

UNIVERSITÀ DEGLI STUDI DEL SANNIO

Dipartimento Di Ingegneria

Corso di Laurea in Ingegneria Informatica

Analisi e validazione di algoritmi di classificazione su un dataset di pazienti Parkinsoniani

RELATORE:
Prof.ssa **Lerina Aversano**

CANDIDATO:
Federico Stocchetti
Matr:863001333

ANNO ACCADEMICO 2020/2021

Indice

Elenco delle figure	ii
Elenco delle tabelle	iii
Introduzione	iv
1 Contesto e problem statement	1
1.1 Informatica medica	1
1.2 Data mining	3
1.3 Machine Learning	5
1.4 Parkinson	7
1.5 Problem Statement e stato dell'arte	8
2 Setup e metodologia di lavoro	11
2.1 Dataset utilizzato	11
2.2 Pre elaborazione dei dati	13
3 Approccio al problema	14
3.1 Linguaggio e librerie	14
3.1.1 Librerie di Python utilizzate	15
3.2 Preparazione dei dati	15
3.3 Bilanciamento del DataFrame	17
3.4 Feature selection	18
3.5 Scelta dei classificatori	22
3.6 Parametri di classificazione	22
3.7 [OPZ] Ensemble di classificatori	22
4 Analisi dei risultati	23
4.1 Recap	23
4.2 Risultati utilizzando tutte le feature	23
4.3 Risultati applicando feature selection a varie threshold	23
4.4 Risultati applicando anche bilanciamento	23
4.5 Eventuali altri risultati: Ensemble, normalizzazione e discretizzazione, altri algoritmi di feature selection	23

INDICE	i
5 Conclusione e sviluppi futuri	24
Bibliografia	25

Elenco delle figure

1.1	Sorgenti dei dati utilizzati	10
3.1	Utilizzo di SMOTE a partire dalla libreria imblearn	18
3.2	Esempio di dendrogramma	20
3.3	Script per la generazione e il taglio del dendrogramma	21
3.4	Selezione di una feature dai cluster di features	21

Elenco delle tabelle

3.1	Feature ricavate alle diverse soglie di correlazione.	22
-----	---	----

Introduzione

Nell'era moderna il progresso tecnologico nel settore dell'informatica coinvolge a cascata il progresso in molti settori adiacenti. In particolare nel settore medico gli avanzamenti tecnologici hanno reso possibili studi, diagnosi, operazioni chirurgiche e trattamenti di efficacia e affidabilità di gran lunga superiori al passato. Nel campo della ricerca medica si sta sempre di più volgendo l'attenzione verso i Big Data sanitari per sviluppare nuove tecniche di medicina di precisione, **da continuare** In questo contesto diviene sempre più importante il processo di Knowledge Discovery, o Data Mining, che permette, a partire da un insieme di dati coerenti, di estrarre contenuto informativo **da continuare**

**** Descrivere brevemente il contenuto della tesi ****

Capitolo 1

Contesto e problem statement

1.1 Informatica medica

"Frase iniziale" - Autore

L'informatica medica è un campo molto ampio, che origina dalla gestione del flusso informativo proveniente da un paziente o cittadino, ma che si è evoluta nel tempo in una disciplina che è essenziale alle attività mediche di ricerca, trattamento del paziente e gestione dei processi medici.

I primi riferimenti all'informatica medica si hanno nella seconda metà del secolo scorso, momento in cui le discipline informatiche iniziano a contaminare vari ambiti per il supporto alla gestione dell'informazione. Qui si originano i primi processi per il supporto al mantenimento dei dati dei pazienti e al loro successivo utilizzo.

Negli anni successivi con l'avanzare dei progressi nel campo informatico si è iniziato a sviluppare i primi strumenti di supporto alle attività mediche come primi software per l'imaging tramite radiografia, strumenti per il design di protesi e software per la ricerca nella bibliografia medica, fino ad arrivare alla diffusione di internet su larga scala che ha permesso di connettere le "isole" di dati medici e di aumentarne così la fruibilità.

Oggi l'informatica medica è un campo di studi ancora in costante evoluzione, dai confini non ben definiti, i cui oggetti di studio non sono più limitati alla raccolta dei dati dei pazienti o al supporto gestionale dei processi medici, ma si espandono nel campo della ricerca farmacologica, della diagnosi assistita da processi automatici e della ricerca medica in generale.

Attualmente le applicazioni notevoli dell'informatica medica sono molteplici e si sviluppano trasversalmente in vari sottocampi:

- CRI o Informatica per la ricerca clinica: ambito dell'informatica medica che mira a migliorare l'efficienza della ricerca clinica tramite l'utilizzo di metodi informatici, ad esempio tramite il supporto alla data collection per gli studi clinici o la creazione di pipeline per la ricerca data driven.
 - Bioinformatica : campo emergente dell'informatica medica in cui convergono informatica, statistica e biologia. Si basa sull'applicazione di tecniche informatiche alla sempre crescente quantità di dati biomedici e genetici per estrarre conoscenza e creare strumenti ad uso di personale medico ma anche pazienti. Vengono considerate popolazioni di pazienti, materiale genetico e dati sulle malattie per applicazioni come la diagnosi assistita, il miglioramento di terapie farmacologiche o il calcolo dei fattori di rischio per l'insorgenza di malattie.
 - Intelligenza Artificiale per la medicina: questo campo si concentra sull'utilizzo di intelligenza artificiale per migliorare i processi medici di diagnosi, trattamento del paziente e sviluppo farmaceutico. La ricerca in questo campo oggi si concentra principalmente sulla creazione di un sistema di supporto alle decisioni cliniche (CDSS) tramite l'utilizzo dei big data sanitari, sulla medicina personalizzata e sulla interpretazione di segnali biologici complessi, come il funzionamento del cervello.
Per via del forte riscontro che si ha applicando metodi di IA alla medicina questo è anche il settore dell'informatica medica che riceve più investimenti privati per la ricerca, con stakeholder del calibro di IBM, Microsoft, Google e Neuralink che competono per sviluppare soluzioni innovative per il mercato medico. In particolare Neuralink è impegnata nella creazione di una interfaccia cervello-macchina con l'obiettivo di un controllo diretto di protesi robotiche come braccia, occhi o orecchie.
 - Telemedicina: questo campo studia le tecniche per il controllo di pazienti a distanza o per permettere di svolgere processi medici di diagnosi e chirurgia in remoto. Questo settore di studi ricopre una importanza sempre crescente per il tentativo di disaccoppiare il luogo in cui si ricevono cure dal personale medico
-

che lo esegue, aprendo nuove possibilità nei paradigmi di cura e monitoraggio del paziente.

- Informatica per l'imaging medico: si occupa del processing delle immagini mediche e della creazione di modelli 3D partendo da rilevazioni effettuate tramite strumenti, allo scopo di assistere il personale medico nelle operazioni di diagnosi e chirurgia.

Rispetto ai primi anni, quindi, cambiano profondamente alcuni paradigmi di sviluppo, si passa dal conservare e distribuire in maniera ottimale i dati ad elaborarli per estrarre conoscenza, dal fornire supporto ai processi medici alla loro automatizzazione e innovazione.

1.2 Data mining

Il settore medico, sulla linea di un trend che accomuna molti altri settori, è sommerso dall'enorme quantità di dati prodotti che ammonta nel 2018 a circa 1218 EB e di cui si proietta una crescita del 36% fino al 2025 [2]. Per sfruttare a pieno le potenzialità che una così grande mole di dati offre per l'ambito medico si usano sempre più frequentemente tecniche di knowledge discovery o data mining, supportate da tecniche di machine learning supervisionato o non supervisionato.

"Data mining, also popularly referred to as knowledge discovery from data (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories or data streams" [3]

Le tecniche di data mining vengono quindi utilizzate per estrarre conoscenza, spesso in maniera automatica, dalle grandi quantità di dati presenti nei database di ospedali e cliniche o provenienti da ricerche mediche su media e larga scala.

Le più utilizzate sono, generalmente:

- Clustering: raggruppamento di oggetti o caratteristiche di un insieme in più sottoinsiemi di elementi coerenti tra di loro (clusters). Gli elementi di un cluster avranno caratteristiche in comune secondo le metriche prese in considerazione,
-

ma al contempo ogni cluster avrà elementi dissimili dagli elementi degli altri cluster

- **Classificazione:** assegnazione di una classe di appartenenza ad elementi di un insieme. Gli elementi di un insieme A vengono etichettati di classe C se possiedono, entro una certa varianza, le caratteristiche che definiscono la classe C. Le caratteristiche della classe possono essere ricavate per inferenza da un insieme di dati già classificati, spesso tramite metodi di machine learning.
- **Regressione:** stima della relazione tra una variabile dipendente, spesso l'output di un sistema, e una serie di variabili indipendenti. Il risultato di questo processo è un modello di regressione, capace di indicare il valore della variabile dipendente a partire da variabili indipendenti.
- **Dependency modelling:** individuazione di regole di associazione, regole logiche basate sull'analisi degli attributi degli elementi di un insieme secondo una serie di *if-then* statements. L'output di questo processo è una serie di regole di associazione per cui, dati due elementi X e Y si può asserire che X implica Y.
- **Anomaly detection:** individuazione degli outliers in un insieme di dati, ovvero di elementi che sono molto diversi dagli altri dello stesso insieme.

A prescindere dalla tecnica utilizzata, il processo di data mining può essere generalizzato nelle seguenti fasi di lavoro:

- **Data collection and integration:** raccolta dei dati rilevanti per l'analisi da eseguire dalle varie fonti e integrazione dei dati in un unico insieme di dati;
 - **Data selection and reduction:** riduzione della dimensionalità dei dati e selezione dei dati da utilizzare rispetto ai dati raccolti. Questa fase, che può sembrare banale, è di vitale importanza poiché contribuisce in maniera importante alla qualità dell'analisi che andremo ad effettuare, secondo il principio di *"garbage in garbage out"*.
-

- Data cleaning and preprocessing: rimozione delle inconsistenze nei dati, delle discrepanze dei valori e "pulizia" dei dati da valori errati. Questa fase prosegue il lavoro della fase precedente di assicurare una analisi su dati di qualità.
- Data transformation and feature reduction: in questa fase vengono selezionate le feature rilevanti dei dati da analizzare, poiché un numero minore di features riduce i tempi di elaborazione, la complessità dell'analisi e permette di produrre analisi più generali. I dati vengono poi trasformati tramite encoding e type conversion per essere accettati come input dell'algoritmo di analisi. In questa fase si possono operare anche discretizzazioni sui dati o normalizzazione in un intervallo ben definito di valori.
- Data mining: questa è la fase di lavoro in cui viene estratta la conoscenza dai dati. E' essenziale scegliere l'algoritmo più adatto ai risultati che vogliamo ottenere, il formato dell'output e i parametri dell'algoritmo di analisi.
- Evaluation and validation: in questa fase avviene una review dei risultati dell'analisi per verificarne la correttezza e la validità. Potrebbe infatti verificarsi la situazione in cui l'analisi effettuata non abbia rilevanza statistica o semplicemente che sia una pessima analisi. Ad esempio potremmo avere un modello di classificazione che produce una accuratezza peggiore di una scelta casuale.

Nonostante questo sia considerato il workflow standard per il data mining esistono alcune variazioni che sono specifiche al campo di applicazione, come ad esempio il Cross Industry Standard Process for Data Mining (CRISP-DM), ma i *core elements* rimangono gli stessi. In questo lavoro di tesi verrà seguito il workflow prima descritto, con particolare enfasi sulla feature selection, essendo la dimensionalità degli attributi del dataset preso in considerazione molto alta.

1.3 Machine Learning

Il machine learning è una tecnica di apprendimento automatico utilizzata in molti domini operativi per risolvere problemi classici di predizione, classificazione e clustering. Il machine learning viene spesso associato all'Intelligenza Artificiale

ma i due concetti non sono perfettamente sovrapposti: il machine learning è una branca dell'intelligenza artificiale, in quanto indica una macchina che è capace di apprendimento automatico a partire da una sufficiente quantità di dati.

Gli algoritmi di machine learning puntano a costruire un modello predittivo o decisionale partendo da un set di dati, detto *training data*. Gli approcci all'apprendimento, quindi alla costruzione del modello, sono molteplici e per via della recente esplosione negli avanzamenti in questo campo alcuni sono oggetto di studio tutt'oggi e non sono consolidati come standard (ad esempio il *Self-supervised learning*). Gli approcci tipici al machine learning, consolidati ed utilizzati universalmente in quest'ambito sono:

- Reinforcement Learning: in questo approccio l'algoritmo decide autonomamente quali azioni o decisioni generano le ricompense maggiori, secondo un processo di *trial and error*. I componenti principali di questo approccio sono: l'agente, che prende le decisioni e impara, l'ambiente e le azioni eseguibili.

A differenza di altri approcci il RL si occupa di decisioni sequenziali, in cui l'azione da compiere dipende dallo stato attuale del sistema e ne determina quello futuro. Per via della sua natura il RL è spesso utilizzato nella teoria dei giochi, nella robotica, nella statistica e nei sistemi multi-agente.

- Unsupervised Learning: è un approccio al machine learning in cui all'algoritmo non vengono fornite etichette per il dataset di training. In questo modo il primo step sarà creare cluster di dati e ricavare attributi e pattern nei dati in maniera inferenziale, senza l'intervento umano, per poi procedere alla creazione del modello.

Per sua natura questo approccio trova applicazione dove è impossibile o non efficiente etichettare i dati manualmente, ad esempio nella Computer vision o Natural Language processing (NLP). Un esempio notevole di applicazione di Unsupervised Learning è GPT-2, modello di NLP sviluppato da OpenAI, addestrato tramite lo scraping di testp da 8 milioni di pagine web.

- Supervised Learning: approccio in cui l'algoritmo costruisce un modello matematico di un insieme di dati che contengono esplicitamente sia gli input che gli output desiderati. L'algoritmo impara così ad ottenere i risultati desiderati da esempi espliciti che mettono in correlazione degli input con tali risultati. Si utilizzano a questo scopo quindi dataset etichettati per risolvere una vasta
-

gamma di problemi, come regressione, predizione e classificazione.

La classificazione tramite classificatori può avvenire secondo vari algoritmi, che hanno diverse performance in base alla conformazione del problema e di come vengono trattati i dati.

Nello specifico si possono individuare tra i classificatori algoritmi basati su:

- Calcolo del gradiente;
- Regressione lineare o non lineare;
- Boosting;
- Calcolo della probabilità e legge di Bayes;
- Alberi di decisione;
- Reti neurali.

Da notare che gli algoritmi di boosting sono una categoria trasversale in quanto mettono insieme più *learners* che utilizzano diversi algoritmi in quello che è noto come *ensemble di classificatori*.

In questo lavoro di tesi verranno utilizzate prevalentemente tecniche di machine learning supervisionato per la classificazione binaria, nello specifico per individuare la presenza o l'assenza del morbo di Parkinson in un paziente.

1.4 Parkinson

Il morbo di Parkinson è una malattia neurodegenerativa che causa una lenta perdita di controllo nei movimenti e in un terzo dei casi anche danni cognitivi. E' una condizione medica molto diffusa, la seconda malattia neurodegenerativa dopo l'Alzheimer e colpisce

- Lo 0,4% delle persone di età inferiore ai 40 anni, 1 persona su 250;
 - L'1% delle persone di età superiore ai 65 anni, 1 persona su 100;
 - Il 10% delle persone di età superiore a 80 anni, 1 persona su 10.
-

L'età media dell'insorgenza della malattia è di circa 57 anni, ma ci sono forme di parkinsonismo giovanile caratterizzate da un esordio della malattia tra i 21 e i 40 anni.

Mentre le cause dell'insorgenza della malattia sono ancora ignote ed oggetto di ricerca, sono note molte delle cause dei sintomi, dovuti alla degenerazione dei gangli basali.

I gangli basali sono aggregati di cellule nervose, situati in profondità nel cervello, che aiutano a coordinare i movimenti muscolari e sopprimere i movimenti involontari. Questo avviene tramite il rilascio di messaggeri chimici detti neurotrasmettitori, come la dopamina, che veicolano il messaggio tra le cellule nervose necessario per inviare un impulso, quindi un segnale, ai muscoli del corpo.

Con l'incorrere della malattia avviene la degenerazione dei gangli basali, che non riescono più a produrre dopamina e di conseguenza a generare gli impulsi per controllare il sistema muscolare, provocando i tipici sintomi associati a questa malattia come tremore, bradicinesia (movimenti lenti), perdita della coordinazione e dell'equilibrio, rigidità e debolezza muscolare. Oltre a questi sintomi troviamo insonnia, stipsi e problemi di deglutizione, ma anche sintomi mentali come depressione, demenza e allucinazioni. I sintomi insorgono gradualmente man mano che la degenerazione delle cellule cerebrali avanza, fino a portare inevitabilmente ad uno stato terminale.

Il Parkinson si presenta quindi come una malattia che coinvolge tutto il corpo, ma di cui sappiamo ancora molto poco. E' infatti ignota la causa della malattia e al momento non esiste alcuna cura per i pazienti parkinsoniani, il cui trattamento si limita alla fisioterapia e a farmaci per rallentare in parte l'avanzamento della malattia.

1.5 Problem Statement e stato dell'arte

Per via della natura della malattia, i pazienti affetti da morbo di Parkinson beneficiano da una diagnosi precoce, per essere un passo avanti alla malattia con i trattamenti[5]. Questo però non è sempre possibile: se è vero che nello stadio avanzato il Parkinson è facilmente diagnosticabile poiché i sintomi sono evidenti e tipici della sua sintomatologia, è vero anche che in una fase iniziale molti dei sintomi sono lievi o quasi inesistenti rendendo molto difficile il processo di diagnosi, che avviene principalmente tramite valutazione dei movimenti del paziente e procede per esclu-

sione di altre malattie neurologiche.

E' bene notare che questi stessi problemi di diagnosi si presentano anche nella forma avanzata della malattia, poiché non esiste attualmente esame o procedura di diagnostica per immagini che può confermare direttamente la diagnosi di Parkinson [4].

Per far fronte a questo problema il settore della ricerca si sta espandendo in nuovi approcci ed in particolare l'informatica medica si sta concentrando su tecniche di *automatic disease detection*, con particolare enfasi su procedure non invasive e telemedicina. Queste tecniche sono prevalentemente basate su algoritmi di machine learning, utilizzati per l'addestramento di un modello predittivo oppure di classificazione che impara da una grande quantità di dati quali sono i pattern da ricercare per individuare la presenza della malattia.

E' bene notare che al momento della stesura di questa tesi non esiste un metodo definitivo per la diagnosi, ma esistono lavori di ricerca che pongono le basi per una sua successiva implementazione tramite l'approccio al problema da vari angoli.

Secondo uno studio condotto a Maggio 2021 infatti, sono 209 i tentativi di successo di implementare una diagnosi del Parkinson in maniera automatica [6] la cui totalità utilizza metodi di machine learning, con una accuratezza che varia tra l'85.6% e il 94.4% a seconda dell'approccio utilizzato.

E' importante ricordare, nel considerare gli approcci utilizzati, che nello studio sopracitato sono analizzati solo i paper in lingua inglese e sono state escluse le pubblicazioni in altre lingue, al fine di facilitare il processo di text mining.

Gli studi presi in considerazione, per quanto simili nell'approccio alla diagnosi tramite machine learning, variano considerevolmente nel tipo di classificatore utilizzato e soprattutto nei dati di partenza utilizzati per il training, che possono essere:

- Registrazioni vocali, nel 26.3% degli studi;
 - Dati sui pattern di movimento, nel 24.4% ;
 - Dati sui pattern di scrittura o disegno a mano libera, nel 7.7% ;
 - Immagini ricavate tramite risonanza magnetica, nel 17.2% ;
 - Immagini ricavate tramite tomografia ad emissione di fotone singolo, nel 6.7%;
-

- Immagini ricavate tramite tomografia a emissione di positroni, nell'1.9% ;
- Combinazione dei precedenti, nell'8.6%.

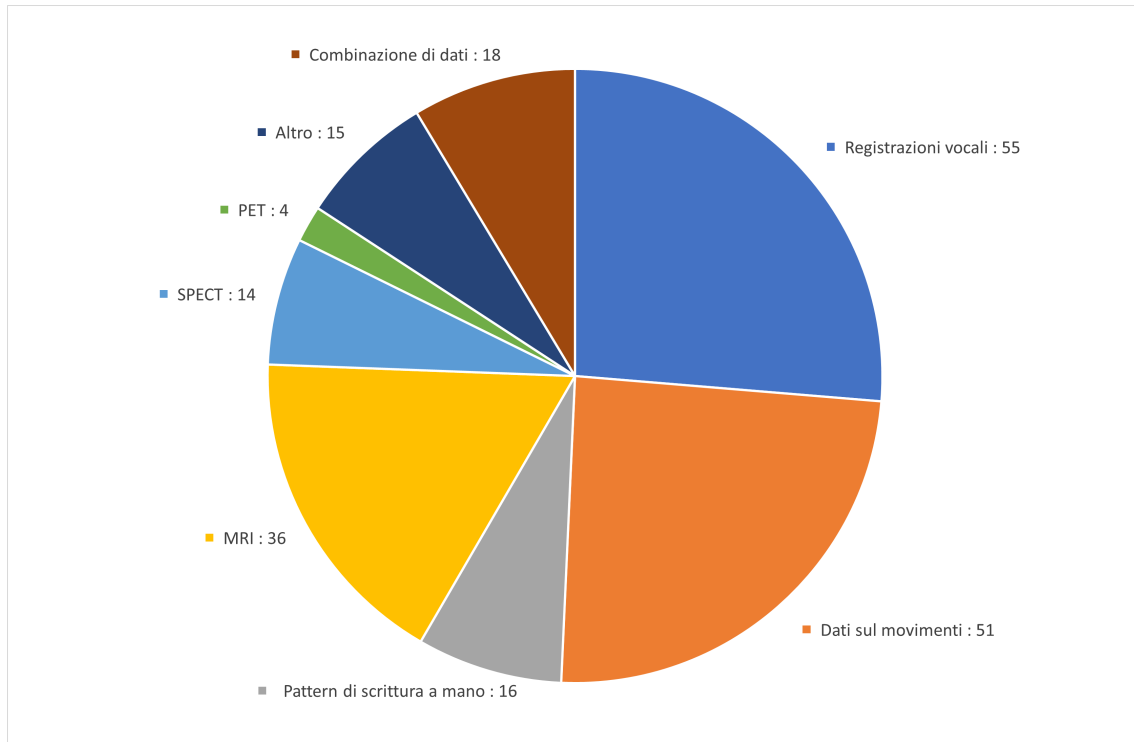


Figura 1.1: Sorgenti dei dati utilizzati

L'utilizzo di features vocali per il training di un algoritmo di machine learning può sembrare ad un primo impatto scorrelato dall'obiettivo di diagnosi del Parkinson. In realtà la malattia porta fin dagli stadi iniziali cambiamenti nella voce dei pazienti, in particolare sul tono, sul volume e sulle frequenze emesse, rendendo molto efficace questo metodo di valutazione

Capitolo 2

Setup e metodologia di lavoro

Come riscontrato nello studio dello stato dell'arte nella ricerca per la diagnosi automatica del Parkinson, un approccio basato sull'analisi di feature vocali genera risultati notevoli in combinazione con l'utilizzo di classificatori. In questo lavoro di tesi verrà analizzato proprio questo approccio, su un dataset contenente feature vocali di pazienti Parkinsoniani.

2.1 Dataset utilizzato

I dati utilizzati provengono da un dataset donato alla University of California Machine Learning Repository da parte del Dipartimento di Neurologia dell'Università di Instambul. I dati sono stati raccolti a partire da 188 pazienti affetti da morbo di Parkinson, 107 uomini e 81 donne con età da 33 a 87 anni, e un gruppo di controllo di 64 individui sani, 23 uomini e 41 donne, con età che variano tra 41 e 82 anni. Le registrazioni vocali sono avvenute in ambiente controllato con un microfono impostato a 44.1 Khz. Ogni partecipante ha effettuato tre ripetizioni della vocale /a/ per un totale di 756 registrazioni vocali.

Il Dataset è composto dalle feature vocali estratte da tali registrazioni da parte degli autori del paper "A Comparative Analysis of Speech Signal Processing Algorithms for Parkinson's Disease Classification and The Use of The Tunable Q-Factor Wavelet Transform"[8], che hanno donato il dataset per pubblico utilizzo alla repository UCI.

All'interno del paper si può leggere come il criterio di estrazione delle feature vocali sia cambiato rispetto ad altri approcci, seguendo le indicazioni di un lavoro prece-

dente che individuava nuovi marker di tipo non lineare nei pattern del parlato dei pazienti parkinsoniani[9], individuabili tramite la trasformata del segnale definita come TQWT (Tunable q-factor Wavelet Transform).

Il dataset ha una dimensionalità molto elevata per via dell'elevato numero di feature estratte, sono infatti 755 gli attributi, categorizzabili in:

- Attributi del paziente: Identificativo del partecipante, genere e classe. Gli attributi genere e classe sono binari, variano tra 0 e 1 per indicare i due generi o l'assenza e la presenza della malattia nel partecipante;
 - Feature appartenenti alla baseline: 21 features che rappresentano le più usate all'interno degli studi su diagnosi basata su features vocali. Includono tremolio (jitter), frequenza fondamentale, parametri sulle armoniche (noise floor, Signal to Noise Ratio) e pitch;
 - Attributi per l'analisi in frequenza: sono in totale 11 e descrivono l'intensità del segnale vocale espressa in dB, le frequenze utilizzate e la larghezza di banda del segnale;
 - Misure sulle corde vocali: sono 22 e vengono stimate a partire dal segnale vocale e dei modelli. Includono parametri che indicano la quantità di rumore emessa dalle corde vocali, durata di apertura e chiusura della glottide e decomposizione empirica in segnali elementari;
 - MFCC: 84 parametri di analisi simili alle misure sulle corde vocali, ma specifiche per individuare come il parkinson modifica il tratto vocale;
 - Features della trasformata del segnale senza TQWT: 182 attributi;
 - Features appartenenti alla TQWT: 432 attributi.
-

Gli autori dello studio che ha pubblicato i dati hanno quindi preso in considerazione una vasta gamma di parametri ai fini dell'analisi, aggiungendo oltre a quelli considerati in altri studi ben 432 parametri appartenenti alla TQWT.

2.2 Pre elaborazione dei dati

Volendo riformulare brevemente l'obiettivo del lavoro che segue, si cercherà di diagnosticare correttamente il morbo di Parkinson solo tramite le informazioni ricavate da un segnale vocale e si metteranno a confronto i vari metodi che è possibile utilizzare per farlo.

Volendo calare questo approccio nel mondo reale possiamo dire che, con una quantità molto maggiore e più diversificata di dati, ad esempio includendo anche pazienti che hanno malattie diverse dal parkinson che influenzano la voce, il modello sviluppato potrebbe essere integrato in un sistema di diagnosi automatica.

Si deve inoltre ricordare che essendo la diagnosi di tipo clinico, ovvero basata sui sintomi, un tool che utilizzi questo modello non si deve in alcun modo sostituire al personale medico nel processo di diagnosi, ma deve soltanto assisterlo nella decisione.

La creazione del modello si presenta quindi come una task di classificazione o predizione, in quanto si può predire il valore dell'attributo classe (presenza o assenza della malattia) ed è bene a questo punto analizzare gli step che serviranno a raggiungere l'obiettivo di una corretta classificazione e predizione.

Qui vengono aggiunti dei paragrafi che mostrano in maniera abbastanza astratta a grandi linee il lavoro svolto dopo.

Capitolo 3

Approccio al problema

In questo capitolo si mostrerà il procedimento operativo utilizzato per il raggiungimento dei risultati finali. In generale il processo si può astarre ad alto livello come composto di soli tre step:

- Pre-elaborazione dei dati;
- Training dei classificatori;
- Testing e analisi dei risultati.

3.1 Linguaggio e librerie

Il linguaggio di programmazione prevalentemente utilizzato in questo lavoro è Python, con una piccola sezione in R.

Questa scelta si allinea con quello che è oggi lo standard nel campo della Data Science, in cui si utilizzano principalmente Python ed R per via di vari fattori, primo su tutti la grande quantità di risorse e letteratura a disposizione per le task più comuni. L'utilizzo di Python rende infatti immediato l'utilizzo di librerie per il caricamento e la visualizzazione dei dati, la manipolazione di strutture dati e l'esecuzione di task di Data Mining e Machine Learning, mettendo al contempo a disposizione dello sviluppatore tutto quello che è necessario per una programmazione sia procedurale che object oriented.

3.1.1 Librerie di Python utilizzate

Nella scrittura dello script di classificazione sono state utilizzate le seguenti librerie Python:

- Pandas: libreria fondamentale per caricamento, manipolazione e analisi dei dati all'interno dell'ambiente di lavoro. Tramite questa libreria possiamo leggere dati da un gran numero di sorgenti in diversi formati, nel caso di questo lavoro in formato CSV.

I dati letti in questo modo sono incapsulati in strutture dati di tipo Series, vettori di dati con un indice associato, o DataFrame, strutture bidimensionali che riflettono la composizione di una tabella indicizzata, con colonne e righe accompagnate da un indice. Le colonne di un DataFrame sono a loro volta delle Series che condividono lo stesso indice per l'ordine degli elementi.

I dati letti possono essere manipolati o analizzati in vari modi tramite operazioni sulle strutture dati che le contengono. Ad esempio possiamo riempire i valori nulli di un DataFrame, convertire i tipi degli oggetti o effettuare operazioni di filtraggio per recuperare un subset dei dati.

- Sklearn: appartenente al progetto Scikit-learn, questa libreria open source contiene algoritmi di apprendimento automatico e machine learning. All'interno del pacchetto sono presenti algoritmi di classificazione, regressione, clustering, macchine a vettori di supporto (SVM), classificazione di tipo bayesiano e probabilistica e algoritmi di k-mean regression e clustering. Questa libreria è progettata per essere utilizzata con le altrettanto note NumPy e SciPy, ed è spesso utilizzata in combinazione con i tool messi a disposizione da Pandas.

3.2 Preparazione dei dati

Per una corretta analisi bisogna verificare che i dati siano di buona qualità, poiché seguendo il principio di *garbage in garbage out* se la qualità dei dati in input è scadente, si otterrà un modello poco significativo per via degli effetti di rumore e dati mancanti.

In questa fase viene seguito il workflow relativo al Data Mining analizzato nel capitolo precedente. Ricordiamo che è composto dalle seguenti fasi:

- Data collection and integration;
- Data selection and reduction;
- Data cleaning;
- Data transformation and preprocessing.

Nella fase di data collection sono stati analizzati i dataset presenti su repository open source come UCI Machine Learning Repository e Kaggle. Sono stati trovati 10 datasets contenenti feature vocali di pazienti parkinsoniani e gruppi di controllo, ma una prima analisi ha rivelato la presenza di dati duplicati e ripubblicati, portando il numero effettivo di datasets disponibili sull'argomento a 4.

I 4 datasets trovati presentano caratteristiche molto differenti: tre di essi infatti fanno riferimento ad un numero esiguo di pazienti (<40) e contengono meno di 100 istanze, mentre quello rimanente contiene i dati di 252 partecipanti con 3 istanze per partecipante, quindi un totale di 756 istanze nel dataset. Le differenze però si trovano anche nelle features utilizzate e nella metodologia: la similitudine tra le features estratte dal segnale vocale tra i vari datasets è quasi zero, con differenze sostanziali sia nelle feature in sé sia per i parametri di estrazione di feature parametrizzate, ad esempio le trasformazioni in frequenza.

A porre un ulteriore ostacolo all'integrazione è la metodologia di acquisizione dei dati, per cui in due dei quattro datasets si sono utilizzati segnali vocali provenienti dalla stessa frase, ripetuta una volta per paziente, mentre negli altri due si è utilizzato rispettivamente la vocale /a/ e una diversa frase. Analizzando bene il contenuto dei datasets quindi, risulta impossibile un corretto processo di integrazione dei dati, poiché se anche fanno tutti riferimento allo stesso tipo di studio, sono stati acquisiti in maniera troppo diversa per essere messi insieme nella costruzione di un modello.

A fronte di questa analisi viene scelto il dataset con il maggior contenuto informativo, il file "pd_speech_features.csv" proveniente dal paper "A Comparative Analysis of Speech Signal Processing Algorithms for Parkinson's Disease Classification and The Use of The Tunable Q-Factor Wavelet Transform"[8], in quanto contiene 756 istanze di 252 partecipanti diversi e 755 features estratte. Questo dataset è stato raccolto con la precisa intenzione di effettuare questo studio, quindi analizzando i

valori troviamo che nessuno di essi è mancante e non ci sono ridondanze (istanze uguali) o dati chiaramente sbagliati, nel limite dell'analizzabile.

3.3 Bilanciamento del DataFrame

Il dataset risulta a questo punto completo per tutti i valori e pulito da errori, ma presenta una caratteristica che può inficiare pesantemente la creazione del modello, ovvero i dati non sono bilanciati. Questo significa che nel dataset sono presenti più istanze di una classe rispetto all'altra e quindi i classificatori allenati con questi dati avranno un *bias* verso la classe più numerosa all'interno del training set.

Per ovviare a questo problema esistono vari approcci, che dividiamo in tre principali categorie:

- **Oversampling:** tecnica in cui si cerca di incrementare artificialmente il numero di istanze della classe meno rappresentata tramite la creazione di istanze sintetiche appartenenti alla stessa classe. Al termine della procedura le classi avranno lo stesso numero di istanze. Il tradeoff in questa tecnica è che per quanto si possano raggiungere risultati molto migliori senza perdere informazioni sul dataset (senza rimuovere istanze), si introducono però nuove informazioni replicando alcune istanze della classe meno rappresentata. Tali informazioni potrebbero non riflettere la realtà e peggiorare il modello oppure essere poco rilevanti e migliorarlo con il solo effetto di bilanciare i dati.
- **Undersampling:** tecnica in cui si cerca di diminuire le istanze della classe/i più numerosa rimuovendone alcune dal dataset casualmente o seguendo un criterio specifico. Questo porta ad avere un numero uguale di istanze delle varie classi a discapito del contenuto informativo presente nelle istanze rimosse.
- **Class weighting:** alle classi viene assegnato un peso, con la classe più numerosa che ha un peso minore della classe meno numerosa. Dando un peso alle varie classi possiamo bilanciare l'analisi proveniente dal dataset senza manipolare le istanze.

In questo lavoro verrà utilizzato una particolare tecnica di oversampling detta SMO-TE (*Synthetic Minority Oversampling Technique*).

Questa tecnica genera nuove istanze della classe di cardinalità minore a partire dal

dataset di partenza. A differenza di altre tecniche di oversampling che possono introdurre overfitting, questa tecnica minimizza il problema inserendo valori nelle feature delle istanze che non cambiano la distribuzione dei valori nel totale. In altre parole non viene aggiunto contenuto informativo che non sia già presente nel dataset di partenza.

```
if balance_data:
    smote = SMOTE('not majority', random_state=1)
    x_train, y_train = smote.fit_resample(x_train, y_train)
    x_test, y_test = smote.fit_resample(x_test, y_test)
```

Figura 3.1: Utilizzo di SMOTE a partire dalla libreria imblearn

3.4 Feature selection

Una seconda problematica da affrontare prima di passare alla fase di training dei classificatori riguarda la feature selection. Le istanze presenti nel dataset utilizzato infatti hanno una alta dimensionalità per quanto riguarda gli attributi e questo, per quanto ci fornisca molti parametri di decisione può inficiare il processo di classificazione in vari modi.

Una alta dimensionalità dei dati porta a due problemi principali:

- Aumento del costo computazionale: la grande quantità di features presenti porta ad allungare considerevolmente i tempi di addestramento e utilizzo del modello, che deve tenere in considerazione molti più parametri.

Puramente come esempio possiamo allenare il classificatore GradientBoostingClassifier con l'intero dataset senza una riduzione dimensionale, impiegando un tempo di 71 secondi. Sulla stessa macchina è stato allenato lo stesso modello, utilizzando solo un subset rilevante delle features in soli 11 secondi, con risultati di performance del modello comparabili. Il modello che utilizza tutto il feature set ha quindi impiegato un tempo del 645% maggiore pur avendo risultati operativi simili.

Bisogna tenere in considerazione anche che se 71 secondi sembra un tempo, in valore assoluto, non esagerato, questo divario va ad aumentare man mano che

vengono aggiunte istanze nel dataset, quindi in un contesto operativo dove si opera con questo stesso workflow ma un numero molto maggiore di dati non effettuare alcun tipo di feature selection porta grandi limitazioni in termini di scalabilità della soluzione.

- **Overfitting:** Un modello allenato con una grande quantità di features può risultare troppo specifico e non generalizzabile, rendendolo inaccurato quando un nuovo set di dati viene proposto. Questo avviene poiché non tutti gli attributi sono significativi al fine della nostra analisi; alcune features non portano contenuto informativo per la task che stiamo svolgendo e si comportano all'atto pratico come rumore, anche se non lo sono.

Un esempio può essere un attributo che indica a che ora è stato prelevato un campione audio: se per valutare in quanto tempo possiamo acquisire abbastanza campioni può essere utile un attributo del genere, porterà sicuramente poche o nessuna informazione per una analisi come quella svolta in questo lavoro, introducendo rumore e contribuendo all'overfitting del modello.

I benefici della feature selection su questo dataset sono quindi chiari, essendo denso di feature. Una selezione delle feature manuale è una delle opzioni migliori se si conosce il dominio applicativo e gli attributi superflui sono noti a priori, ma una selezione solo manuale non è quasi mai possibile, se non in contesti specifici; per questo ci si affida a metodi di feature selection automatica, che possono essere di tipo unsupervised o supervised.

Si parla di unsupervised feature selection se si applica un algoritmo di machine learning che cerca di comprendere autonomamente l'apporto informativo di ogni feature e scartare quelle poco importanti, mentre si parla di supervised feature selection se tale processo è guidato.

Esistono molti metodi per la feature selection in letteratura ed in generale non vi è uno standard che si può applicare in tutti i casi, ma bisogna valutare caso per caso quale è il metodo più appropriato. Nel paper da cui sono tratti i dati viene utilizzato un algoritmo chiamato mrMr (minimum redundancy Maximum relevance). Questo algoritmo ha guadagnato popolarità negli ultimi anni dopo essere stato utilizzato con successo in uno studio recente [9] e si basa sul selezionare un feature set in cui le feature hanno il maggiore impatto sulla variabile (maximum relevance) target ma sono correlate al minimo tra di loro (minimum redundancy).

In questo lavoro viene utilizzato un algoritmo che segue lo stesso principio, implementato in maniera differente: si effettua un dendrogramma delle feature, che viene tagliato ad una certa altezza per selezionare un subset ottimale delle stesse. In altre parole, il dendrogramma ci permette di effettuare il clustering delle features e di capire quanto esse siano ridondanti sulla base della mutua correlazione.

Tagliando il dendrogramma viene selezionata casualmente da ogni cluster una fea-

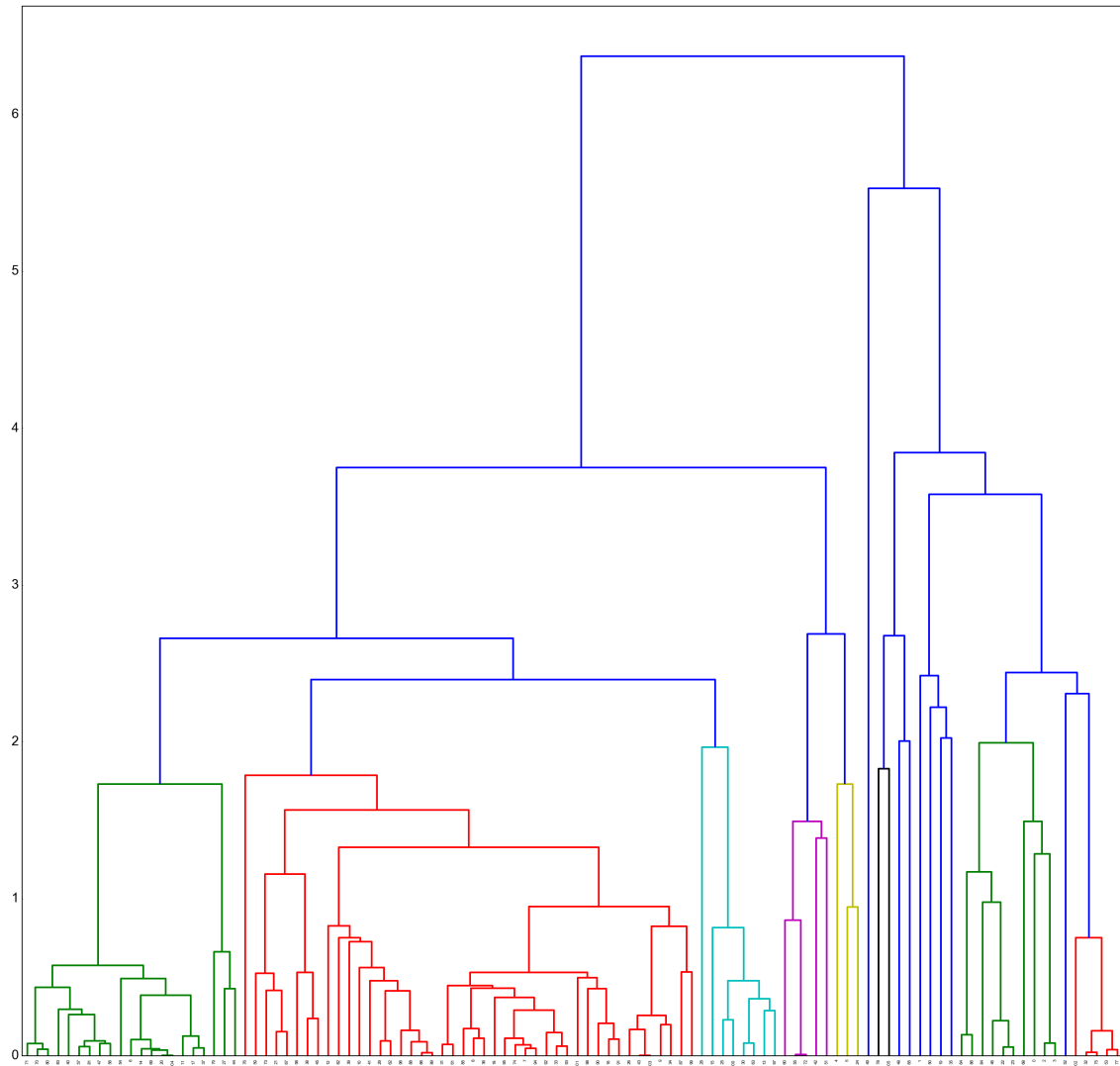


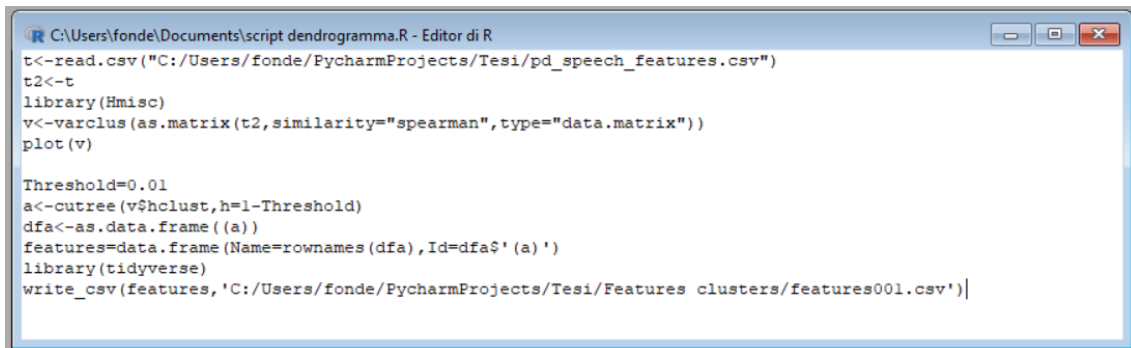
Figura 3.2: Esempio di dendrogramma

ture. Il fatto che il taglio sia casuale non inficia la rilevanza di tale feature poiché il dendrogramma è costruito raggruppando features simili, che hanno lo stesso effetto sulla variabile target.

La creazione ed il taglio del dendrogramma sono stati effettuati con R, per facilitar-

ne il processo vista l'ampia presenza di letteratura e risorse a riguardo. Il taglio è stato effettuato a diverse altezze dell'albero in modo da testare gli effetti di avere un numero maggiore o minore di features con un diverso grado di correlazione. Il taglio è stato effettuato alle soglie di correlazione : 0.9 , 0.7 , 0.5 , 0.3 , 0.2 , 0.1 .

Il risultato del taglio del dendrogramma è un file che contiene i cluster di feature,



```

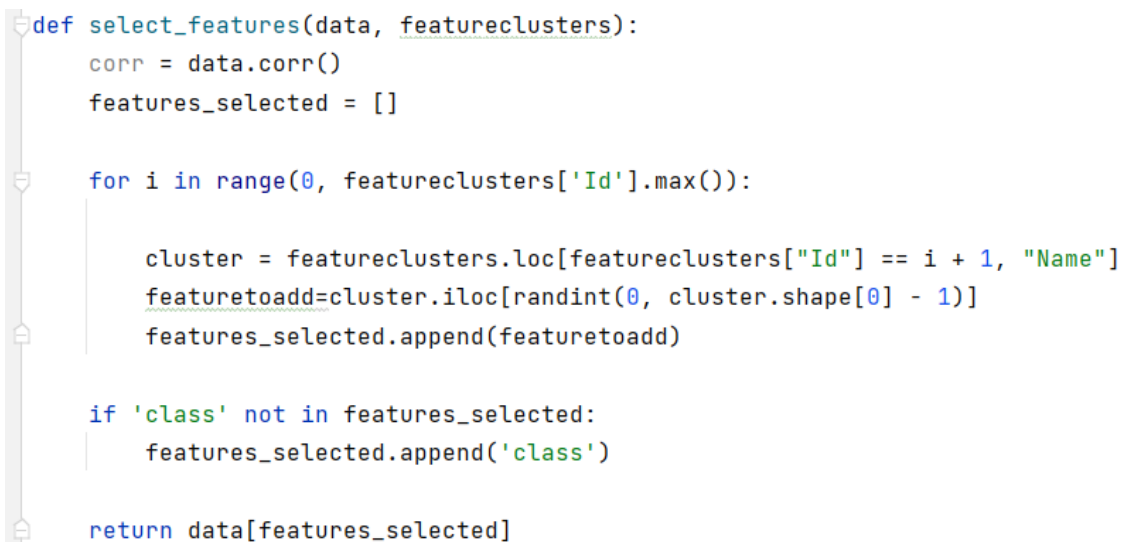
C:\Users\fonde\Documents\script dendrogramma.R - Editor di R
t<-read.csv("C:/Users/fonde/PycharmProjects/Tesi/pd_speech_features.csv")
t2<-t
library(Hmisc)
v<-varclus(as.matrix(t2,similarity="spearman",type="data.matrix"))
plot(v)

Threshold=0.01
a<-cutree(v$hclust,h=1-Threshold)
dfa<-as.data.frame((a))
features=data.frame(Name=rownames(dfa),Id=dfa$(a))
library(tidyverse)
write_csv(features,'C:/Users/fonde/PycharmProjects/Tesi/Features_clusters/features001.csv')
  
```

Figura 3.3: Script per la generazione e il taglio del dendrogramma

rami del dendrogramma che sono stati collassati in foglia ad una certa soglia di correlazione dallo script di cui sopra. Per estrarre una feature dal cluster utilizziamo Python, in quanto il resto del lavoro sarà svolto in questo ambiente ed è conveniente ai fini operativi poter manipolare questi dati in Python nelle diverse fasi del processo.

Il risultato del taglio riduce il numero di feature sensibilmente. Di seguito sono



```

def select_features(data, featureclusters):
    corr = data.corr()
    features_selected = []

    for i in range(0, featureclusters['Id'].max()):

        cluster = featureclusters.loc[featureclusters["Id"] == i + 1, "Name"]
        featuretoadd=cluster.iloc[randint(0, cluster.shape[0] - 1)]
        features_selected.append(featuretoadd)

    if 'class' not in features_selected:
        features_selected.append('class')

    return data[features_selected]
  
```

Figura 3.4: Selezione di una feature dai cluster di features

mostrate il numero di features per il taglio a diverse soglie di correlazione

Come si può notare, più si abbassa la soglia di correlazione a cui effettuiamo il

Soglia di correlazione	Numero di features
0.9	456
0.7	342
0.5	278
0.3	224
0.2	203
0.1	179

Tabella 3.1: Feature ricavate alle diverse soglie di correlazione.

tagio, meno feature vengono selezionate. Questo accade perché ad una bassa soglia di correlazione sono di più le features che vengono considerate correlate e quindi appartenenti allo stesso cluster, creando quindi meno clusters ma con un numero maggiore di features al loro interno. Estrahendo una sola feature per cluster questo risulta direttamente nell'avere meno features.

3.5 Scelta dei classificatori

3.6 Parametri di classificazione

3.7 [OPZ] Ensemble di classificatori

Capitolo 4

Analisi dei risultati

4.1 Recap

4.2 Risultati utilizzando tutte le feature

4.3 Risultati applicando feature selection a varie threshold

4.4 Risultati applicando anche bilanciamento

4.5 Eventuali altri risultati: Ensemble, normalizzazione e discretizzazione, altri algoritmi di feature selection

Capitolo 5

Conclusione e sviluppi futuri

Bibliografia