo modello è stato evidente che lo sbilanciamento dei dati (molti parti vaginali, meno cesarei e pochissimi con ventosa) rendesse difficile, se non quasi impossibile, la previsione dei parti con ventosa. Questo è stato dovuto allo sbilanciamento ma anche alla distribuzione di questi dati (LDA li ha collocati nella fascia intermedia del grafico, completamente sovrapposti con le altre classi). Vista la volontà dei dottori di scoprire le cause del cesareo e non dei parti con ventosa, si è deciso di inglobare parti vaginali e con ventosa in un'unica classe.

Su queste due classi di output si è applicato anche il Decision Tree, un classificatore supervisionato che, oltre a cercare una divisione tra i tipi di parto, permette di evidenziare i valori soglia di queste variabili e connetterne logicamente anche più di una.
Anche in questo caso sono emerse come *feature* principale quelle estratte con LDA, dando però un ordine di lettura migliore a queste (come evidenziato nel capitolo 4.4, nella sottosezione dedicata all'interpretazione del grafico).

In conclusione, lo studio ha permesso di trovare caratteristiche dei parti che effettivamente aumentano l'incidenza dei parti cesarei. Su alcune di queste non è possibile intervenire, come la durata dell'analgesia, l'altezza o i parti cesarei subiti nel passato. Altre invece dipendono maggiormente dall'intervento medico. Da quanto è emerso, un numero maggiore di boli e l'inizio dell'analgesia dopo il travaglio potrebbero diminuire la necessità di ricorrere al parto cesareo urgente.

Ringraziamenti

Per questa tesi desidero ringraziare il Prof. Davide Maltoni le cui conoscenze sono state indispensabili per questo lavoro e la Dottoressa Sara Montagna per l'attenzione rivoltami durante questi mesi e per l'estrema disponibilità. A quest'ultima faccio anche i più sentiti auguri per la futura nascita del figlio.

Un ringraziamento particolare va alla mia famiglia che nonostante non nutra lo stesso mio amore per lo studio, mi ha permesso di affrontare questo percorso nel migliore dei modi. In particolare a mia mamma, per avermi insegnato a sorridere in ogni momento, a mio padre, per la trasparenza con cui mi ha fatto sempre vedere il mondo e a mio fratello, per non essersi accontentato.

Infine il ringraziamento più grande va a Christian che ogni giorno, da più di due anni, ispirandomi mi spinge a migliorare e senza il quale probabilmente la mia strada sarebbe stata completamente diversa.

Bibliografia

[1] *Ue lancia piano intelligenza artificiale da 20 miliardi l'anno*, ANSA, 7 Dicembre 2018.

[2] Jacques Bughin, Eric Hazan, Sree Ramaswamy, Michael Chui, Tera Allas, Peter Dahlström, Nicolaus Henke, Monica Trench, *Artificial intelligence: The next digital frontier?*, McKinsey&Company, 2017.

[3] Eric J. Topol, *High-performance medicine: the convergence of human and artificial intelligence*, Nature Medicine, 2019.

[4] H. Kozima, C. Nakagawa, Y. Yasuda, *Interactive robots for communication-care: a case-study in autism therapy*, 2005 IEEE International Workshop on Robots and Human Interactive Communication, 2005.

[5] Google Health, *www.ai.googleblog.com/2018/10/applying-deep-learning-to-metastatic.html*

[6] Intuitive Surgical, *www.intuitive.com*

[7] LeanTaaS Inc., *www.leantaas.com/?press-release=uchealth-improves-operating-room-efficiency-lowering-wait-times-and-improving-patient-experience-with-leantaas*

[8] Pavel Hamet, Johanne Tremblay, *Artificial intelligence in medicine*, Metabolism, 2017.

[9] *Uber non perseguibile per incidente in Arizona*, ANSA, 6 Marzo 2019.

[10] Rapporto CeDAP, *La nascita in Emilia-Romagna. 15° Rapporto sui dati del Certificato di Assistenza al Parto. Dati anno 2017*, Sistema Informativo Politiche per la Salute e Politiche Sociali, 2018.

[11] Valentina Arcovio, *L'Oms richiama l'Italia «Troppi parti cesarei»*, Il Messaggero, 25 Febbraio 2018.

[12] Osservatorio Nazionale sulla Salute nelle Regioni Italiane, *Rapporto Osservasalute 2018 - La sanità italiana nel confronto europeo*, 2019.