Highlights

Sfruttare la keystrokes dynamics per la profilazione dell'utente: lingua parlata, provenienza geografica e corso di scrittura

Carmela Pia Senatore,Gennaro Capaldo

- Individuazione di modelli di classificazione in grado di discriminare gli utenti sulla base di dinamiche di tasti
- Ottenere una classificazione rispetto ai caratteri demografici
- Comprensione delle features più importanti nella classificazione

Sfruttare la keystrokes dynamics per la profilazione dell'utente: lingua parlata, provenienza geografica e corso di scrittura

Carmela Pia Senatore[a,c], Gennaro Capaldo[b,d]

[a] *Università degli Studi di Salerno, Dipartimento di Informatica*
[a] *Università degli Studi di Salerno, Dipartimento di Informatica*

ARTICLE INFO

Keywords:
keystrokes dynamics
classification
soft biometrics
SMOTE
demographic classification

ABSTRACT

L'identificazione di specifiche caratteristiche degli utenti online, quali genere, età, lingua parlata, area geografica e corso di scrittura, possono essere fondamentali in contesti di sicurezza e per sviluppare applicazioni commerciali. Un elemento chiave in questo contesto è la capacità di riconoscere e classificare le caratteristiche distintive di un utente attraverso l'analisi di tastiera.

Nel corso del progetto, partendo da un dataset disponibile pubblicamente e utilizzando algoritmi di machine learning, abbiamo analizzato le caratteristiche dinamiche relative alla pressione dei tasti e siamo riusciti a ottenere un buon tasso di accuratezza nel riconoscimento di caratteristiche demografiche. I risultati del seguente progetto mostrano che è possibile discriminare per area geografica di provenienza, lingua parlata e corso di scrittura avendo solo a disposizione le *caratteristiche di keystrokes* per utente. Sulla base dei modelli più performanti per ogni task, si è cercato di analizzare l'impatto delle variabili e di identificare quali tra esse siano particolarmente discriminanti.

1. Introduzione

Il campo della *keystroke dynamics* si colloca all'intersezione di numerose discipline quali la biometria, la cibernetica e la psicologia cognitiva. Questo ambito di studio si focalizza sull'analisi del comportamento di digitazione degli utenti come tratto biometrico comportamentale, esaminando aspetti come la velocità, il ritmo, e la pressione esercitata sui tasti durante la scrittura al computer. Tali caratteristiche vengono prese in considerazione dagli algoritmi di battitura volti a discriminare tra i soggetti. Ogni individuo, infatti, manifesta un pattern unico di digitazione, che può essere descritto attraverso una serie di metriche quantificabili. Queste peculiarità rendono la dinamica della digitazione non solo uno strumento per l'identificazione personale e la verifica di autenticità, ma anche un potenziale mezzo di classificazione basato su caratteristiche demografiche.

La ricerca sulla dinamica della digitazione ha guadagnato popolarità negli ultimi decenni grazie alla crescente necessità di sistemi di sicurezza non invasivi e facilmente integrabili. Inoltre, il continuo miglioramento delle tecnologie di sensori e di analisi dati ha permesso di esplorare nuovi modi di applicazione di queste tecniche. La semplicità con cui i dati di digitazione possono essere raccolti, unita alla loro natura intrinsecamente ricca di informazioni, offre un vasto campo di indagine per migliorare tanto la sicurezza informatica quanto la personalizzazione delle interazioni digitali.

Keystroke Dynamics per il Riconoscimento e la Verifica dell'Utente

Uno degli usi più comuni della dinamica della digitazione è il riconoscimento dell'utente. Le aziende e le organizzazioni possono utilizzare questi dati per verificare l'identità di un individuo in maniera discreta e continua, senza richiedere interventi diretti o periodici da parte dell'utente. Questo tipo di verifica può servire come un ulteriore livello di sicurezza, integrando o sostituendo metodi tradizionali come le password o i PIN. In questo contesto, la keystrokes dynamics agisce come una forma di firma invisibile, difficilmente replicabile da attori esterni.

La verifica dell'autenticità, invece, si occupa di controllare che l'interazione con un sistema sia effettivamente condotta dall'utente autorizzato in tutti i momenti di utilizzo. Gli approcci impiegati, di solito per la verifica della persona, analizzano le informazioni comportamentali (ad esempio gesti tattili, sequenze di tasti) memorizzate dai dispositivi durante la normale interazione utente-dispositivo e controllano l'identità dell'utente in background. [1]

Keystroke Dynamics per la Classificazione e la Discriminazione dei Soggetti

Il focus principale del presente lavoro, tuttavia, riguarda l'utilizzo della keystrokes dynamics per la classificazione dei soggetti secondo caratteristiche demografiche quali genere, età, lingua parlata e area geografica di provenienza. Questo approccio sfrutta le sottili differenze nei modi di digitazione che possono riflettere l'identità culturale, sociale e/o personale di un individuo. Ad esempio, le variazioni nella velocità e nella pressione dei tasti possono indicare non solo l'età e il genere, ma anche la familiarità con la tecnologia o l'influenza di specifici contesti linguistici e culturali.

Uno dei principali vantaggi della keystroke dynamics è che non richiede attrezzature o hardware speciali. Utilizza invece i semplici dati già generati dalla tastiera dell'utente durante la digitazione. Ciò significa che la keystroke dynamics può essere facilmente implementata in una varietà di impostazioni. Dati i bassi costi di implementazione, i quali richiedono la sola definizione di una struttura necessaria per la registrazione delle informazioni, questa risulta meno

*Corresponding author
c.senatore50@studenti.unisa.it (C.P. Senatore);
g.capaldo12@studenti.unisa.it (G. Capaldo)
ORCID(s):

costosa rispetto ai sistemi tradizionali per l'acquisizione di dati biometrici.

Un altro vantaggio è che non richiede la raccolta di dati sensibili. A differenza di altre forme di riconoscimento biometrico, come il riconoscimento facciale, la keystroke dynamics non utilizza informazioni personali o immagini, ma analizza semplicemente il modo in cui un utente digita, in maniera non invasiva e trasparente per lo stesso.

Un altro vantaggio è il fatto che questa analisi si basa sul tipo di comunicazione più utilizzato, il testo. Ciò significa che gli algoritmi di keystrokes dynamics possono essere applicati a un'ampia gamma di interazioni online, rendendolo uno strumento potente su diverse piattaforme e applicazioni.

Tuttavia, i problemi derivati dall'utilizzato di keystrokes data sono molteplici, di seguito i principali emergenti dalla letteratura.

Il primo problema riguarda la tipologia di tastiera utilizzata. I sistemi di rilevamento dei keystrokes su tastiere diverse possono produrre risultati diversi. È stato dimostrato che l'utilizzo della tastiera meccanica richiede, da parte dell'utente, uno sforzo minore di pressione dei tasti rispetto a quella a membrana. In questo caso, i dati raccolti potrebbero essere compromessi in termini di velocità di scrittura. [2]

Il secondo problema riguarda il layout della tastiera. Un'utente che utilizza quotidianamente un tipo di layout di tastiera potrebbe non ottenere le stesse performance su un altro.

Il terzo problema è relativo a fattori esterni che possono influenzare l'utente durante la scrittura di frasi: distrazioni, infortuni, stanchezza o condizioni ambientali.

Partendo da questo background, l'obiettivo principale di questo progetto è quello di comprendere se esistono fattori discriminanti biometrici, utilizzando la dinamica dei tasti, in grado di distinguere utenti e rivelare parte dell'identità di una persona sconosciuta. In tal senso, l'approccio proposto si basa sull'idea che i caratteri demografici possono avere un impatto unico sul modo in cui gli utenti digitano. L'ipotesi è che gli individui tendano a trasferire caratteristiche personali specifiche al loro comportamento di battitura, come la velocità di battitura e i ritardi nei diagrammi. Analizzando queste caratteristiche è possibile sviluppare un metodo per discriminare genere, età, lingua parlata, continente di provenienza e corso di scrittura.

L'analisi su dati di keystrokes per scopi di classificazione demografica apre numerose possibilità applicative. Ad esempio, può migliorare l'accuratezza dei sistemi di raccomandazione personalizzati, ottimizzare l'interfaccia utente in base alle preferenze e alle capacità dell'utente o persino contribuire a studi comportamentali su larga scala.

Le sezioni successive di questo documento offrono una panoramica esaustiva sull'argomento. Nella sezione successiva (2), verrà esposto un breve excursus sul problema applicativo proseguendo attraverso una sintesi delle metodologie implementate. A seguire, nella sezione 3, viene presentato il dataset e le operazioni di data wrangling con le strategie adottate. La sezione 4 , poi, si concentrerà sull'analisi esplorativa del dataset necessaria per una comprensione delle features. La sezione 5 spiega il protocollo sperimentale. La sezione 6 gli esperimenti eseguiti, con conseguente presentazione e analisi dei risultati raggiunti . La sezione 7 trarrà le conclusioni dello studio.

2. Metodologia

Per keystrokes dynamics, ci si riferisce al processo di misurazione e valutazione del ritmo di battitura umano sui dispositivi digitali. Una forma di impronta digitale viene creata dall'interazione umana con il dispositivo di acquisizione (solitamente un telefono cellulare, computer o un touch screen). Si ritiene che queste firme siano ricche di qualità cognitive, che sono abbastanza uniche per ogni individuo e hanno un enorme potenziale come identificatore personale. [3] Le caratteristiche più importanti e ampiamente utilizzate nell'analisi keystroke della pressione dei tasti sono la durata della pressione dei tasti e le latenze dei diagrammi. La durata della pressione dei tasti si riferisce alla quantità di tempo in cui un tasto viene tenuto premuto prima di essere rilasciato, mentre le latenze dei diagrammi si riferiscono alla quantità di tempo che intercorre tra l'uso di due tasti consecutivi. L'analisi implementata si basa sull'idea che ogni individuo digita in un modo unico e che questo stile di digitazione può essere utilizzato come forma di identificazione biometrica, come un'impronta digitale o la scansione del riconoscimento facciale o come mezzo per classificare gli utenti. Nel caso specifico ci si concentrerà sulla distinzione di aspetti determinanti e significativi per la classificazione. La metodologia utilizzata comprende tre fasi distinte. La *prima fase* prevede la raccolta dei dati, in cui i dati keystrokes vengono registrati dai partecipanti. Nella *seconda fase*, ai dati raccolti vengono applicate tecniche di estrazione e selezione delle caratteristiche. Infine, nella *terza fase,* i modelli di machine learning vengono addestrati e valutati utilizzando le funzionalità selezionate.

Nella prima fase, il dataset utilizzato e la procedura di data collection, richiede l'installazione di keylogger sui computer dei partecipanti, i quali eseguendo dei task producono dei file utili all'analisi. In ciascun file sono registrate le azioni svolte sulla tastiera.

Nella seconda fase di la features extraction, partendo dai dati di ciascun partecipante, è possibile estrarre delle caratteristiche utili all'implementazione dei modelli di machine learning. Le informazioni di ciascun partecipante vengono aggregate rispetto alle informazioni demografiche.

Infine, nella terza fase, viene presentata l'estrazione dei migliori risultati dopo l'utilizzo di modelli di machine learning.

3. Dataset

Il dataset utilizzato, "*136 M keystrokes*", disponibile pubblicamente, si compone di circa 169 mila osservazioni (utenti) su 16 variabili. Il dataset è stato raccolto nel contesto di uno studio innovativo sulla biometria comportamentale, in particolare mirato a esplorare se il modo unico di digitare di un individuo può essere utilizzato per il riconoscimento

Table 1
Descrizione delle variabili estrapolate dal Keylogging per utente

Variabile	Descrizione
PARTICIPANT_ID	Identificativo unico per ciascun partecipante.
TEST_SECTION_ID	Identificativo associato alla frase specifica che il partecipante doveva trascrivere, utilizzato per collegare ciascun tentativo alle 15 frasi presentate.
SENTENCE	La frase assegnata per la trascrizione.
USER_INPUT	La risposta effettiva data dal partecipante, utile per analizzare gli errori di digitazione.
KEYSTROKE_ID	Identificativo unico di ogni pressione di tasto.
PRESS_TIME	Timestamp dell'evento di pressione di ogni tasto espresso in millisecondi.
RELEASE_TIME	Timestamp dell'evento di rilascio di ogni tasto, espresso in millisecondi.
LETTER	Lettera digitata.
KEYCODE	Il codice JavaScript del tasto premuto.

Table 2
Descrizione dettagliata delle variabili del dataset

Variabile	Descrizione
PARTICIPANT_ID	Identificativo unico per ciascun partecipante.
GENDER	Sesso biologico.
AGE	Età dell'utente.
COUNTRY	Paese di provenienza.
NATIVE LANGUAGE	Lingua parlata dall'utente.
LAYOUT	Layout della tastiera usato (QWERTY, AZERTY, QWERTZ).
HAS TAKEN TYPING COURSE	Indica se il partecipante ha mai seguito un corso di dattilografia.
FINGERS	Numero di dita utilizzate per digitare.
TIME SPENT TYPING	Ore trascorse a scrivere al computer ogni giorno.
KEYBOARD TYPE	Tipo di tastiera utilizzata (desktop, laptop, fisica piccola, touch).
ERROR RATE	Tasso di errore di scrittura.
AVG WPM 15	Media di parole al minuto diviso 15 frasi.
AVG IKI	Intervallo medio tra le pressioni dei tasti.
ECPC	Correzioni di errore per carattere.
KSPC	Battiture per carattere.
ROR	Rollover ratio, misura la frequenza con cui i tasti sono premuti in sovrapposizione rispetto a quelli adiacenti.

personale. Questo campo di indagine apre interessanti possibilità di sicurezza e personalizzazione, basandosi sul presupposto che ogni persona ha un "impronta digitale" di digitazione distinta e riconoscibile. Il task viene pubblicizzato su diversi siti web, ogni utente sceglie di sottoscrivere la propria partecipazione. Gli utenti che accedono alla pagina web possono scegliere tra test di battitura standard di un minuto, con una parte di testo fissa, o l'esperimento presentato. La maggior parte dei partecipanti sono giovani provenienti dagli Stati Uniti che sono interessati a testare e migliorare le proprie capacità di battitura.

Ogni partecipante allo studio è stato invitato a trascrivere una serie di frasi mostrate casualmente su uno schermo. Durante questa attività, un sistema di keylogging ha accuratamente registrato diversi parametri relativi alla digitazione di ciascun soggetto. [4] Nella *Table 1* i parametri per ogni soggetto.

Il dataset completo integra le informazioni raccolte con dati demografici e di comportamento di digitazione, che includono le variabili descritte nella *Tabella 2*.

DATA CLEANING

Il dataset , presentato nella 2, presenta valori mancanti oltre che valori settati su None, si è proceduto con l'eliminazione degli stessi.

Dalle analisi descrittiva del dataset, 3 si osservano dati utili per rilevare anomalie o irregolarità che potrebbero influenzare negativamente le analisi. Un aspetto critico emerso riguarda l'età dei partecipanti: il valore massimo registrato è di 120 anni. Considerando che il 75% dei partecipanti ha un'età compresa tra 0 e 29 anni, valori così elevati sono altamente improbabili e potrebbero indicare errori di inserimento o outlier. Per garantire la coerenza e la qualità dei dati, è stato deciso di considerare solo i partecipanti con un'età compresa tra 10 e 90 anni, in linea con le soglie utilizzate nella letteratura di riferimento e nel paper di base.

Inoltre, per quanto riguarda la variabile che misura il tempo quotidiano trascorso a scrivere al computer, è stato notato che alcuni partecipanti hanno riportato valori che superano le 24 ore giornaliere o altri che hanno valore negativo. Questo ovviamente non è possibile, pertanto è stato deciso di considerare valide solo le risposte in cui il numero di ore giornaliere di digitazione è inferiore a 24 e maggiore o uguale a 0. Questo filtro ulteriore assicura che i dati utilizzati nelle analisi siano realistici e veritieri.

Per la variabile *Fingers*, che indica il numero di dita usate per digitare, sono stati eliminati i valori che non rientrano nel range possibile di 0 a 10 dita. Questo passaggio assicura che i dati rimangano realistici e coerenti con la fisiologia umana.

Table 3
Statistiche Descrittive del Dataset

Statistic	AGE	TIME_SPENT_TYPING	ERROR_RATE	AVG_WPM_15	AVG_IKI	ECPC	KSPC	ROR
Count	146576	146576	146576	146576	146576	146576	146576	146576
Mean	24.47	11616	1.164	50.92	237.60	0.06289	1.172	0.2552
Std	10.51	2979363	1.425	20.30	111.77	0.04474	0.09445	0.1724
Min	0.00	-5	0.000	3.91	21.65	0.00000	1.012	0.0000
25%	17.00	1	0.316	35.21	160.23	0.03226	1.107	0.1097
50%	22.00	2	0.722	48.80	207.27	0.05306	1.152	0.2268
75%	29.00	5	1.498	64.48	283.92	0.08219	1.214	0.3720
Max	120.00	1006950000	24.96	152.56	1825.37	1.90909	4.969	0.9170

DATA WRANGLING

A partire dai dati presenti nel csv per ogni utente in 1, si è proceduto con la creazione di nuove variabili, tenendo conto della letteratura (1) :

- Key Latency: rappresenta il tempo tra due pressioni successive di un tasto. Questa viene calcolata come la differenza tra Key Released di un tasto precedente e Key pressed del tasto successivo.
- Average Keypress: è rappresentativo del tempo medio di pressione, ovvero la differenza tra Key Pressed di un tasto e Key Released dello stesso.
- Di-gram UP UP : analizza il tempo tra il rilascio di un tasto e il rilascio del successivo, importante per valutare la coordinazione nella digitazione;
- Di-gram DOWN DOWN: esprime, in media, il tempo che intercorre tra la pressione del tasto precedente e la pressione del tasto successivo.

  Inoltre, per ognuno di queste, si tiene conto delle principali statistiche di sintesi: massimo, minimo, mediana, primo quartile, terzo quartile.

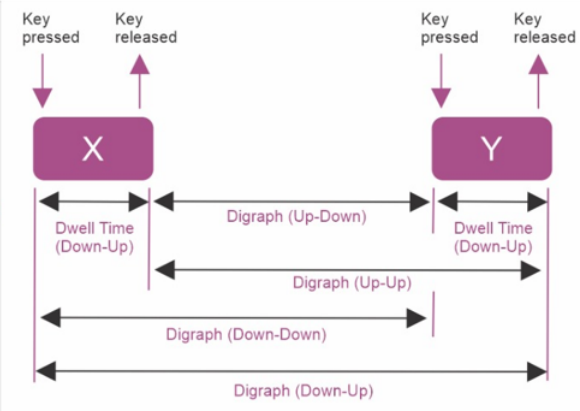

Figure 1: Estrazione delle features

CATEGORIZZAZIONE

Nel trattamento delle variabili categoriali, sono state adottate alcune strategie per semplificare l'analisi, eliminare le anomalie e migliorare la gestione del dataset in vista dell'utilizzo dello stesso per task di classificazione. Nello specifico:

1. *Raggruppamento geografico*: Per la variabile "*COUNTRY*" che indica il paese di provenienza dei partecipanti, si è deciso di aggregare i paesi in macroaree geografiche. Il numero iniziale di paesi di provenienza contava circa 210 paesi. Questo approccio è stato scelto per ridurre il numero eccessivo di categorie uniche, facilitando così le analisi comparative tra diverse regioni. Le macro aree geografiche individuate sono:
   - America;
   - America Latina;
   - Europa;
   - Africa;
   - Asia;
   - Oceania

   Nel dataset, inoltre, sono presenti alcuni paesi che originariamente non sono stati assegnati a nessuna macroarea geografica definita. Per questi casi, l'assegnazione è stata determinata sulla base della posizione geografica del paese di appartenenza. Si è poi deciso di eliminare 4 osservazioni le cui utenze sono provenienti dall'Antartide.
2. *Semplificazione della lingua nativa*: La variabile relativa alla lingua conta circa 150 lingue sparse in tutto il mondo. Considerando che l'83% dei partecipanti ha come lingua madre l'inglese, abbiamo optato per una binarizzazione della variabile "NATIVE_LANGUAGE". Pertanto, sono state create due categorie: "INGLESE" (codificata come 1) e "NON INGLESE" (codificata come 0).
3. Creazione della variabile categorica per l'età: ai fini della classificazione è necessario stabilire un range limite per l'assegnazione delle classi per età. La suddivisione è avvenuta con il seguente criterio logico: ragazzi dai 10 ai 22 anni; ragazzi dai 23 ai 35 anni; adulti con età maggiore di 35 (ma comunque massimo 90 anni).

4. E D A

Nell'ambito dell'analisi è stato condotto uno studio per esaminare le associazioni tra diverse variabili di interesse. Per valutare il legame e il grado di associazione si ricorre ad analisi esplorativa e all'analisi delle corrispondenze multiple.

L'età media dei partecipanti è di 24.5 anni, con uno scarto standard di 10.1 anni. L'età minima registrata è di 10 anni e la massima di 90 anni, concentrando il 75% dei partecipanti sotto i 29 anni. La distribuzione di genere tra i partecipanti è relativamente equilibrata, con una lieve prevalenza femminile (53.08%) rispetto ai maschi (46.92%). La maggioranza dei partecipanti (67.57%) non ha mai frequentato un corso di dattilografia, mentre il 32.43% ha seguito un tale corso.

Circa la metà dei partecipanti (49.14%) usa da 9 a 10 dita per digitare, indicando una prevalenza di tecniche di digitazione standard. Altri gruppi includono coloro che usano 7-8 dita (16.67%), 3-4 dita (13.48%), 5-6 dita (11.18%) e 1-2 dita (9.52%).

L'Area geografica di provenienza presenta una forte concentrazione di partecipanti dagli Stati Uniti (61.72%), seguiti da India (8.24%), Filippine (6.51%), Canada (4.65%) e Regno Unito (3.76%). Altre nazioni sono rappresentate in proporzioni molto minori, con alcuni paesi che hanno una rappresentazione praticamente nulla.

Inoltre, l'inglese è la lingua madre predominante tra i partecipanti, rappresentando l'82.94% del totale. Altre lingue significative includono il Tagalog (2.63%), lo Spagnolo (2.12%), il Cinese (1.56%) e l'Hindi (1.40%). Si possiede un totale di 162 lingue, ma molte di queste sono presenti solo in una frazione minima dei partecipanti, dimostrando una vasta diversità linguistica ma con una forte inclinazione verso l'inglese. La maggior parte dei partecipanti predilige l'utilizzo di laptop (52.26%) e tastiere complete (45.71%) per la digitazione. Le tastiere piccole fisiche e quelle touch sono molto meno comuni, rappresentando solo l'1.10% e lo 0.93% dei partecipanti, rispettivamente.

Per quel che concerne il comportamento di digitazione dei partecipanti: in media, i partecipanti trascorrono circa 3.2 ore al giorno a digitare, con un massimo di 24 ore. La media della velocità di digitazione è di 51.08 parole al minuto, ma con una variazione significativa (std = 20.26 WPM), indicando diversi livelli di abilità di digitazione. L'errore medio è dell'1.16%, con un massimo che raggiunge quasi il 25%. L'intervallo medio tra la pressione dei tasti è di 236.35 millisecondi, evidenziando differenze nei tempi di reazione tra i partecipanti. Da ulteriori analisi, i partecipanti, in media, fanno correzioni in circa il 6.27% dei caratteri digitati e producono 1.17 battiture per carattere. La media del rollover ratio è 0.26, con alcuni partecipanti che mostrano un uso quasi perfetto del rollover (max = 0.917). I tempi medi tra pressioni e rilasci consecutivi dei tasti sono simili, con medie attorno a 380 millisecondi, ma con un'ampia variazione.

ACM : Analisi delle corrispondenze multiple

Table 4
Statistiche di backgound sui partecipanti

Variabile	RisultatiRemark	
Donne	52.7%	
Uomini	48.3	
Età: Media	24.5	75% tra 11-30 anni
Paesi	218	61% ca. dall'America
Tempo medio scrittura	3.2	64% < 2 ore, 14% > 6 ore

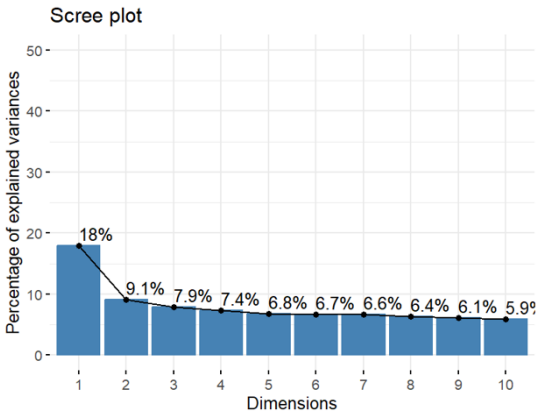

Figure 2: Scree Plot ACM

In seguito, è stato applicato l'analisi delle corrispondenze multiple il cui fine ultimo risiede nell'esplorazione e visualizzazione delle associazioni tra le categorie delle variabili prese in esame. Per il solo scopo esplorativo le variabili numeriche sono state trasformate in categoriali distinguendo utenti i cui valori sono maggiori o minori della media della stessa variabile.

La Fig. 2 mostra lo screeplot delle dimensioni ottenute in seguito all'applicazione della analisi delle corrispondenze multiple. La visualizzazione rende chiaro che la varianza e l'inerzia totale spiegata da ciascun fattore diminuisce a mano a mano che queste diventano maggiori in numero. Il punto in cui la curva inizia a livellarsi è definito "elbow point" indicando il numero di dimensioni significative, questo potrebbe essere individuato tra la componente numero 4 e la componente numero 8.

La visualizzazione bidimensionale degli assi principali è presentata nella Fig. 3 . Il grafico dell'analisi delle corrispondenze multiple viene creato proiettando le variabili categoriche su un piano bidimensionale in modo da visualizzare le loro relazioni. La rappresentazione è composta da un piano cartesiano con due assi principali. L'inerzia totale spiegata dal grafico dell'analisi delle corrispondenze multiple basato sui primi due assi fattoriali, ovvero la misura del grado di dispersione del profilo attorno al profilo medio, è del circa 30%. Questo consente di individuare facilmente le relazioni tra le righe e le colonne della tabella di dati, facilitando l'analisi e la comprensione. Di seguito quelle essenziali e da prendere in considerazione:

1. L'elevato contributo tra Rollover ratio < media, Parole per minuto < media, IKI > media e utilizzo di 1-2 dita per la scrittura. Queste relazioni potrebbero essere strettamente correlate fra di loro: utenti che utilizzano poco frequentemente la tecnica del rollover durante la scrittura sono, in genere, persone non specializzate. Data la loro scarsa velocità di scrittura, la latenza tra un tasto e il successivo è molto alta. Di solito, questo comportamento è frequentemente associato a utenti che utilizzano 1-2 dita durante la scrittura.
2. La relazione tra sesso biologico femminile e tempo medio di pressione dei tasti > media.
3. La vicinanza tra gli utenti la cui lingua nativa è inglese, tasso di errore < media, tempo dedicato alla scrittura < media. Gli utenti che posseggono un'elevata padronanza della lingua inglese hanno meno probabilità di commettere errori anche senza dedicare necessariamente un cospicuo contributo di ore giornaliere alla scrittura.
4. In maniera analoga al punto (2), la vicinanza tra soggetti di sesso maschile e tempo di pressione < media.
5. Gli utenti che non hanno svolto corsi di scrittura, utilizzo di (3-4)(5-6) dita, tasso di errore > media e non parlano nativamente la lingua inglese. Gli utenti la cui lingua nativa non è inglese hanno un tasso di errore di errore superiore alla media in quanto utilizzano dalle 3 alle 6 dita per scrivere. Ciò probabilmente è dovuto al fatto che le persone che non hanno mai seguito un corso, hanno meno familiarità con la coordinazione delle dita nella scrittura di frasi che non appartengono alla loro lingua madre.
6. La prossimità, analogamente al punto (1), tra rollover ratio> media, iki < media, parole per minuto > media.

5. Protocollo Sperimentale

In questa sezione ci si concentrerà sulla descrizione accurata del metodo di valutazione e addestramento proposto per le soft biometrics. Prima di generare i modelli, per le possibili variabili, sono stati sintetizzati features relative a massimo, minimo, mediana, primo quartile e terzo quartile, oltre che deviazione standard. Nella fase sperimentale sono stati implementati 7 modelli di machine learning per problemi applicativi binari e multiclasse: Modello di regressione Logistico, Random Forest, Gradient Boosting, AdaBoost, SVC, MLP e XGBoost. E' stata implementata la ricerca degli iperparametri ottimali attraverso cross-validation. La lista degli iperparametri è presente nella Tabella 5 .

Il problema applicativo non presenta più di una istanza per ciascuno dei soggetti per cui non è stato necessario implementare *subject indipendent*. Tuttavia, ci si è comunque attenuti alle linee guida per lo splitting dei dati in training set e test facendo si che le osservazioni presenti nel training set non siano presenti di nuovo nel test set. La costruzione dei modelli di classificazione ha tenuto conto della sola variabile demografica da classificare come variabile di output e le

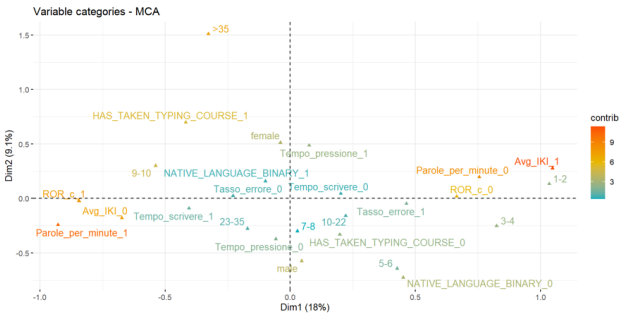

Figure 3: Biplot degli Assi Principali

variabili relative alla keystrokes dynamics come variabili indipendenti. Ci si è concentrati sui seguenti metodi applicativi:

1. Per tutte le variabili demografiche utili alla classificazione, che siano binarie o multi-classe, si è proseguito con l'addestramento dei modelli e l'individuazione del migliore sulla base delle metriche proposte (weighthed accuracy e f1-score).
2. Il nostro dataset presenta inizialmente un marcato sbilanciamento nelle variabili demografiche, il che si traduce in una bassa precisione nella classificazione delle categorie minoritarie come positive e una scarsa recall per le stesse quando considerate negative. Per contrastare questo fenomeno, le categorie con poche osservazioni sono state eliminate o inglobate in una categoria superiore (per ognuna delle implementazioni è presente una descrizione dettagliata). È stato implementato, inoltre, il metodo di oversampling SMOTE, che genera dati sintetici per la classe minoritaria, equilibrando così la distribuzione delle classi.
3. Un'ulteriore strategia testata è stata quella di suddividere il dataset in fold bilanciati, valutando l'efficacia dei modelli tanto sui dati bilanciati quanto su quelli non modificati da dati artificiali. Il risultato presentato è la media delle performance dei singoli modelli sui fold bilanciati.

Per quanto riguarda la divisione del dataset, abbiamo optato per una stratificazione mediante la funzione *train_test_split*, che ci ha permesso di mantenere le proporzioni originali delle categorie nella variabile dipendente. Questo approccio ha assicurato una distribuzione equa delle categorie nel training set e nel test set, rispettivamente costituiti dall'80% e dal 20% delle osservazioni.

L'approccio sistematico e giustificato garantisce l'integrità e l'efficacia del nostro protocollo sperimentale, fornendo un solido fondamento metodologico per le nostre analisi successive.

6. Risultati

Gli esperimenti sono stati condotti su una macchina equipaggiata con un processore Intel Core i9-14900K, che

Table 5
Parametri per i modelli ML

Modello	Parametri
Logistic Regression	solver: [bfgs, sag, saga]
Random Forest	{n_estimators: [100, 300, 500], max_depth: [None, 10, 20, 30], max_features:[sqrt, log2]}
Gradient Boosting	{n_estimators: [100, 200, 300], max_depth: [3,5,7], learning_rate:[0.01,0.05,0.1]}
Ada Boost	{n_estimators: [50,100,150], learning_rate:[0.01,0.1,0.001]}
SVC	None
MLP	{hidden_layer_sizes: [(50,), (100,), (100, 100)], activation.: ['relu', 'tanh'], alpha: [0.0001, 0.001, 0.01]}
XGBoost	{n_estimators: [100, 200, 300], max_depth: [3,6,9], learning_rate:[0.01,0.05,0.1]}

dispone di 24 core suddivisi in 8 Performance-cores (P-core) e 16 Efficiency-cores (E-core), operanti a una frequenza di 6.0 GHz, con supporto per un totale di 32 thread. La macchina è dotata di 32 GB di memoria RAM DDR5 a 7200 MHz e di una scheda grafica NVIDIA GeForce RTX 4070 Ti Super con 16 GB di memoria GDDR6 dedicata. La scheda madre utilizzata è una MSI MAG Z790 TOMAHAWK MAX WIFI, mentre il sistema di raffreddamento è un dissipatore a liquido NZXT Kraken 360. L'alimentazione è garantita da un alimentatore MSI MPG A1000G PCIE5. Il sistema operativo installato è Windows 11 Pro. Per l'esecuzione degli esperimenti è stato utilizzato Python nella versione 3.10.14.

L'addestramento effettivo ha tenuto conto di 5 task di classificazione: genere, età, corso di scrittura, lingua parlata e area geografica di provenienza. Ognuno dei modelli ha preso in input le variabili che esprimono i comportamenti di scrittura, di seguito elencati: 'ERROR_RATE','AVG WPM 15', 'AVG_IKI', 'ECPC', 'KSPC', 'ROR', 'MEDIAN DOWN DOWN', 'MAX DOWN DOWN', 'MIN DOWN DOWN', 'STD DOWN DOWN', 'PRIMO Q DOWN DOWN', 'TERZO Q DOWN DOWN', 'MEDIAN UP UP', 'MAX UP UP', 'MIN UP UP', 'STD UP UP', 'PRIMO Q UP UP', 'TERZO Q UP UP', 'MEDIAN KEYPRESS', 'MAX KEYPRESS', 'MIN KEYPRESS', 'STD KEYPRESS', 'PRIMO Q KEYPRESS', 'TERZO Q KEYPRESS'.

L'individuazione di modelli di machine learning attraverso la keystroke dynamics per il genere e l'età non ha prodotto risultati utili per lo studio. Per il primo, seppur gli oggetti per le classi sono abbastanza bilanciati, i risultati in termini di accuratezza sono molto bassi. Per il secondo, l'addensamento dell'età degli utenti tra 10 e 22 anni, ha reso la variabile fortemente sbilanciata. Per qualunque categorizzazione si optasse, due classi o tre classi, e qualunque metodologia di addestramento venisse implementata i risultati hanno restituito risultati pessimi. Di seguito, i risultati migliori per i restanti task di classificazione.

Per l'area geografica, i migliori risultati si ottengono per la categorizzazione America e Non America. La Fig. mostra l'importanza delle variabili sulla base degli SHAP values per l'area geografica. In particolare, la massima accuratezza viene raggiunta dal Gradient Boosting con tecnica di oversampling SMOTE, visto lo sbilanciamento tra le classi, con accuratezza prossima al 77%. Dal grafico è possibile notare che la caratteristica più influente è KSPC, numero medio di battute necessarie per digitare ogni carattere. Seguono Error Rate, il tasso di errore durante la digitazione, e Min Down Down , che riflette il tempo minimo tra la pressione di due tasti consecutivi. Infine, ECPC e Rollover Ratio.

Per la lingua parlata, categorizzata in Inglese/Non Inglese, il miglior modello è il Gradient Boosting con l'applicazione di oversampling SMOTE. La classe, fortmente sbilanciata, ha mostrato risultati pessimi con il modello addestrato sui folds bilanciati mentre buoni sui dati originali ma con tendenza a preferire la classe maggioritaria (inglese). Il tasso di accuratezza del modello è circa dell' 88%. Dalla Fig. 5, le variabili più importanti sono ERROR RATE e ECPC, con shap value piuttosto elevati.

Per frequenza del corso di scrittura, Fig. 6, anch'esso fortemente sbilanciato, il miglior tasso di accuratezza è stato raggiunto dal Gadient Boosting addestrato sul dataset a cui è stato applicato SMOTE. Il valore di accuratezza è pari circa al 75%. Il contributo delle variabili è decisamente più basso rispetto ai task precedenti ma le variabili discriminanti rimangono le stesse.

7. Conclusioni

L'identificazione di specifiche caratteristiche degli utenti online, quali genere, età, lingua parlata, area geografica e corso di scrittura, può essere fondamentale per fini di

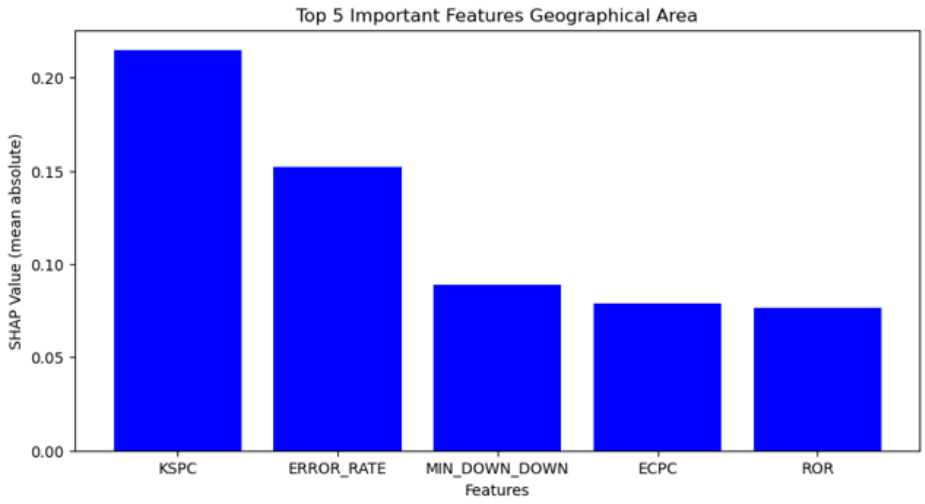

Figure 4: Features Importance Area Geografica

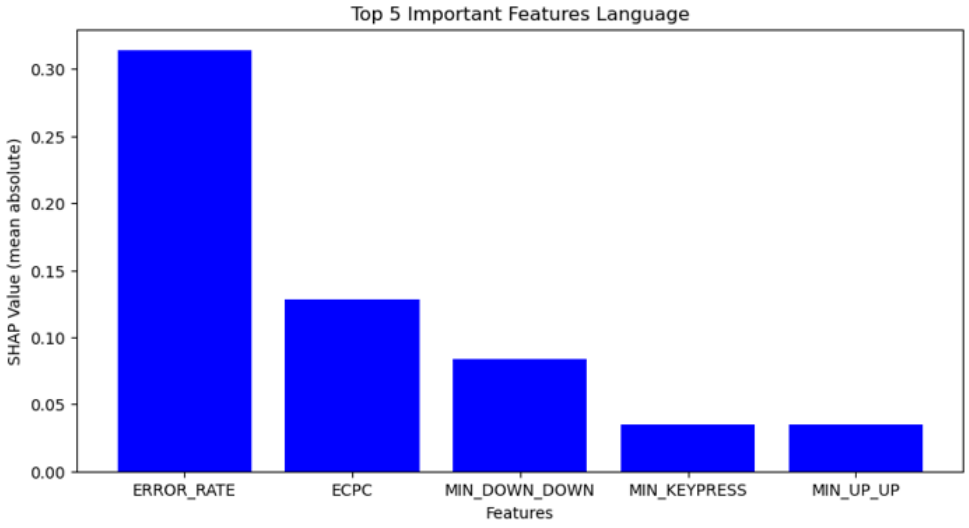

Figure 5: Feature Importance Lingua Parlata

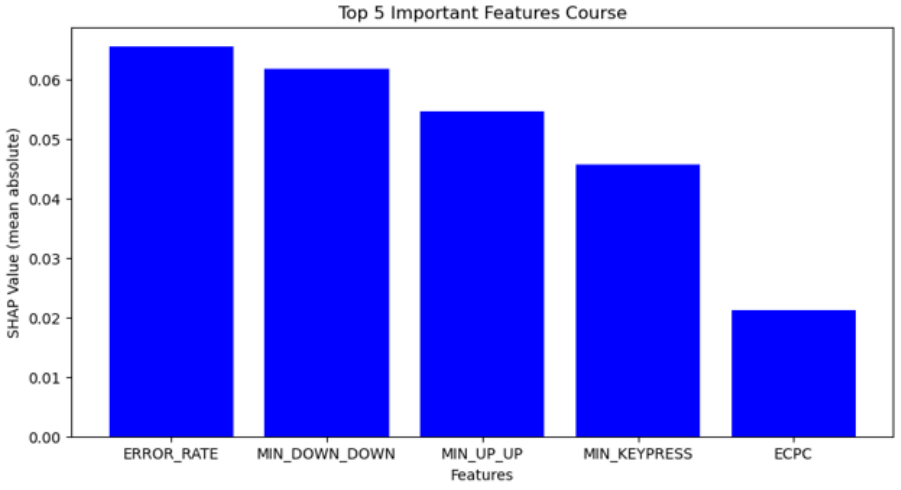

Figure 6: Features Importance Corso di scrittura

sicurezza e per applicazioni commerciali. In questo progetto, vengono proposte delle metodologie per rivelare caratteristiche demografiche dell'utente, attraverso kesytroke dynamics. Per lo scopo del progetto, viene utilizzato il dataset "136 M Keystrokes" disponibile pubblicamente. Dal dataset sono state estratte le feature di tastiera più comuni, tra queste: tempo di latenza, tasso di errore, tempo di pressione, tempo minimo di digitazione e rollover ratio. Nella fase sperimentale sono stati testati diversi algoritmi di machine learning per la classificazione di genere, età, area geografica di appartenenza, lingua parlata e corso di scrittura. Per genere ed età nessuna delle metodologie proposte è stata in grado di ottenere risultati discreti. Al contrario, per lingua parlata, area geografica e corso di scrittura il Gradient Boosting con l'applicazione di SMOTE ha mostrato tasso di accuratezza più alto. L'accuratezza dei modelli è stata ulteriormente rafforzata dall'uso di SHAP per interpretare l'importanza delle variabili. Il tasso di errore e il numero di Tasti per caratteri si sono rivelati come fattori discriminanti importanti in tutti e tre i modelli.

Uno degli aspetti critici emersi dall'analisi è il forte sbilanciamento del dataset, che è composto prevalentemente da utenti americani, giovani e anglofoni. Questo ha introdotto un bias significativo, limitando la generalizzazione dei risultati e compromettendo la capacità di classificare con precisione utenti di diverse lingue e aree geografiche. Di conseguenza, la classificazione è stata possibile solo per le categorie lingua parlata inglese/non inglese e area geografica America/non America, mentre i tentativi di discriminare tra un numero maggiore di lingue o aree geografiche hanno mostrato capacità predittive scarse. I risultati ottenuti sono, però, promettenti. A partire da questo lavoro è possibile definire degli obiettivi futuri quali, per primo, la registrazione di più utenti provenienti da più parti del mondo in quantità, quantomeno, desiderabili e eque, vista la prevalenza di americani in questo dataset; secondo, considerare l'integrazione della keystroke dynamics con altre forme di biometria comportamentale per aumentare l'accuratezza e la robustezza dei sistemi; terzo, sviluppare sistemi che utilizzano keystroke dynamics in tempo reale per la verifica continua dell'identità.

A. Appendice

E' possibile consultare il notebook commentato con le operazioni svolte per la presentazione del progetto e il pdf dettagliato con le analisi sulla selezione dei modelli non riportato in questo paper.

References

[1] Acien, A., Morales, A., Vera-Rodriguez, R., Fierrez, J. (2019). Keystroke Mobile Authentication: Performance of Long-Term Approaches and Fusion with Behavioral Profiling. In: Morales, A., Fierrez, J., Sánchez, J., Ribeiro, B. (eds) Pattern Recognition and Image Analysis. IbPRIA 2019. Lecture Notes in Computer Science(), vol 11868. Springer, Cham. https://doi.org/10.1007/978-3-030-31321-0_2

[2] Pham, Tri & Kelling, Nicholas. (2015). Mechanical and Membrane Keyboard Typing Assessment Using Surface Electromyography (sEMG). Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 59. 912-915. 10.1177/1541931215591268.

[3] Teh, Pin Shen & Teoh, Andrew & Yue, Shigang. (2013). A Survey of Keystroke Dynamics Biometrics. TheScientificWorldJournal. 2013. 408280. 10.1155/2013/408280.

[4] Vivek Dhakal, Anna Maria Feit, Per Ola Kristensson, and Antti Oulasvirta. 2018. Observations on Typing from 136 Million Keystrokes. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). Association for Computing Machinery, New York, NY, USA, Paper 646, 1–12. https://doi.org/10.1145/3173574.3174220