UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO

DIPARTIMENTO DI INFORMATICA

CORSO DI LAUREA TRIENNALE IN INFORMATICA

TESI DI LAUREA

IN

RETI DI CALCOLATORI

RICONOSCIMENTO DI EMOZIONI DI SVILUPPATORI SOFTWARE BASATO SU SENSORI BIOMETRICI NON INVASIVI

RELATORI:

Prof.ssa Nicole Novielli

Prof. Filippo Lanubile

LAUREANDA:

Daniela Grassi

ANNO ACCADEMICO 2019-2020

Indice

1. Introduzione

Le emozioni pervadono ogni aspetto della nostra vita. Ogni istante è caratterizzato da sfumature di sentimenti, ondate di emozioni, stati di umore e affetto. Il mondo della letteratura, del teatro e del cinema è caratterizzato essenzialmente da opere che ruotano intorno a questioni emozionali e che esplorano la vita interiore ed emotiva dei loro personaggi. In effetti, le emozioni sono oggetto di numerosi studi multidisciplinari. In particolare, in ambito informatico già a metà degli anni ’90, Rosalind Picard introduce il concetto di “*Affective Computing*”. [25] Secondo Picard la macchina deve interpretare gli stati emotivi degli esseri umani adattando il proprio comportamento ad essi e alle loro emozioni. In altre parole, l’oggetto di ricerca del “*Affective Computing*” è la progettazione e lo sviluppo di sistemi in grado di riconoscere, interpretare e simulare emozioni umane. Gli sviluppatori software possono avvalersi del riconoscimento emotivo sia per poter rendere i propri progetti adattivi rispetto alle emozioni dell’utente, sia per migliorare la propria esperienza lavorativa. Infatti, emozioni positive sono favorevoli per il benessere e la produttività degli sviluppatori, mentre emozioni negative, come rabbia o stress, sono dannose per il rendimento lavorativo e rallentano lo sviluppo di programmi software. [12]

Nel presente lavoro ci siamo interessati di indagare la stretta correlazione che si pone tra emozioni e cambiamenti fisiologici, in particolare durante le attività lavorative nell’ambito dell’ingegneria del software. Questo studio si compone di due sperimentazioni: la prima ha l’obiettivo di identificare quale sia la migliore configurazione per addestrare un classificatore che, a partire dalle feature biometriche, sia in grado di riconoscere e classificare le emozioni in tre classi relative alla valenza emotiva: *positive*, *negative* e *neutral*. Il dataset utilizzato è stato raccolto in laboratorio presso l’università degli studi di Bari, ed è stato elaborato studiando diversi set di dati (con un numero crescente di casi *neutral*) e diverse partizioni del set di training e di testing: 70-30, 80-20 e 90-10.

La seconda sperimentazione sfrutta i risultati della prima per poter costruire un classificatore con setting ottimale che sia in grado di riconoscere sia la valenza che l’attivazione emotiva considerando un set di dati raccolto durante lo studio sul campo (*field study*) condotto dalla dott.ssa Daniela Girardi. Le metriche considerate sono *precision*, *recall*, *F1*, *accuracy* e *weighted Cohen's Kappa*.

Il lavoro di tesi strutturato come segue. Il secondo capitolo fornisce una panoramica teorica sul riconoscimento delle emozioni, introducendo il modello circonflesso di Russell e le emozioni in ambito teorico, il terzo capitolo illustra le varie misure biometriche e il loro utilizzo nell'ingegneria del software. Il quarto capitolo espone le modalità di raccolta dei dati e la conseguente configurazione. Il quinto capitolo descrive la metodologia di classificazione utilizzata, facendo riferimento ai modelli, agli algoritmi e alle metriche. L'analisi dei risultati viene, infine, esposta nel capitolo sei e nel capitolo sette seguito dal capitolo conclusivo.

2. Stato dell'arte e modelli teorici di riferimento

2.1 Modello teorico delle emozioni.

Diversi sono i modelli teorici sviluppati dagli psicologi per descrivere e categorizzare le emozioni. Tra queste, le teorie discrete propongono un sistema di classificazione di tipo categoriale, dove le emozioni sono classificate come entità discrete, indipendenti le une dalle altre e facilmente distinguibili. Nell'ultimo decennio però è stato proposto un approccio alle emozioni di tipo dimensionale che ne facilita l'identificazione e la caratterizzazione [4]. Uno tra i precursori di questo approccio è James Russell, con il modello circomplesso delle emozioni [2]. In particolare, Russell sostiene che gli stati affettivi sono riconducibili a due principali sistemi neurofisiologici: *Valence* e *Arousal.* La *Valence* (valenza emotiva) si riferisce alla piacevolezza di uno stato emotivo; stati come gioia o divertimento sono associati a un valore alto di *Valence*, mentre quelli non piacevoli corrispondono a dei valori bassi della stessa. L'*Arousal* descrive il livello di attivazione dello stato emotivo che va da inattivo (basso *Arousal*), come noia o sonnolenza, fino ad attivo (alto *Arousal*), come eccitazione o rabbia [3]. La gioia, ad esempio, è concettualizzata come uno stato emotivo connotato da *Valence* positiva e da un livello di *Arousal* moderato.

Il modello circonflesso di Russell è mostrato in figura 2.1.

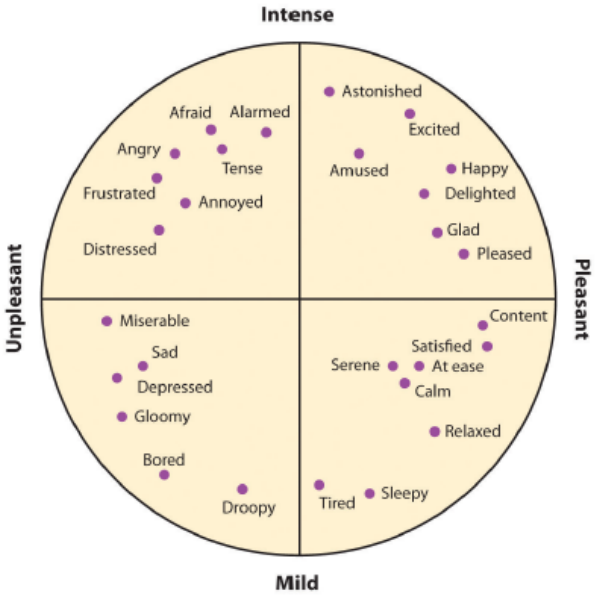


Figura 2.1*: Il modello circomplesso di Russell* [34]

Così come avviene per le diverse gradazioni di colore, allo stesso modo, le emozioni non sono chiaramente distinguibili e separabili le une dalle altre. Chi, ad esempio, prova un'emozione di gioia, probabilmente avvertirà anche altre sensazioni positive (ad esempio, sorpresa o allegria, eccitazione, euforia, soddisfazione, fierezza, etc.), che, infatti nel modello circomplesso sono poste molto vicine tra loro.

3. Riconoscimento di emozioni basato su analisi biometrica

Secondo l'autore del modello circomplesso, la combinazione di *Valence* e *Arousal*, associate alla risposta fisiologica scaturita dalla stimolazione elicitante e dalla percezione cognitiva, darebbero origine all' emozione [2]. Il corpo umano, in sostanza, reagisce quando avverte un'emozione: il cuore batte più velocemente, i palmi delle mani cominciano a sudare, i muscoli si contraggono o si rilassano. I sensori biometrici forniscono nuove opportunità per misurare i cambiamenti fisiologici che possono essere legati a mutamenti psicologici [8]. Inoltre, questi risultano molto utili perché permettono di operare anche in ambienti non controllati, dove non è necessario che l'utente sia osservato in uno spazio definito, mentre esegue compiti specifici [6].

In questo lavoro di tesi sono stati considerati i seguenti parametri biometrici:

2.2.1 Frequenza Cardiaca

La frequenza cardiaca è il numero di battiti del cuore al minuto. Insieme alla temperatura corporea, la pressione sanguigna e il ritmo respiratorio, è una delle funzioni vitali. [7] È una misura variabile, che può aumentare se si presentano situazioni inattese o eventi improvvisi o che può decelerare se ci si trova in situazioni rilassanti.[6] È ricavata attraverso algoritmi di conversione applicati al segnale catturato dal pletismografo, un sensore ottico posizionato, di solito, su un dito. [3]

2.2.2 Elettroencefalografia

L'elettroencefalografia, comunemente detto EEG, è la misurazione, attraverso l'applicazione di un certo numero di elettrodi, dell'attività elettrica del cervello. Gli elettrodi posizionati sul cuoio capelluto di un individuo misurano le fluttuazioni di tensione che riflettono l'attività neurale. Quando gli elettrodi sono posizionati correttamente, forniscono informazioni sulle diverse attività in ogni regione del cervello. Questi dati possono essere utilizzati per indagare i processi cognitivi dell'individuo. [30] Lo studio di Fritz et al. [8] riconosce come le bande di frequenza cerebrali, spesso indicate come *alfa, beta, gamma, delta* e *theta*, all'interno dei dati EEG possono essere collegate a svariati stati mentali e a diversi tipi di emozioni. Ad esempio, una diminuzione dell'attività *alfa* e un aumento dell'attività *theta* è sintomo di un aumento dell'attenzione e dello sforzo di memorizzazione.

2.2.3 Risposta galvanica della pelle

La risposta galvanica della pelle (GSR), anche chiamata attività elettro-dermica (EDA), è la misura delle variazioni delle caratteristiche elettriche della pelle, come ad esempio la conduttanza, a seguito della variazione della sudorazione del corpo umano. [3] Essa rappresenta l'inverso della resistenza elettrica, in quanto l'apertura delle ghiandole sudoripare eccrine favorisce il passaggio della corrente attraverso il derma. Più una persona suda, maggiore sarà la conduttanza, registrata soprattutto a livello palmare e della pianta dei piedi. L'Attività elettro-dermica ha un'influenza incisiva sull' *Arousal*,

sull’attenzione, su stress e ansia [8], infatti Nourbakhsh et al. [9] mostrano come il dominio di frequenza dell'attività elettro-dermica sia indicativo per individuare la difficoltà durante le attività di lettura e compiti aritmetici.

2.2.4 Attività elettrica muscolare

L’elettromiogramma (EMG) è il segnale elettrico prodotto dalle fibre muscolari quando queste ricevono dai motoneuroni gli impulsi elettrici che ne causano la contrazione. Le contrazioni muscolari generano elettricità, questa si propaga lungo i tessuti, le ossa e nelle zone cutanee vicine. Tuttavia, questo segnale è misurabile anche quando non ci sono contrazioni visibili, ad esempio quando si cerca di controllare il corpo affinché non si verifichino determinati comportamenti [3]. l’EMG un ottimo metodo per monitorare il processo cognitivo comportamentale e per la predizione delle emozioni.

2.3 Analisi biometrica per lo sviluppo software

Nell'ingegneria del software, le misurazioni biometriche possono essere utilizzate per ottenere informazioni sui processi cognitivi ed emotivi degli sviluppatori software, inoltre, possono essere utilizzate per fornire un supporto immediato agli sviluppatori, ad esempio, impedendo che questi introducano bug nel codice [30].
Poiché gli sviluppatori dedicano una notevole quantità di tempo alla lettura e alla scrittura del codice, un impiego dei sensori biometrici nell’ingegneria del software riguarda come gli sviluppatori leggono e comprendono il codice sorgente. Diversi studi hanno utilizzato l'eye tracking per analizzare il modo in cui gli sviluppatori leggono gli algoritmi e il codice sorgente, per esaminare su cosa si concentrano maggiormente durante la lettura di frammenti di codice e come la lettura del codice si confronta con la lettura del testo naturale. Una delle scoperte interessanti è che la lettura del testo naturale avviene in gran parte linearmente, da sinistra a destra e dall'alto verso il basso. Per il codice sorgente, l'eye tracking ha rivelato che gli esperti leggono il codice sorgente in modo meno lineare rispetto ai principianti [32]. Un altro studio dimostra che i segmenti di codice iniziali (ad esempio, in una funzione) sono letti più volte ricevendo maggiore attenzione, mentre le parti successive possono essere solo scremate [33].

Un altro uso dei sensori biometrici nell'ingegneria del software si ritrova nell'ambito della "*code comprehension*"; infatti la comprensione del codice è tra i task più impegnativi per gli sviluppatori software. Un sondaggio condotto dalla NASA ha dichiarato che quando si sviluppa software la comprensione del codice è molto più importante della correttezza funzionale [26]. Similmente, aziende come Facebook e Google richiedono sempre più la revisione del codice per i nuovi aggiornamenti; questo spiega l'importanza della "*code comprehension*" nell'ambito pratico e di ricerca. [27]

Un esempio di analisi in questo ambito è rappresentato dallo studio di Floyd et al. [27], i quali hanno analizzato la comprensione del codice tramite l'utilizzo dell'fMRI (Risonanza Magnetica funzionale) per determinare quali tipi di task gli sviluppatori stiano svolgendo in relazione anche alle loro competenze. Lo studio analizza tre tipi di task: comprensione di codice, revisione di codice e revisione di una prosa. Questo ha rilevato che le rappresentazioni neurali del nostro cervello quando visioniamo prosa e quando visioniamo codice sono diverse, infatti è stato possibile costruire un modello che classifica i compiti dei partecipanti in base all'attività cerebrale. Da questa analisi è emerso che le rappresentazioni neurali sono modulate dall'esperienza, infatti nei programmatori più esperti la differenza tra le due rappresentazioni è minima, ciò significa che questi considerano il codice come linguaggio naturale.

In un altro studio fMRI, Siegmund et al. hanno dimostrato come durante la comprensione del codice *bottom up* (ovvero l'interpretazione del codice dichiarazione per dichiarazione) gli sviluppatori adottano metodi di *chunking semantico* e sono facilitati dalle firme dei metodi, come ad esempio i prototipi in C++, e dagli idiomi di programmazione comuni. [31]

L'analisi biometrica tramite sensori è utile anche ad individuare episodi di *burnout* e *turnover* che possono essere causati da un forte stress lavorativo. Infatti, durante lo sviluppo software i programmatori provano un'ampia gamma di emozioni: si passa facilmente da momenti di grande contentezza per aver trovato soluzioni ad un problema a profonda frustrazione o rabbia nel sentirsi incapaci di procedere con il lavoro. [10]

Questi sentimenti si ripercuotono sulla qualità del software sulle interazioni sociali nell'ambiente lavorativo. Per esempio, nello studio di Muller e Fritz [10] sono stati utilizzati l'fMRI e l'Eye-tracking, per monitorare quali attività alterino le emozioni e in quale range, durante attività di sviluppo software.

Nel nostro studio che mira a riconoscere le emozioni sono stati utilizzati sensori biometrici poco invasivi: il braccialetto Empatica E4, per la misurazione della frequenza

cardiaca e GSR, e il caschetto BrainLink, per il rilevamento dell'EEG. Questi dispositivi sono visibili nelle figure successive.

Figura 2.1: *Braccialetto Empatica E4*

Figura 2.3: *Caschetto Brainlink*

4. Dataset

La sperimentazione prende in considerazione due dataset: uno acquisito in laboratorio presso l'università degli studi di Bari, il secondo ottenuto durante lo studio sul campo condotto dalla dott.ssa Daniela Girardi presso aziende di sviluppo software. [35]

3.1 Studio in laboratorio

3.1.1 Raccolta Dati

La sperimentazione originale [13] da cui derivano i dati, analizza il legame tra emozioni e progresso percepito, e il grado di precisione dei sensori biometrici supportati dall'analisi delle espressioni del volto.
Nello studio [13] è stato utilizzato il feedback biometrico per addestrare e valutare un classificatore che utilizza tecniche di machine learning per distinguere le emozioni positive da quelle negative.
L'esperimento di raccolta dati è suddiviso in due fasi e in entrambe le fasi i partecipanti, selezionati tra gli studenti di informatica dell'Università degli Studi di Bari, indossano i sensori biometrici Empatica E4 e Brainlink, e sono ripresi da una webcam.
Nella prima fase il compito è quello di osservare alcuni video che permettano di calibrare la baseline biometrica in situazioni neutrali.
Nella seconda fase ai partecipanti è assegnato un task di programmazione da svolgere. Durante l'esecuzione del task ogni partecipante è interrotto circa ogni cinque minuti. Nel corso di ogni interruzione ogni partecipante compila il questionario *Self-Assessment Manikin* (SAM) in cui indica qual è il suo stato emotivo e le difficoltà riscontrate durante lo svolgimento del task. Alla fine, viene mostrato ai partecipanti un video rilassante della durata di due minuti.

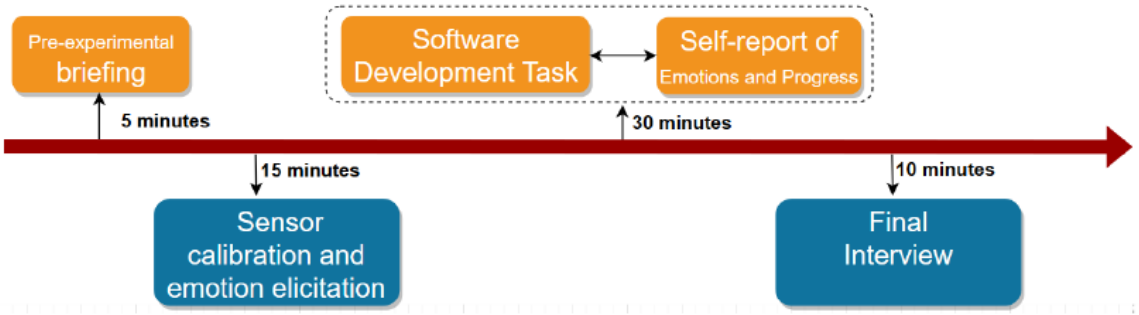


Figura 3.1: *La timeline dello studio* [13]

Una volta ottenuto il dataset, nello studio originale sono stati scelti otto algoritmi di machine learning per la costruzione di un classificatore (Naive Bayes, K-Nearest Neighbor, C4.5- like trees, SVM, Multi-layer Perceptron for neural network e Random Forest). Inoltre, sono stati considerati due differenti *validation setting: LORO (leave-one-run-out) cross-validation* e *Hold-out*, le metriche considerate sono state *precision*, *recall F1-score e accuracy*.

I migliori classificatori, per quanto riguarda l'Hold-Out Setting sono il K-Nearest Neighbor e il Random Forest, mentre per LOSO sono il Support Vector Machines e il Multilayer Perceptron.

Le performance del classificatore sono riportate in figura 3.2

I risultati [13] dimostrano che è possibile riconoscere le emozioni di sviluppatori software durante le attività di programmazione, soprattutto impiegando un set minimale di sensori biometrici non invasivi. In particolare, si vede come la biometria può migliorare la vita del team di sviluppatori software includendo le informazioni emotive raccolte quotidianamente oppure alla fine di una iterazione o di uno sprint.

	Hold-out setting Train: 90% + LOO cross validation Test: 10% (10 times)						Leave-one-subject out setting Train: all-1 subject + LOO cross validation Test: 1 held-out subject (23 subjects)					
Valence												
Devices	**Alg.**	**Prec**	**Rec**	**F1**	**Accuracy**	**stdev**	**Alg.**	**Prec**	**Rec**	**F1**	**Accuracy**	**stdev**
Full set	knn	.68 (+.34)	.60 (+.10)	.60 (+.19)	.72 (+.04)	.12	svm	.48 (+.14)	.62 (+.12)	.53 (+.12)	.69 (+.01)	.25
Empatica	knn	.70 (+.36)	.59 (+.09)	.59 (+.18)	.71 (+.03)	.07	svm	.45 (+.11)	.61 (+.11)	.50 (+.09)	.68 (–)	.27
Brainlink	rf	.54 (+.20)	.54 (+.04)	.52 (+.11)	.66 (-.02)	.07	mlp	.66 (+.32)	.64 (+.14)	.64 (+.23)	.71 (+.03)	.22
Baseline		*.34*	*.50*	*.41*	*.68*			*.34*	*.50*	*.41*	*.68*	
Arousal												
Full set	rf	.62 (+.31)	.61 (+.11)	.59 (+.21)	.65 (+.04)	.05	svm	.46 (+.15)	.59 (+.09)	.50 (+.12)	.61 (+.05)	.25
Empatica	knn	.67 (+.36)	.58 (+.08)	.55 (+.17)	.65 (+.04)	.10	J48	.40 (+.09)	.59 (+.09)	.49 (+.11)	.62 (–)	.25
Brainlink	rf	.66 (+.35)	.59 (+.09)	.58 (+.20)	.63 (+.01)	.12	nb	.62 (+.31)	.63 (+.13)	.61 (+.23)	.63 (+.01)	.17
Baseline		*.31*	*.50*	*.38*	*.62*			*.31*	*.50*	*.38*	*.62*	

Figura 3.2: *Migliori performance del classificatore per Valence e Arousal* [13]

3.1.1 Creazione di un gold standard con Affectiva

I video raccolti durante la sperimentazione originale [13] sono stati analizzati in uno studio successivo [14] per la creazione di un gold standard, e la successiva realizzazione di un classificatore.

Il dataset è costituito da 20 video, poiché in alcuni non è stato possibile riconoscere bene il volto, di questi ne sono stati considerati solo 12. I video sono stati analizzati tramite il tool Affectiva, questa tecnologia permette alle applicazioni software di utilizzare una webcam per tracciare la mimica facciale del soggetto, misurando anche il livello di sorpresa, soddisfazione e confusione. A partire da queste misurazioni, Affectiva calcola il punteggio di *Valence*, che varia da 100 a -100 e di *engagement*, questo deriva da una media pesata dei valori di alcune espressioni*: brow raise, brow furrow, nose wrinkle, lip corner depressor, chin raise, lip pucker, lip press, mouth open, lip suck, smile*.

L'utilizzo di Affectiva ha permesso di identificare gli episodi caratterizzati da valori molto alti o molto bassi di *Valence*, in corrispondenza dei quali sono state selezionate le feature biometriche. Le predizioni di Affectiva in combinazione con il questionario SAM, compilato dai partecipanti per il rilevamento delle emozioni costituiscono il *gold standard.*

3.1.2 Rielaborazione del Dataset

Il dataset è composto dalle misurazioni biometriche raccolte nello studio descritto nel capitolo 3.1 utilizzando Empatica E4 supportato dal tool Affectiva. Tuttavia, la sperimentazione pregressa prende in considerazione solo i valori *positive* e *negative*, per questa indagine il dataset include anche i valori *neutral*.

Il dataset conta di 17710 misurazioni biometriche di cui 1581 riportano *Valence* positiva mentre 1420 negativa, i restanti 14709 sono neutrali. C'è una significativa sproporzione del numero di esempi relativi a ciascuna classe, infatti i valori *positive* e *negative* sono quantitativamente inferiori rispetto al numero di quelli *neutral*: il dataset risulta, quindi, sbilanciato verso questi ultimi. Per riequilibrarlo sono stati aggiunti un numero di *neutral* crescente, scelti in maniera casuale. In questo modo i dataset risultano sei, in ognuno i *neutral* presenti sono multipli di 1400: primo dataset 1400, secondo 2800, terzo 4200, così crescendo fino a 7000 del sesto dataset.

Questi costituiscono gli input per il modello di classificazione monolitico, la cui struttura è visibile in figura 4.1.

Per il classificatore in pipeline, visibile in figura 4.2, il primo sotto-modello considera come input gli stessi dataset, dove la *Valence positive* e quella *negative* sono state sostituite con "*not neutral*". Quindi, in questo set ci saranno solo due etichette per la *Valence*: "*neutral*" e "*not neutral*". Al contrario, per il secondo sotto-modello, il dataset esclude l'etichetta *neutral* e utilizza solo le due etichette *positive* e *negative*.
Si è adottando il metodo "*hold-out*", il quale prevede che il set di dati sia suddiviso in due sottoinsiemi, uno funzionale all'addestramento del classificatore, l'altro alla validazione. Ognuno dei dataset è stato suddiviso in training e testing; le percentuali considerate per la distribuzione dei dati sono 70, 80 e 90 per il set di training e rispettivamente 30, 20 e 10 per il set di testing.

3.2 Studio sul campo

L'obiettivo dello studio sul campo è in linea con quello di questo studio: migliorare la cognizione delle emozioni degli sviluppatori attraverso sensori biometrici. Nello specifico, il proposito finale è quello di riconoscere al meglio le emozioni negative al fine di dare alcuni suggerimenti come ad esempio "fai una pausa".

3.2.1 Raccolta dati

I dati raccolti durante lo studio sul campo sono relativi alle misurazioni biometriche e ai self report degli sviluppatori software partecipanti alla sperimentazione. Sono stati utilizzati il sensore Empatica E4 perché semplice e poco invasivo, e il software *WorkAnalytics* per la raccolta dei self-report. I dati sono stati collezionati per due settimane durante le quali i partecipanti indossavano il braccialetto Empatica E4 accendendolo e spegnendolo all'inizio e alla fine della giornata lavorativa, e rispondevano a domande pop-up notificate da *WorkAnalytics*. In particolare, gli sviluppatori davano brevi informazioni sull'ultima attività svolta tra le seguenti: *coding, reading/writing e-mails, helping colleagues, networking*. Inoltre, fornivano l'auto-valutazione delle due dimensioni del modello circonflesso di Russel [2]: *Valence* e *Arousal*, attraverso il questionario Self-Assessment Manikin (SAM).

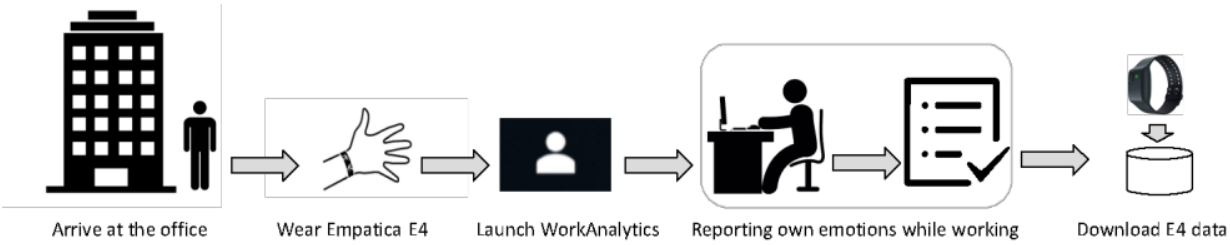


Figura 3.3: *Workflow della sperimentazione* [35]

3.2.2 Creazione di un gold standard con Self-Report

Il gold standard per i dati raccolti durante lo studio sul campo è rappresentato dai self-report volti ad indagare gli stati emotivi elicitati dalle attività svolte dai partecipanti. Come detto nel paragrafo 3.2.1, durante l'interruzione i partecipanti rispondono a due domande: la prima riguardante l'attività svolta, in modo che sia possibile ricavare il contesto in cui l'emozione si manifesta; la seconda relativa al livello di *Valence* e *Arousal.* Quest'ultima è rappresentata utilizzando una scala Likert a nove punti, tramite il Self-Assessment Manikin (SAM). Il SAM è uno strumento costituito da tre scale: una per la *Valence*, una per l'*Arousal* e, infine, una per la dimensione del controllo. La particolarità di tale strumento sta nel fatto che la scala di valori non è espressa con parole testuali, ma attraverso immagini congiunte a valori numerici (da uno a nove). Per ogni dimensione affettiva, sono definite cinque figure: la *Valence* è raffigurata a partire da una figura sorridente sino ad una infelice, mentre l'*Arousal* da una figura eccitata sino ad una che esprime rilassatezza.

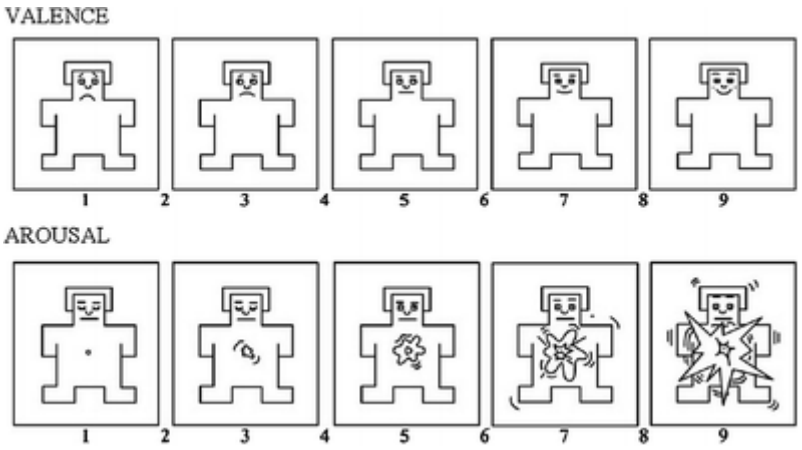


Figura 3.4: *Self-Assessment Manikin*, adattata da [36]

3.2.3 Rielaborazione del Dataset

Durante le due settimane di studio sul campo è stata raccolta una grande mole di dati biometrici che includono sia la valutazione della *Valence* sia dell*'Arousal*. Si è deciso di selezionare delle "finestre temporali", per l'estrazione di ciascuna feature in modo da poter trasformare i flussi di dati continui delle feature in variabili discrete. Le finestre temporali selezionate corrispondono a 10 secondi e 3 minuti. [28] Dal campionamento secondo questi intervalli temporali risulta la seguente distribuzione:

Valence					
Time interval		*Negative*	*Neutral*	*Positive*	Overall
10 seconds	Overall	83 (11%)	234 (31%)	**442 (58%)**	759
	Training (90%)	75 (11%)	211 (31%)	398 (58%)	684
	Test (10%)	8 (11%)	23 (31%)	44 (58%)	75
3 minutes	Overall	82 (11%)	232 (31%)	**435 (58%)**	749
	Training (90%)	74 (11%)	209 (31%)	392 (58%)	675
	Test (10%)	8 (11%)	23(31%)	43 (58%)	74
Arousal					
Time interval		*High*	*Neutral*	*Low*	Overall
10 seconds	Overall	183 (24%)	**303 (40%)**	273 (36%)	759
	Training (90%)	165 (24%)	273 (68%)	245 (36%)	684
	Test (10%)	18 (24%)	30 (68%)	27 (36%)	75
3 minutes	Overall	182 (24%)	**297 (40%)**	270 (36%)	749
	Training (90%)	164(24%)	267 (68%)	243 (36%)	674
	Test (10%)	18 (24%)	30 (68%)	27 (36%)	75

Tabella 3.1 *Distribuzione del dataset (Field Study)*

Dalle tabelle si evince lo squilibrio tra le classi del dataset: per la *Valence* si ha una preponderanza di casi *positive,* mentre per *Arousal* di casi *neutral*. Per bilanciare il dataset le modalità più utilizzate sono i *sampling method,* ossia metodi di campionamento che permettono di modificare il dataset di partenza in modo da riequilibrarlo. Tra questi si è deciso di applicare l'algoritmo SMOTE. Il termine SMOTE [29] è l'abbreviazione di *Synthetic Minority Oversampling Technique* ed è un algoritmo che appartiene agli *Oversampling method*. Attualmente questo metodo è forse il più utilizzato perché, anziché replicare semplicemente i campioni della classe meno rappresentata, ne introduce di nuovi. Questi ultimi sono creati dall'interpolazione di vari elementi presenti tra la classe che si vuole riprodurre. Sebbene, infatti, SMOTE non crei semplicemente copie dei campioni già esistenti, si è voluto verificare che i valori di accuratezza ottenuti non fossero *overoptimistic*. Si è, dunque, proceduto a ribilanciare il training set totale, in modo che tutte le classi constassero del 33% del training set.

5. Costruzione dei classificatori

Per apprendimento supervisionato si intende la metodologia di machine learning che elabora automaticamente previsioni sui valori di uscita di un sistema rispetto ad un input sulla base di una serie di esempi ideali, costituiti da coppie di *input* e di *output*, che vengono inizialmente forniti.

4.1 Feature

L'estrazione delle feature è un processo che permette di derivare l'informazione utile contenuta in grandi quantità di dati. Attraverso essa, sono stati determinati gli attributi che maggiormente descrivono le misurazioni ricavate dai sensori biometrici. A seguito dello studio condotto da Girardi et al. [12] si sono considerate solo le misurazioni estratte tramite Empatica E4, in quanto quelle derivate dal caschetto Brainlink impattano negativamente, seppur di poco, le performance del classificatore, inoltre non sono compatibili con il dataset collezionato durante lo studio sul campo.
Per quanto riguarda la prima sperimentazione descritta in questo lavoro di tesi, sono stati presi in considerazione i segnali raccolti nei 10 secondi prima dell'interruzione dei soggetti. Per sincronizzare le misurazioni dei segnali biometrici con l'autovalutazione emotiva, si sono salvati il timestamp dell'interruzione (t_interruption), il timestamp per l'intervallo di tempo relativo ad ogni interruzione, ovvero 10 secondi prima del self-report (t_start) ed è stato selezionato ciascun campione del segnale registrato tra t_start e t_interruption. Il segnale è stato normalizzato in base alla baseline di ciascun partecipante utilizzando Z-score [12]. La baseline è calcolata avvalendosi degli ultimi 30 secondi dei video elicitanti, mostrati prima del task. Per massimizzare le informazioni sul segnale e ridurre il rumore causato da movimenti, si sono applicate più tecniche di *filtering*. [12]
Relativamente al segnale di conduttanza cutanea (EDA), esso è costituito dalla somma di due componenti: la componente Fasica che rappresenta le attivazioni delle ghiandole sudoripare indotte dagli stimoli, e la componente Tonica che invece rappresenta le variazioni di umidificazione degli strati della pelle. Le due costituenti sono state estratte come feature tramite l'algoritmo cvxEDA, mentre per il segnale BVP sono state estratte le bande di frequenza in diversi intervalli usando un algoritmo di *pass-band filtering*. Per la frequenza cardiaca (HR), come per le altre misure fisiologiche, sono state ricavate

feature statistiche come media, minimo, massimo, varianza e deviazione standard rispetto alla baseline. Le feature sono visibili nella tabella seguente.

EDA	Media componente tonica AUC componente fasica Minimo, massimo, media, somma dei picchi della componente fasica
BVP	Minimo, massimo, somma ampiezza dei picchi, media ampiezza picchi (differenza tra baseline e task)
HR	Media, deviazione standard (differenza tra baseline e task)

Tabella 4.1: *Feature utilizzate per l'addestramento del classificatore*

In un secondo esperimento, sono state aggiunte nuove feature al fine di replicare la classificazione eseguita nello studio di Jacques et al. [15]. Le feature considerate sono legate al segnale EDA ed includono la conduttanza cutanea (SC) in *microSiemens*, l'accelerometro a 3 assi e la temperatura.

La conduttanza cutanea nello specifico misura il grado di attivazione del sistema nervoso simpatico, che aumenta in caso di eventi stressanti o situazioni di emergenza, perciò può essere molto utile l'individuazione dei picchi del segnale al fine di determinare e prevenire condizioni di *burnout* e *turnover*. Per la computazione del segnale SC si è applicato un *low-pass filter* ad 1Hz e successivamente la normalizzazione tramite la formula: $\frac{SCi-\mu}{\max - min}$ che considera la media (μ), il minimo e il massimo di ogni partecipante. I picchi sono stati individuati sulla base di diversi criteri come l'ampiezza, la durata, e la forma del segnale, inoltre sono state incluse feature relative al numero di picchi, le misure statistiche, e l'area sotto la curva del segnale.

Per limitare il rumore a cui è sottoposto il segnale dell'EDA si è calcolato la *magnitudine* relativa alle misurazioni dell'accelerometro tramite la formula: $Mag = \sqrt{acc_x^2 + acc_y^2 + acc_z^2}$. Questa misura è stata normalizzata, sottratta ad 1 e poi il suo inverso moltiplicato per il segnale SC; lo scopo è quello di ridurre l'incremento dell'attività elettro-dermica dovuto a fattori diversi da stress, ansia e agitazione.

Queste feature hanno un'influenza negativa sulle performance del classificatore, infatti non sono state considerate rilevanti per la costruzione del modello.

4.2 Classificatori

Durante questo studio sono stati addestrati e valutati due diversi classificatori che, a partire dalle feature biometriche distinguono la valenza emotiva in tre classi: *positive*, *negative* e *neutral*, e l'attivazione emotiva in *high*, *low* e *neutral*

Il primo classificatore è rappresentato da un modello monolitico, il quale classifica le emozioni direttamente nelle tre categorie (Monolithic; Figura 4.1).

Il secondo classificatore prende in esame un modello in pipeline, composto a sua volta da due diversi classificatori che lavorano in sequenza, dove l'output del primo è di input al secondo.

Il primo sotto-modello del classificatore in pipeline (Model neutral - non neutral; Figura 4.2) distingue le feature in *neutral* e *not neutral*.

Le feature catalogate come *not neutral* sono l'input per il secondo sotto-modello (Model polarity; Figura 4.2), il quale le classifica in *positive* e *negative* per la *Valence* (*high* e *low* per l*'Arousal*).

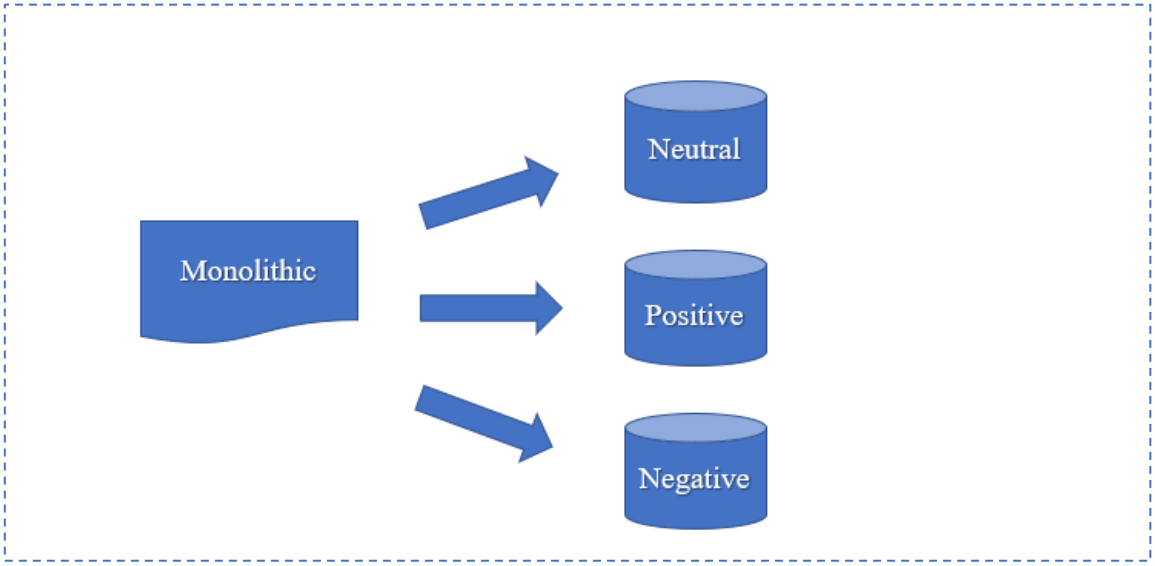


Figura 4.1: *Classificatore monolitico*

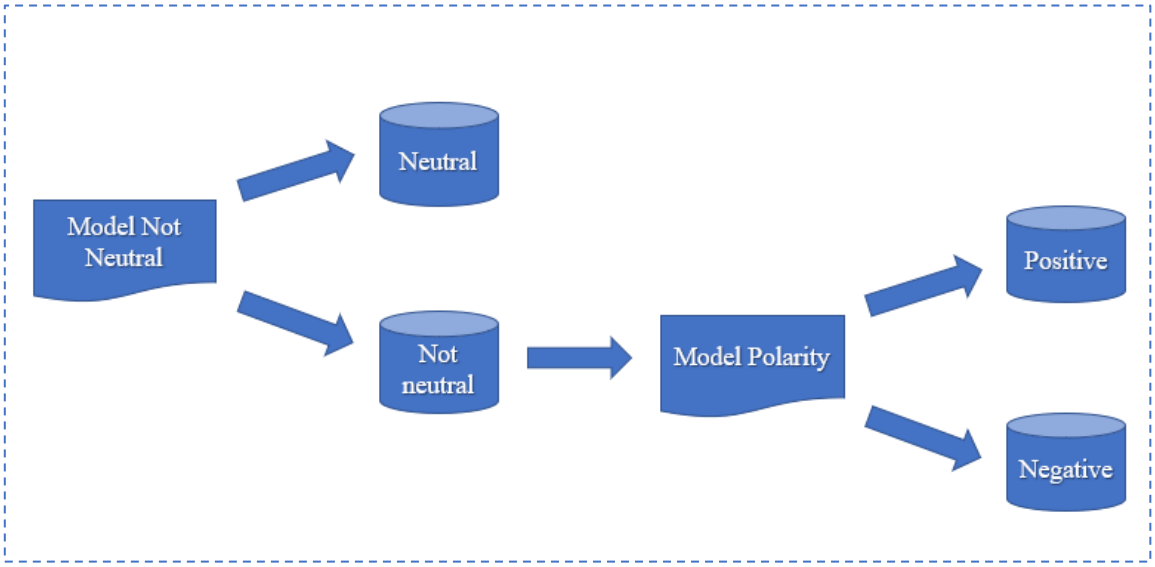


Figura 4.2: *Classificatore in pipeline*

4.3 Classificazione

Utilizzando il package "*Caret*" del linguaggio R, tutti i modelli descritti nel paragrafo precedente sono stati addestrati e valutati con le medesime tecniche. La scarsa disponibilità di dati ha richiesto l'applicazione della tecnica di validazione "*k-fold cross validation*", mentre gli algoritmi utilizzati sono stati il *random forest* (RF) e il *support vector machine* (SVM) e J48, quest'ultimo utilizzato unicamente sui dati raccolti durante lo studio sul campo.

4.3.1 K-fold cross validation

La *cross validation* è una tecnica utilizzata per valutare le performance di un classificatore. Il classificatore è addestrato su un sottoinsieme dei dati di input, chiamato *training set*, e validato attraverso il sottoinsieme complementare dei dati, chiamato *validation set.* L'idea di base è quella di incrociare ripetutamente il validation set e il training set, in modo tale che ogni istanza sia valutata.
Con la tecnica della *cross validation* si divide l'insieme dei dati in k partizioni (dette *fold*) delle stesse dimensioni, indipendenti tra loro. Ad ogni ripetizione della procedura una partizione viene utilizzata per il test e le rimanenti per il training, con cui si costruisce il modello, questa tecnica viene comunemente chiamata "*k-fold cross validation".*
Nel nostro caso il processo viene iterato dieci volte considerando k = 10, cambiando ogni volta il *fold* escluso così da saggiare tutti i *fold*.

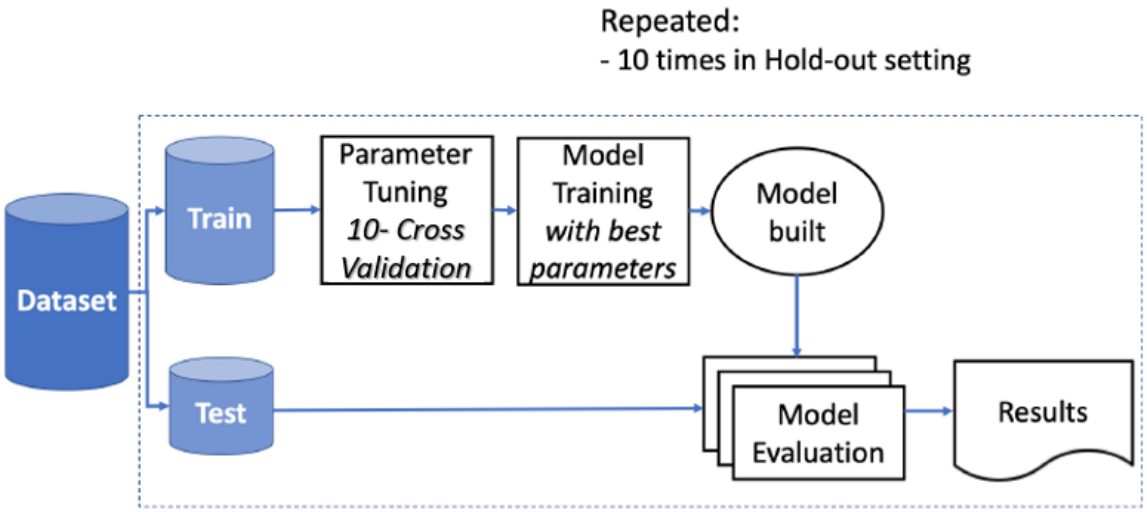


Figura 4.3: *Setting per la classificazione*

4.3.2 Algoritmi

Gli algoritmi utilizzati per la classificazione sono il *Random Forest* (RF) e il *Support Vector Machine* (SVM). Il Random Forest è un modello ottenuto dall'aggregazione tramite *bagging* di alberi di decisione. Esso è un meta-stimatore che si adatta ad una serie di alberi decisionali addestrati su vari sotto-campioni del dataset e utilizza la media di ogni singolo output di ogni albero per migliorare l'accuratezza predittiva e il controllo dell'*overfitting*. Il Support Vector Machine, invece, ha l'obiettivo di identificare l'iperpiano che meglio divide i vettori di supporto in classi. I nuovi esempi sono mappati nell'iperpiano e le predizioni delle categorie alle quali appartengono viene fatta individuando il lato dell'iperpiano nel quale ricadono. L'algoritmo SVM ottiene la massima efficacia nei problemi di classificazione binari.
La scelta di questi due algoritmi è dovuta ai risultati ottenuti in uno studio simile condotto da Jaques et al. [15], il quale indica questi come i migliori modelli nella classificazione degli stati emotivi a partire da feature biometriche. Al contrario, nel nostro studio i classificatori costruiti utilizzando l'algoritmo SVM hanno prodotto risultati poco considerevoli, in quanto su dataset sbilanciati tendono a predire tutte le emozioni come appartenenti ad un'unica categoria. Il confronto tra le macro-metriche di quest'ultimo classificatore con il classificatore che utilizza RF sono visibili in Tabella 4.2.

Classifier	Algorithm	macroPrecision	macroRecall	macroF1	accuracy
Best monolithic (90-10, 2800 neutral)	RF	0.622517	0.5980323	0.6071705	0.6275862
Best pipeline (90-10, 1400 neutral)	RF	0.6619297	0.6713224	0.6552936	0.6772727
Best pipeline (90- 10, 2800 neutral)	SVM	0,3809496	0,3915332	0,3542097	0,3810345
Best monolithic (90-10, 1400 neutral)	SVM	0,4237315	0,41359	0,3891002	0,425

Tabella 4.2: *migliori performance dei classificatori sul test set*

4.3.2 Metriche di valutazione

L'utilizzo della "*10-cross Validation"* con 10 ripetizioni ha prodotto 10 modelli addestrati per ogni dataset, di cui sono state calcolate le metriche di *precision*, *recall*, *F1*, e *accuracy* per le classi di *positive*, *negative* e *neutral*, e *high*, *low* e *neutral*.

La *precision* è il rapporto tra il numero delle previsioni corrette di una classe (veri positivi) sul totale delle volte che il modello lo prevede. Il *recall* rappresenta la sensibilità del modello, è il rapporto tra le previsioni corrette per una classe sul totale dei casi in cui si verifica effettivamente. L'*F1* è calcolata come la media armonica tra *precision* e *recall*. [19]

Le performance complessive sono state calcolate adottando la micro-media come metrica aggregata e la macro-media, con cui *precision* e *recall* vengono prima valutati localmente per ogni classe e poi globalmente calcolando la media dei risultati per le diverse categorie.

Una volta addestrati i classificatori, per ognuno di essi è stata calcolata una metrica combinata, data dalla normalizzazione e conseguente somma delle metriche *F1* e *accuracy*.

Il classificatore che riportava il valore più alto secondo quest'ultima metrica è stato selezionato per la predizione dei risultati sul set di test. In seguito, sono state calcolate le metriche di *precision*, *recall*, *F1*, *accuracy* e le relative macro e micro-medie, per consentire un rapido confronto delle prestazioni complessive, inoltre, è stata calcolata la matrice di confusione. Per valutare l'accordo tra gli output dei classificatori e le gold label, è stata calcolata la *weighted Cohen's Kappa*, [16] che determina l'accordo tra due istanze seguendo uno schema pesato; lo schema adottato penalizza la confusione fra le emozioni positive e negative. Infatti, sono state calcolate le percentuali di disaccordo tra positive vs negative (severe disagreement), tra positive vs neutral, e negative vs neutral (mild disagreement). Per la computazione della *wighted Kappa* È stato assegnato un peso = 2 per il severe disagreement e un peso = 1 per il mild disagreement, come indicato nella tabella 4.3. Per l'interpretazione della κ, si è adottato il metodo proposto da Viera e Garret [17], i quali suggeriscono di considerare l'accordo come quasi assente se la $\kappa \leq 0$, leggero se $0.01 \leq \kappa \leq 0.20$, medio se $0.21 \leq \kappa \leq 0.40$, moderato se $0.41 \leq \kappa \leq 0.60$, sostanziale se $0.61 \leq \kappa \leq 0.80$ e praticamente perfetto $0.81 \leq \kappa \leq 1$.

	Negative	**Neutral**	**Positive**
Negative	0	1	2
Neutral	1	0	1
Positive	2	1	0

Tabella n.4: *Schema pesato per la computazione di κ.*

6. Riconoscimento di emozioni degli sviluppatori: studio in laboratorio

5.1 Analisi dei risultati

Si riportano i risultati relativi al processo di classificazione, descritto in 4.2.
Per il classificatore in pipeline il setting che riporta una migliore prestazione secondo le metriche descritte nel capitolo 4.2.2 è quello relativo alla suddivisione 90-10, addestrato sul dataset che include 1400 *neutral* utilizzando il *random forest*.
In merito all'accordo tra le *prediction* e le *gold label*, questo classificatore riporta i migliori risultati; infatti, la *weighted Cohen's Kappa* risulta "*sostanziale*" secondo il metodo proposto da Viera e Garret [17] ed è pari a 0.61. Inoltre, risulta il più sensibile nel riconoscere i *negative* dai *positive* e viceversa dato che il *recall* delle due classi risulta essere pari a 0.80 circa.
Lo stesso setting risulta migliore anche per il classificatore monolitico, il quale però, raggiunge performance leggermente più basse rispetto a quello in pipeline. Le metriche sono visibili nelle tabelle 5.1 e 5.2.

Performance's monolithic classifier (2800 neutral, 90-10)				
Confusion Matrix:				
	Predicted			
Actual		*negative*	*neutral*	*positive*
	negative	78	41	23
	neutral	28	205	47
	positive	9	68	81
Metrics per class:				
	Precision	Recall	F1	Accuracy
negative	0.6782609	0.5492958	0.6070039	0.6275862
neutral	0.6528662	0.7321429	0.6902357	0.6275862
positive	0.5364238	0.5126582	0.5242718	0.6275862
Metrics macro:				
	MacroPrecision	MacroRecall	MacroF1	Accuracy
	0.622517	0.5980323	0.6071705	0.6275862
Overall (Micro-average): 0.6275862				
Weighted Kappa: 0.42				
Perfect agreement cases: 58%				
Severe disagreement cases: 11%				
Mild disagreement cases: 31%				

Tabella 5.1: Performance del classificatore monolitico addestrato con RF, 2800 neutral, setting 90-10

Performance's pipeline classifier (1400 neutral, 90-10)				
Confusion Matrix:				
	Predicted			
Actual		*negative*	*neutral*	*positive*
	negative	120	14	8
	neutral	45	52	43
	positive	9	23	126
Metrics per class:				
	Precision	Recall	F1	Accuracy
negative	0.6896552	0.8450704	0.7594937	0.6772727
neutral	0.5842697	0.3714286	0.4541485	0.6772727
positive	0.7118644	0.7974684	0.7522388	0.6772727
Metrics macro:				
	macroPrecision	macroRecall	macroF1	accuracy
	0.6619297	0.6713224	0.6552936	0.6772727
Overall (Micro-average): 0.6772727				
weighted kappa: 0.61				
Perfect agreement cases: 68%				
Severe disagreement cases: 4%				
Mild disagreement cases: 28%				

Tabella 5.2: *Performance del classificatore in pipeline addestrato con RF, 1400 neutral, setting 90-10*

5.2 Utilizzo del classificatore per predire emozioni di sviluppatori software

Il classificatore con le performance migliori (il *best model*) è stato utilizzato per predire le emozioni di sviluppatori software impiegati in diverse aziende. I dati utilizzati sono quelli ricavati durante lo studio sul campo effettuato dalla dott.ssa Daniela Girardi (Si veda capitolo 3). Il dataset si presenta adeguato alle predizioni emotive tramite il *best model* addestrato in questo studio di tesi; la classificazione però non ha ottenuto i risultati sperati in quanto le performance risultano poco significative. [Tabella 5.3]

	Positive	Negative	Neutral	
Positive	4	6	4	
Negative	4	5	6	
	True Positive	True Negative	False Positive	False Negative
Positive	4	11	4	10
Negative	5	8	6	10
	Precision	Recall	F1	Accuracy
Positive	0,5	0,285714	0,363636	
Negative	0,454545	0,333333	0,384615	
Macro	0,477273	0,309524	0,374126	0,310345

Tabella 5.3: *Metriche del best model utilizzato sul dataset del field study*

La motivazione di tali performance si ritrova nel fatto che ogni soggetto ha una baseline emotiva diversa, ovvero le reazioni fisiologiche agli eventi esterni sono diverse da individuo a individuo. Per questo, il classificatore addestrato su misurazioni riguardanti gli studenti può avere scarse performance su dati rilevati dagli sviluppatori software collocati in un contesto lavorativo e non sperimentale. Inoltre, i dipendenti, a differenza degli studenti, svolgono diverse attività oltre a scrivere codice; come rispondere ad e-mail, comunicare con i colleghi, partecipare ai meeting.

5.3 Discussione

Le emozioni sono costituite da più componenti: prima di tutto mutamenti fisiologici, come la frequenza cardiaca e la risposta galvanica della pelle; in secondo luogo la cosiddetta "preparazione all'azione", ad esempio le espressioni facciali; terzo le emozioni coinvolgono l'esperienza cosciente, cioè come un'emozione ci fa *sentire.* [20] Secondo gli studi di Gross e Pekrun [21, 22] tra queste componenti ci deve essere un certo indice di accordo; di conseguenza, per migliorare la comprensione teorica e pratica delle emozioni è in crescita l'importanza attribuita alla necessità di andare oltre le semplici auto-valutazioni emotive. [23]

Una delle questioni più dibattute nell'area di ricerca del *Affective Learning* riguarda l'accordo tra le differenti componenti delle emozioni; cioè se queste debbano restituire informazioni emozionali uguali o diverse. Secondo l'esempio riportato da Pekrun [22] la coerenza tra le componenti emozionali esiste. Infatti, immaginiamo di essere uno studente poco prima dell'esame: le mani sudano, il cuore batte più velocemente, ci sentiamo agitati, preoccupati per il fallimento e vorremmo fuggire. Le sensazioni descritte mostrano una risposta coerente fra le componenti delle espressioni emozionali: queste includono l'aspetto affettivo (agitazione), cognitivo (preoccupazione per il fallimento), fisiologico (aumento del battito cardiaco e dell'attività sudoripara), motivazione (impulso alla fuga) ed espressivo (espressioni di preoccupazione). Altri ricercatori, invece, affermano come uno stretto accoppiamento tra tutte le componenti emozionali non esista necessariamente. [24]

L'approccio che si è utilizzato in questo studio è di tipo multimodale, cioè la misurazione delle emozioni è avvenuta con strumenti di natura diversa; ciò ci ha permesso di osservare se i metodi utilizzati fossero tutti in accordo tra di loro. Il confronto è avvenuto tra l'auto-valutazione emotiva (data dal questionario SAM) e le predizioni del classificatore basato sulle espressioni del volto. I video che ritraggono gli studenti sono stati analizzati e messi a confronto con le auto-valutazioni emotive riportate dagli stessi (i *self-report*). Le espressioni che gli studenti assumono sono prevalentemente di attenzione e concentrazione, che vengono classificate dal tool Affectiva come neutrali, ma nel caso di un *self-report* con *Valence* positiva non vi è corrispondenza con l'output di Affectiva. Inoltre non si ritrova la corrispondenza inversa, ovvero, nel caso in cui il partecipante suggerisca qualche espressione serena o fiduciosa, non vi è un *self-report* con *Valence*

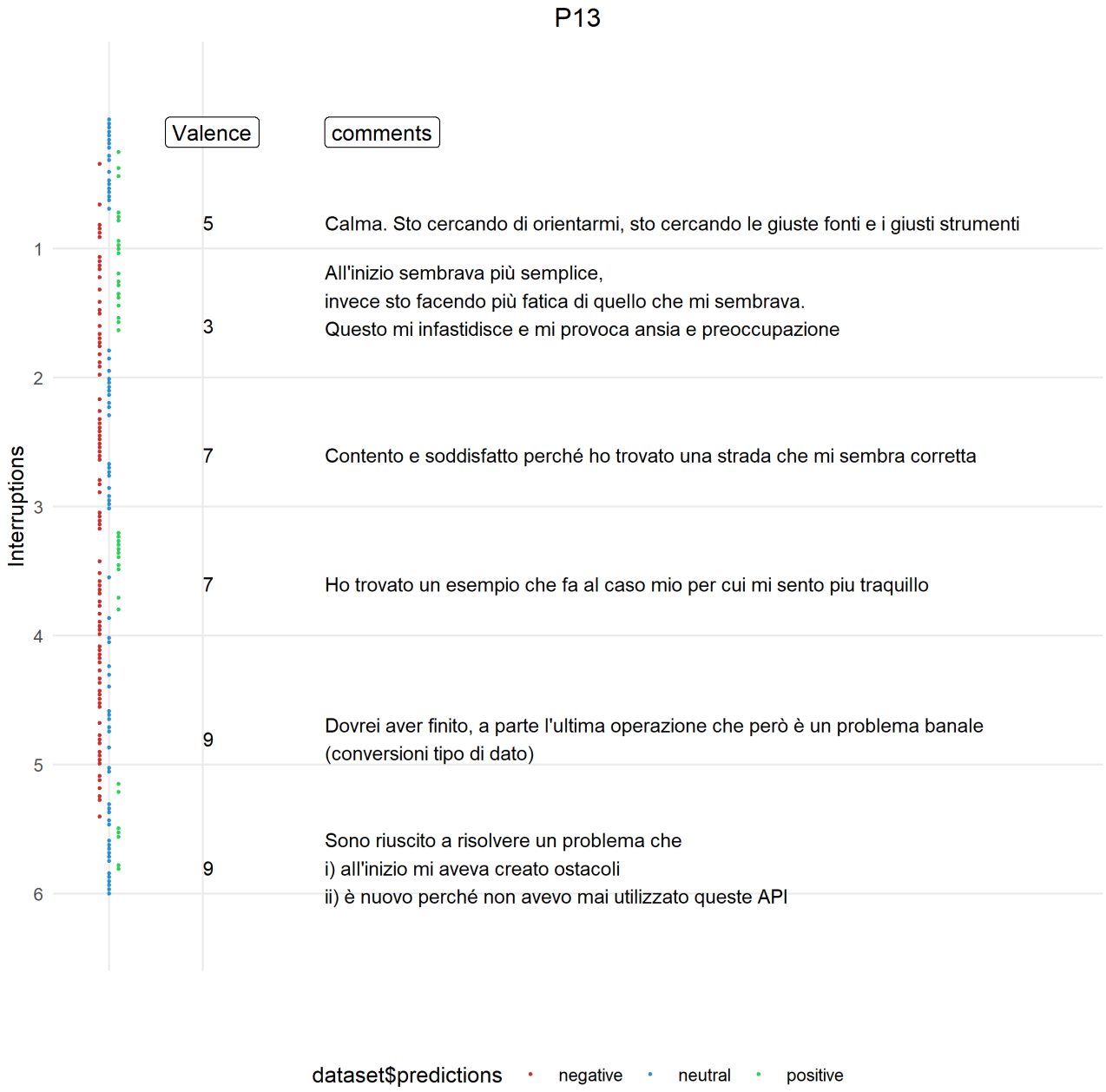
positiva. Questo si spiega con il fatto che il *self-report* è il risultato di una valutazione cognitiva, mentre l'espressione facciale è il riflesso di un'emozione istantanea.

Harley et al. [23] hanno condotto uno studio multimodale, simile a quello descritto in questo lavoro di tesi, con l'obiettivo di valutare gli strumenti di rilevamento di emozioni (riconoscimento facciale, self-report, GSR) e il loro accordo, considerando non la *Valence*, ma l'*Arousal* come dimensione di riferimento. Questi hanno documentato come, nell'ambito del machine learning, tutti i confronti sono fatti considerando il *ground truth* (cioè confrontandolo con dati di training), mentre i metodi utilizzati per le predizioni non vengono confrontati tra di loro; infatti in uno studio multimodale che considera i *self-report* come *ground truth* è raro che vi si trovi l'accordo tra i risultati ottenuti utilizzando criteri diversi. Per esempio, non si ritrova un accordo tra i risultati ottenuti monitorando la postura e l'attività elettrodermica oppure la frequenza cardiaca e l'espressione facciale [23].

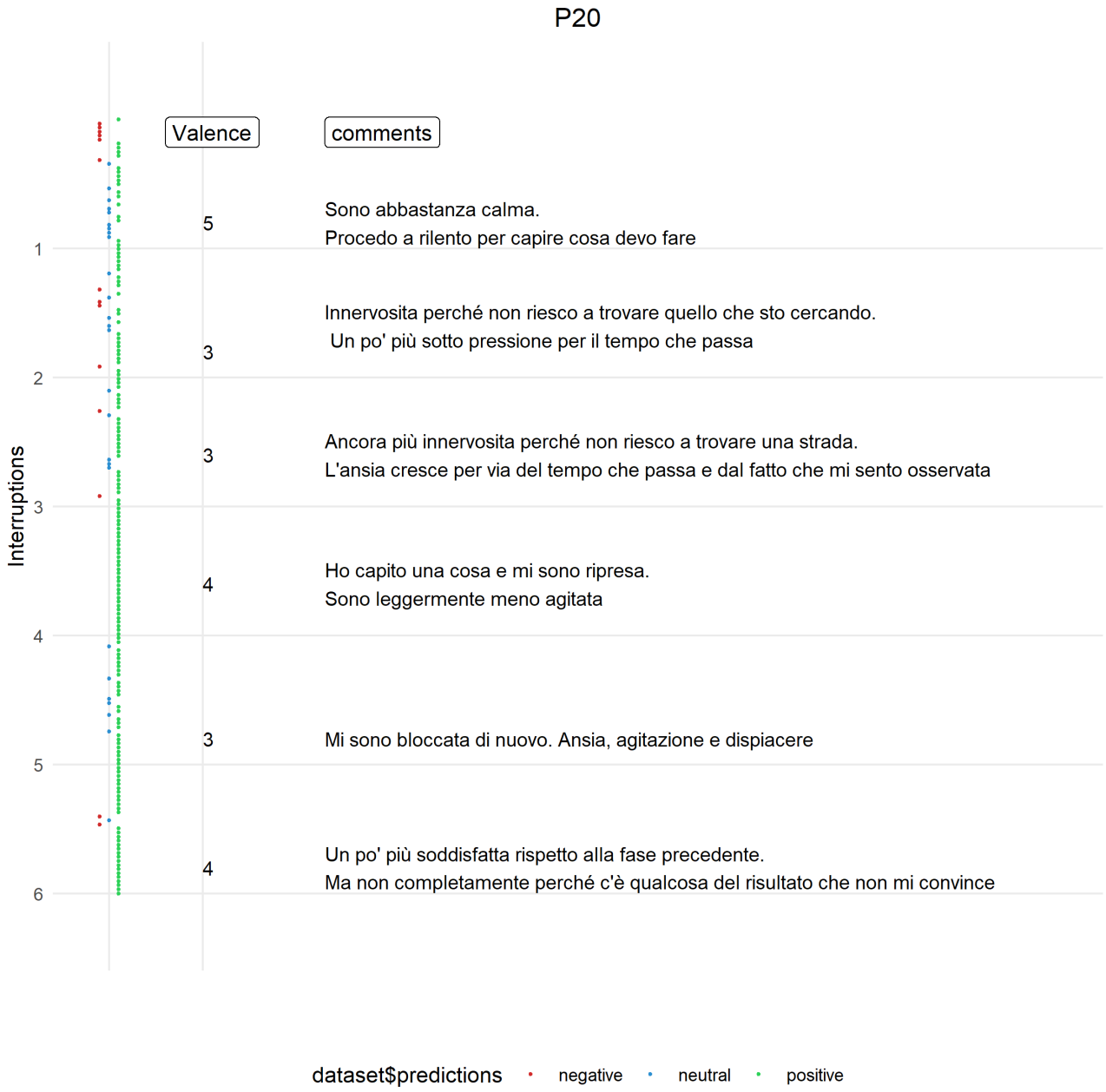
Questo può essere spiegato con il fatto che alcune modalità garantiscono predizioni più accurate per alcuni tipi di emozioni: per esempio è più facile rilevare la *Valence* con l'utilizzo del riconoscimento del volto e l'*Arousal* tramite l'attività elettro-dermica. Purtroppo però, non è possibile ritrovare quali sono le modalità maggiormente complementari per rilevare determinate emozioni perché non è possibile misurare l'accordo tra le modalità. [23]

Inoltre, un'altra motivazione plausibile al fatto che non si ritrovi accordo tra gli strumenti utilizzati in questo studio, Affectiva e *self-report*, riguarda le autovalutazioni emotive da parte degli studenti, queste infatti sono fatte prevalentemente in base all'*Arousal* ("sono più agitato/sono più calmo"), che non è rilevata da Affectiva.

Nei grafici successivi sono riportate: le predizioni del classificatore eseguite durante i venti minuti prima di ogni interruzione (punto verde: emozione positiva; rosso: emozione negativa; blu: emozione neutra), la *Valence* auto-valutata e il commento riguardante l'esecuzione del task scritto dal partecipante.



In questo esempio durante i primi cinque minuti, il partecipante n.13 riporta una *Valence* neutra in linea con le predizioni del classificatore. Nelle successive interruzioni (2-5), invece, non vi è corrispondenza tra le predizioni e i commenti; infatti, dove il partecipante dice di essere "contento e soddisfatto" il classificatore predice *negative.* Nell'ultima interruzione il classificatore restituisce *neutral*, mentre il partecipante riporta il più alto livello di *Valence*.



Il partecipante n.20 riporta prevalentemente emozioni negative (“ansia, agitazione, dispiacere”) che sono travisate dal classificatore, il quale predice prevalentemente emozioni positive.

7. Riconoscimento di emozioni degli sviluppatori: studio sul campo

6.1 Analisi dei risultati

Per quanto riguarda la seconda sperimentazione di questo studio di tesi, si riportano i risultati ottenuti con il dataset relativo allo studio sul campo, dopo aver applicato il processo di classificazione descritto nella sezione 4. Il setting utilizzato per questa classificazione è esclusivamente 90-10 coerentemente alla sperimentazione avvenuta in laboratorio, in quanto risulta essere la miglior ripartizione.

Result Hold-out setting (Valence)					
Classifier	Alg	macroPrec	macroRec	macroF	Accuracy
Baseline	RF	0.20	0.33	0.25	0.59
3 labels (10 seconds) - best	**RF**	**0.44**	**0.46**	**0.45**	**0.55**
3 labels (10 seconds) - average	RF	0.39	0.38	0.38	0.50
3 labels (3 minutes) - best	RF	0.47	0.49	0.48	0.57
3 labels (3 minutes) - average	RF	0.36	0.36	0.36	0.46
Result Hold-out setting (Arousal)					
Classifier	Alg	macroPrec	macroRec	macroF	Accuracy
Baseline	RF	0.13	0.33	0.19	0.40
3 labels (10 seconds) - best	**RF**	**0.56**	**0.58**	**0.56**	**0.56**
3 labels (10 seconds) - average	RF	0.45	0.44	0.44	0.45
3 labels (3 minutes) - best	RF	0.56	0.54	0.55	0.54
3 labels (3 minutes) - average	RF	0.44	0.43	0.43	0.44

Dalle tabelle si evince che il "*best model*" è rappresentato dal classificatore monolitico sia per la *Valence* che per l'*Arousal*, con algoritmo Random Forest. Si noti come questi migliorino le performance rispetto ad un classificatore naïve.

6.2 Discussione

Le seguenti tabelle sono relative alle performance del "*best model*" per ogni classe.

Performance by class for best valence classifier (3 labels; 10 seconds)				
		Predicted		
		negative	*neutral*	*positive*
Actual	*negative*	3	2	3
	neutral	4	9	10
	positive	5	8	30
	Precision	Recall	F	Accuracy
negative	0.25	0.38	0.30	
neutral	0.47	0.39	0.43	
positive	0.70	0.70	0.70	
Macro-avg	0.47	0.48	0.48	0.57
Micro-avg	0.57	0.57	0.57	0.57

Tabella 6.1*: Performance per ogni classe del best model per la Valence*

Performance by class for best arousal classifier (3 labels; 10 seconds)				
		Predicted		
		high	*low*	*neutral*
Actual	*high*	12	3	3
	low	3	17	7
	neutral	7	10	13
	Precision	Recall	F	Accuracy
high	0.55	0.67	0.60	
low	0.57	0.63	0.60	
neutral	0.57	0.43	0.49	
Macro-avg	0.56	0.58	0.56	0.56
Micro-avg	0.56	0.56	0.56	0.56

Tabella 6.2: *Performance per ogni classe del best model per l'Arousal*

Per quanto riguarda la *Valence* la classe dei negative è quella di maggiore interesse dal punto di vista applicativo, ma allo stesso tempo quella meno rappresentata nel dataset. Di conseguenza, il classificatore classifica erroneamente i negative confondendoli con i positive o i neutral, mentre metà dei casi neutral è classificata come positive. Come per la *Valence* il problema principale è riuscire a riconoscere i casi di *Arousal* low, anch'essi meno rappresentati nel dataset. Questi vengono maggiormente confusi con i casi neutral, i quali vengono, invece, confusi con low e high.

8. Conclusioni

Con questo lavoro di tesi si è cercato di comprendere quale fosse la migliore configurazione per addestrare un classificatore che riconoscesse le emozioni in base alle tre classi relative alla *Valence*: *positive, negative, neutral* e alle tre classi relative all' *Arousal*: *high, low* e *neutral*.

Lo studio è composto di due sperimentazioni.

Per la prima sperimentazione, la raccolta dati ha adoperato i sensori biometrici come Empatica E4 e BrainLink, e acquisizione dell'espressione del volto tramite webcam. Le acquisizioni video sono state analizzate tramite il tool Affectiva che in base all'espressione facciale determina *Valence* ed *Engagement*. L'output del tool Affectiva congiuntamente alle autovalutazioni emotive riportate dai partecipanti ha costituito il *gold standard.* Infatti, l'annotazione tramite Affectiva è servita a identificare gli episodi caratterizzati da valori molto alti o molto bassi di *Valence* per poi procedere con l'estrazione delle feature biometriche corrispondenti a tali istanti di tempo. Il dataset è composto da queste misurazioni biometriche di cui 1581 con *Valence* positiva 1420 negativa, i restanti 14709 neutra, risulta, quindi, molto sbilanciato verso i *neutral*. Di conseguenza, il primo quesito di ricerca è stato quello di indagare quanti valori *neutral* dovessero essere considerati nel dataset. Si è deciso di creare più dataset includendo un numero di *neutral* crescenti, partendo da 1400 fino a 7000 per avere un bilanciamento con il numero di *negative*. Il secondo quesito riguarda il *setting* con cui addestrare e validare il classificatore, si sono considerati tre diversi *setting:* 90-10, 80-20, 70-30. Queste riguardano le percentuali dei dati che si trovano rispettivamente nel dataset di training e in quello di test. Per rispondere ai quesiti si sono addestrati due classificatori: uno monolitico che esegue la classificazione direttamene nelle tre label (*positive, negative, neutral*), il secondo in pipeline. Quest'ultimo è costituito a sua volta da due classificatori, uno riconosce le emozioni *neutral* e *not-neutral*, l'altro distingue i *not-neutral* in *positive* e *negative.*

Per questa prima sperimentazione il classificatore migliore risulta essere quello in pipeline addestrato su un dataset che include 1400 *neutral* sul *setting* 90-10. Successivamente si è cercato l'accordo tra le predizioni del classificatore e le autovalutazioni emotive riportate dagli studenti, si è visto, però come vi è incoerenza tra le due.

La seconda sperimentazione sfrutta i risultati della prima per poter costruire un classificatore in grado di riconoscere le emozioni di sviluppatori software, utilizzando il dataset collezionato durante lo studio sul campo condotto dalla dott.ssa Daniela Girardi. La raccolta dati impiega unicamente il sensore Empatica E4 e non il riconoscimento del volto come nella prima sperimentazione, di conseguenza il gold standard è costituito dai self-report e non da Affectiva. Il dataset è stato campionato utilizzando due diversi intervalli di tempo: 10 secondi e 3 minuti, in linea con lo studio condotto da Zuger et al. [28] Il classificatore realizzato rispecchia il setting di quello monolitico della prima sperimentazione con la differenza che questo predice anche l'*Arousal* oltre che la *Valence.* Il classificatore migliore, in questo caso, risulta essere quello che considera un campionamento di 10 secondi sia per quanto riguarda la *Valence* che per l'*Arousal*

In futuro si intende replicare lo studio addestrando i classificatori partecipante per partecipante. Il dataset sarà composto dalle misurazioni biometriche raccolte unicamente per un singolo partecipante, in maniera tale da calibrare la classificazione su ogni baseline emotiva.

Replicando lo studio sarà possibile valutare altri approcci, ad esempio individuando la combinazione di *Valence* e *Arousal* predicendo in quale quadrante del modello circomplesso si ritrova l'emozione.

9. Ringraziamenti

In primo luogo vorrei esprimere la mia gratitudine per la Prof.ssa Nicole Novielli, una guida fondamentale nella stesura di questo lavoro di tesi. In particolare, la ringrazio per la fiducia riposta in me sin dal primo momento e per avermi fatto partecipe di questo lavoro di ricerca.

Ringrazio il Prof. Filippo Lanubile per la massima professionalità e disponibilità, per i preziosi consigli e i calorosi incoraggiamenti.

Ringrazio la Dott.ssa Daniela Girardi per la sua gentilezza e la sua pazienza; si è sempre mostrata disponibile ad aiutarmi fornendomi tutti gli elementi necessari per lo svolgimento di questo lavoro di tesi.

10. Bibliografia

[1] J. LeDoux, "*Il cervello emotivo. Alle origini delle emozioni*", Baldini Castaldi Dalai Editore, Milano 2003, p. 24.

[2] J.A. Russell. "Core affect and the psychological construction of emotion". 2003

[3] D. Girardi, F. Lanubile, and N. Novielli, "*Emotion detection using noninvasive low cost sensors*," 2017 7th Int. Conf. Affect. Comput. Intell. Interact. ACII 2017, vol. 2018-Janua, no. April 2018, pp. 125–130, 2018.

[4] Un modello dimensionale delle emozioni: integrazione tra le neuroscienze dell'affettività, lo sviluppo cognitivo e la psicopatologia [online] Available: https://www.cognitivismo.com/2012/08/01/un-modello-dimensionale-delle-emozioni-integrazione-tra-le-neuroscienze-dellaffettivita-lo-sviluppo-cognitivo-e-la-psicopatologia/

[6] Daniela Girardi. *Classificazione dei livelli di attivazione e valenza emotiva mediante sensori biometrici.* 2015-2016

[7] Frequenza Cardiaca. [online] Available: https://it.wikipedia.org/wiki/Frequenza_cardiaca

[8] T. Fritz, A. Begel, S. C. Müller, S. Yigit-Elliott, and M. Züger. "*Using psycho-physiological measures to assess task difficulty in software development*". 36th International Conference on Software Engineering, pp. 402-413, 2014.

[9] N. Nourbakhsh, Y. Wang, F. Chen, and R. A. Calvo. *Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks*. 24th Australian Computer-Human Interaction Conf., OzCHI '12, pages 420–423, New York, NY,USA, 2012. ACM.

[10] S. C. Müller and T. Fritz, "*Stuck and frustrated or in flow and happy: Sensing developers' emotions and progress*," Proc. - Int. Conf. Softw. Eng., vol. 1, pp. 688–699, 2015.

[11] N. Novielli, F. Calefato, and F. Lanubile, "*The challenges of sentiment detection in the social programmer ecosystem,*" 7th Int. Work. Soc. Softw. Eng. SSE 2015 - Proc., no. May 2017, pp. 33–40, 2015.

[12] D. Girardi, N. Novielli, D. Fucci, F. Lanubile. "*Recognizing Developers' Emotions while Programming*". In Proceedings of the 42th International Conference on Software Engineering (ICSE 2020) October, 2020

[14] Silvia Manca. "*Rilevazione biometrica delle emozioni degli sviluppatori software: inclusione delle espressioni facciali."* 2018-2019

[15] N. Jaques, S. Taylor, A. Azaria, A. Ghandeharioun, A. Sano, R. Picard; "*Predicting students' happiness from physiology, phone, mobility, and behavioral data*"

[16] Cohen. 1968. Weighted kappa: *Nominal scale agreement provision for scaled disagreement or partial credit*. Psychological Bulletin, 70, 4, 213-220

[17] A.J. Viera, J.M. Garrett. 2005. Understanding interobserver agreement: the kappa statistic. Family Medicine, 37,5, 360–363.

[18] "Cross Validation." [Online]. Available: https://www.cs.cmu.edu/~schneide/tut5/node42.html. [Accessed: 24-06-2020].

[19] David L. Poole and Alan K. Mackworth. "*Artificial Intelligence: Foundations of Computational Agents*" 2nd Edition. Section 7.2.2 Types of Errors.

[20] M.W. Eysenck. "*Psicologia Generale*" 2006. pp. 351.

[21] James J. Gross "*The Future's So Bright, I Gotta Wear Shades*". *Emotion* 13, 359-365.

[22] Pekrun, R. (2006). "*The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice*." Educational Psychology Review, 18, 315–341.

[23] J. M. Harley , F. Bouchet, M. S. Hussain, R. Azevedo, R. Calvo. "*A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system*". Computers in Human Behavior 48 (2015) pp. 615–625

[24] S. D'Mello, R. Dale, A. Graesser "*Disequilibrium in the mind, disharmony in the body*" 23 May 2011

[25] Rosalind Picard. "*Affective Computing*", 1995.

[26]] NASA Software Reuse Working Group. Software reuse survey. [Online] Available: http://www.esdswg.com/ softwarereuse/Resources/library/working group doc%uments/ survey2005, 2005.

[27] B. Floyd, T. Santander, W. Weimer, "*Decoding the representation of code in the brain: An fMRI study of code review and expertise"* 39th International Conference on Software Engineering (ICSE 2017)

[28] M. Zuger, T. Fritz, A. Mayer "*Sensing Interruptibility in the Office: A Field Study on the Use of Biometric and Computer Interaction Sensors*" 2018

[29] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer (2002*), SMOTE: Synthetic Minority Over-sampling Technique* In Journal of Artificial Intelligence Reasearch p.16

[30] F. Fagerholm, T. Fritz "*Biometric Measurement in Software Engineering*" (2020) In Contemporary Empirical Methods in Software Engineering pp 151-172

[31] J. Siegmund, N. Peitek, C. Parnin, S. Apel, J. Hofmeister, C. Kästner, A. Begel, A. Bethmann, A. Brechmann. (2017) "*Measuring neural efficiency of program comprehension*." In: Proceedings of the 2017 11th joint meeting on foundations of software engineering (ESEC/FSE 2017). ACM, New York, pp 140–150.

[32] T. Busjahn, R. Bednarik, A. Begel, M. Crosby, J. H. Paterson, C. Schulte, B. Sharif, S. Tamm, (2015) "*Eye movements in code reading: relaxing the linear order"*. In: 2015 IEEE 23rd international conference on program comprehension, pp 255–265.

[33] A. Jbara, D. G. Feitelson (2017*) "How programmers read regular code: a controlled experiment using eye tracking*". Empir Softw Eng 22(3):1440–1477

[34] , B. Park, E. Jang, S. Kim, C. Huh, J. Sohn, Seven emotion recognition by means of particle swarm optimization on physiological signals Networking, Sensing and Control (ICNSC), 9th IEEE International Conference on Year 2012;

[35] D. Girardi, F. Lanubile, N. Novielli, L. Quaranta, A. Serebrenik, "Towards Recognizing the Emotions of Developers Using Biometrics: The Design of a Field Study". In Proceedings of the Fourth International Workshop on Emotion Awareness in Software Engineering (SEmotion 2019), May 28, 2019

[36] D. Graziotin, X. Wang, P. Abrahamsson "*Understanding the Affect of Developers: Theoretical Background and Guidelines for Psychoempirical Software Engineering"* 2015 ACM 7th International Workshop on Social Software Engineering (SSE 2015), pp. 25-32, 2015