SAPIENZA
UNIVERSITÀ DI ROMA

Sapienza Università degli studi di Roma
Facoltà di Medicina e Psicologia
Dottorato in Psicologia Dinamica e Clinica, XXXII ciclo

Tesi di Dottorato

Investigazione di alcuni fenomeni di response bias nei questionari di personalità self-report mediante algoritmi di apprendimento automatico

Dottorando
Dott. Pierpaolo Calanna

Tutor
Prof. Marco Lauriola

Co-Tutor
Prof. Aristide Saggino

Ringraziamenti

Ringrazio la Dott.ssa Paolotti, il Gen. Abbenante e il Gen. Arduino per aver reso possibile l'esperienza del dottorato; il prof. Lauriola — mio tutor — per avermi fatto crescere come professionista; la società *Giunti Psychometrics* (in particolare la Dott.ssa Furlan), il prof. Saggino, il prof. Tommasi, il prof. Sellbom e la prof.ssa Van Dongen per aver creduto nel progetto e contribuito alla raccolta dei dati; infine, ringrazio Claudia, Salvatore, Rina, Antonio, Emilia, Emma, Francesca e Milo per avermi sopportato nell'ultimo periodo.

Indice

Elenco delle tabelle

Elenco delle figure

Acronimi

16PF	16 Persoality Factors.
ABLE	Assessment of Background and Life experiences.
ASPD	Disturbo Antisociale di Personalità.
AUC	Area under the ROC curve.
BFQ2	Big Five Questionnaire 2.
BPI	Basic Personality Inventory.
CAPP	Comprehensive Assessment of Psychopathic Personality.
CAPP-SR	Comprehensive Assessment of Psychopathic Personality - Self Report.
CPI	California Personality Inventory.
DIF	Differential Item Functioning.
DTF	Differential Test Functioning.
EPA	Elemental Psichopathy Assessment.
ERM	Minimizzazione del rischio empirico.
FN	False Negative.
FP	False Positive.
FPR	False Positive Rate.
HPI-R	Hogan Personality Inventory - Revised.
HPSI	Holden Psychological Screening Inventory.
IPIP	International Personality ItemPool.
IRT	Item Response Theory.
LCA	Latent Class Analysis.
LRSP	Levenson Self-Report Psychopathy Scale.

ML	Machine Learning.
MM-IRT	Mixed Model - Item Response Theory.
MMPI	Minnesota Multiphasic Personality Inventory.
NEO-FFI	NEO Five Factors Inventory.
NEO-PI-R	NEO Personality Inventory - Revised.
PCL	Psychopathic Checklist.
PCL-R	Psychopathic Checklist - Revised.
PPI	Psychopathic Personality Inventory.
PPI-R	Psychopathic Personality Inventory - Revised.
SRP-III	Self-Report Psychopathy Scale.
TN	True Negative.
TP	True Positive.
TRIPM	Triarchic Psychopathy Measure.

Riassunto

Nel contesto della valutazione della personalità con scale *self-report*, il termine *faking* si riferisce ai tentativi messi in atto dagli individui di alterare le risposte agli item di un test al fine di costruire e comunicare un'immagine di sé non veritiera, funzionale al raggiungimento di scopi personali (che spesso confliggono con il processo misurativo). Le scale *Lie* sono una delle strategie di rilevazione del *faking* più diffuse e si basano sull'ipotesi che quest'ultimo possa essere equiparato a un fenomeno lineare/quasi-lineare quantificabile attraverso l'accumulazione "indiziaria" dei suoi effetti sulle risposte specifiche agli item *Lie*. Il superamento di una soglia di attenzione di natura normativa determina la presa di decisione in merito all'attendibilità delle risposte in generale. Due recenti studi hanno affrontato il problema da una prospettiva diversa. Kuncel e Borneman (2007) e Kuncel e Tellegen (2009) hanno mostrato: (a) che il *faking* può verificarsi a livello delle singole opzioni *likert* degli item; (b) e che la sua natura può essere intrinsecamente non lineare. Queste due proprietà si traducono in matrici di risposte caratterizzate da "idiosincrasie" numeriche che possono fungere da marcatori dei comportamenti distorsivi se studiate nella loro individualità, ma che vengono "oscurate" dal meccanismo aggregativo dei punteggi di scala[1]. Con il presente lavoro abbiamo tentato di sviluppare una tecnica innovativa per la rilevazione del *faking* attraverso l'analisi dei *pattern* di risposta agli item. Al fine di raggiungere tale scopo, ci siamo posti due domande esplorative: (1) è possibile impiegare gli algoritmi di *machine learning* (ML) per rilevare la presenza di *faking*? (2) I classificatori ML possono sostituire efficacemente le scale *Lie*? Per rispondere alle precedenti domande, abbiamo realizzato due studi empirici; nel primo, è stato impiegato il questionario di personalità BFQ2 (Caprara, Barbaranelli, Borgogni & Vecchione, 2007) con

[1] Per illustrare il concetto con un esempio banale, è facile osservare come $1+9$ e $5+5$ assommino ugualmente a 10, ma è altrettanto facile constatare come tali addizioni veicolino significati profondamente diversi, la prima di un forte contrasto, la seconda di perfetta "medietà".

un campione di studenti universitari, nel secondo il *Psychopathic Personality Inventory - Revised* (PPIR; Lilienfeld e Widows, 2005) con un campione di studenti universitari e uno di pazienti psichiatrici. Relativamente al secondo lavoro, la decisione di adottare il PPIR è scaturita dalla constatazione — suffragata dalla letteratura — che gli individui con personalità psicopatica possono esibire condotte distorsive e manipolatorie e dunque gli strumenti *self-report* atti a misurarne l'organizzazione caratterologica rappresentano un buon "banco di prova" per qualunque tecnica di rilevazione del *faking*. L'impostazione generale di entrambi i lavori ha previsto le seguenti fasi: (1) manipolazione diretta del *faking* al fine di ottenere un *dataset* di profili di personalità sia attendibili che distorti (*honest* vs *fake*); (2) implementazione di due o più algoritmi ML in grado di rilevare la presenza di *faking* o nei punteggi di scala o nei *pattern* di risposta; (3) comparazione del miglior algoritmo ML (scelto tra quelli implementati al punto precedente) con il classificatore di riferimento (CBC) basato sui punteggi delle scale *Lie* e relativi *cutoff* normativi. I risultati delle due indagini empiriche hanno confermato l'efficacia dei classificatori ML: la loro performance nel rilevare i cosiddetti *faker* si è rivelata superiore a quella conseguibile con le sole scale *Lie*. Quando le prestazioni sono state valutate in termini di previsioni errate, i classificatori ML si sono rivelati, ancora un volta, migliori delle tecniche basate sulle scale di controllo. Nonostante alcune limitazioni dovute alla manipolazione diretta dei comportamenti distorsivi e ai campioni di partecipanti sbilanciati in termini di età e genere, l'approccio qui proposto consentirebbe di ridurre la lunghezza dei questionari *self-report* di personalità eliminando gli item delle scale di controllo. Forse anche in modo più interessante, tale approccio potrebbe essere usato per "aggiungere" un meccanismo di rilevazione del *faking* ai *self-report* che sono sprovvisti di strategie per la detezione degli stili di risposta distorsivi.

Il presente documento è strutturato in tre capitoli. Nel primo (INQUADRAMENTO TEORICO) vengono presentati — entro i rispettivi ambiti teorici — il *faking*, il disturbo di personalità psicopatica e il *machine learning* (con particolare riferimento agli algoritmi effettivamente impiegati nella ricerca, ovvero alle tecniche di insieme basate sugli alberi decisionali). Nel capitolo di mezzo (STUDI) sono illustrati gli obiettivi, i metodi, le procedure e i risultati dei due studi realizzati. Infine, nell'ultimo capitolo (CONCLUSIONI), sono riportare le riflessioni di carattere conclusivo in merito ai risultati ottenuti.

Capitolo 1

Inquadramento teorico

1.1 *Faking*

Pagato il doveroso tributo di riconoscenza ai primi scienziati che sostennero la necessità di misurare le differenze individuali con tecniche standardizzate, tra cui Wundt, Galton e Cattel, il *testing* psicologico moderno ha raccolto il suo maggiore impeto dallo sviluppo della scala d'intelligenza di Binet-Simon, pubblicata per la prima volta in Francia nel 1905 (Binet, 1905). Qualche anno più tardi, un approccio simile fu adottato dal governo degli Stati Uniti in campo militare (e.g., test di intelligenza *Army Alpha* e *Army Beta*, 1917) e nella gestione degli imponenti flussi migratori, in linea con lo spirito pragmatico americano. Uguale interesse suscitò la misurazione della personalità: è del 1919 il primo questionario di valutazione del carattere, il *Personal Data Sheet* di Woodworth (1920), concepito come strumento di *screening* per indagare la stabilità emotiva dei coscritti alla leva, mentre è del 1943 il primo inventario per l'*assessment* dei disturbi psichici, il *Minnesota Multiphasic Personality Inventory* (MMPI; Hathaway e McKinley, 1943), strumento che prevedeva già nella sua prima edizione un protocollo atto all'individuazione dei comportamenti di risposta distorsivi. Si può insomma affermare che la preoccupazione intorno alla veridicità delle risposte ai test è sorta immediatamente e altrettanto immediatamente si è imposta all'attenzione degli studiosi (Baer, Wetter, Nichols, Greene & Berry, 1995). Alla base di questa preoccupazione vi sono due assunti teorici da salvaguardare: il primo è che i punteggi di un test devono rappresentare una stima *unbiased* dei costrutti analizzati (i.e., bisogna tenere sotto controllo tutte le fonti di variazione sistematica e cioè non

imputabili all'errore casuale), il secondo è che tale stima — per essere davvero utile — deve consentire la previsione di un comportamento futuro (Ziegler, MacCann & Roberts, 2012). L'eventualità che un individuo decida di mentire a un questionario, attribuendosi qualità non realmente possedute o sottacendo caratteristiche di sé ritenute scomode, compromette i due assunti sopra citati e mina le fondamenta del processo misurativo nelle sue componenti diagnostiche e prognostiche. Una simile eventualità non è poi così remota quando si considerino le statistiche sulla prevalenza del fenomeno. Negli scenari cosiddetti *high stakes*, in cui la posta in gioco è alta, la contraffazione delle risposte può riguardare fino al 50% degli individui in esame (per una rassegna vedi Hall & Hall, 2012; Ones, Viswesvaran & Reiss, 1996).

Nello sforzo di circoscrivere i confini del problema, gli esperti anglosassoni hanno introdotto il termine *faking* per indicare l'atto cosciente di falsificare le risposte ai test. Sulla scorta delle riflessioni di Ziegler et al (2012) è possibile fornire la seguente definizione formale: *il faking è una strategia adottata dall'individuo per fornire descrizioni di sé inaccurate al fine di conseguire un vantaggio personale. Il faking genera differenze sistematiche nei punteggi che non sono dovute al costrutto in esame* (p. 8). Esplicitando meglio la precedente definizione, è possibile evidenziare i seguenti punti. Il *faking*: (a) è un comportamento intenzionale; (b) può avere una natura omissiva (tacere le informazioni) e/o commissiva (alterare le informazioni); (c) può portare a una descrizione del sé positiva (*faking good*) o negativa (*faking bad*); (d) è volto al raggiungimento di scopi personali di natura estrinseca[2]; (e) possiede una qualità eminentemente interpersonale in quanto forma di comunicazione (anche se parzialmente compromessa); (f) scaturisce dall'interazione dell'individuo con l'ambiente. È necessario ribadire ancora una volta che si tratta di una strategia consapevole: non è possibile infatti attribuire l'etichetta di *faking* a quelle auto-descrizioni inaccurate che le persone forniscono in perfetta buona fede o per ignoranza di sé. Questa importante distinzione è alla base, ad esempio, del modello bipartito di Paulhus sulla desiderabilità sociale (1988) che prevede una componente definita *Impression Management* e un'altra denominata *Self Deception*. La prima si avvicina al concetto di *faking* per come è stato precedentemente definito, mentre la seconda è una forma di auto-percezione distorta, dovuta o a una carenza di *insight* o a un meccanismo

[2] Questa precisazione si rende necessaria per distinguere il *faking* dai disturbi fittizi, dove la falsificazione di sintomi fisici o psicologici avviene senza un chiaro incentivo esterno; la motivazione di questo comportamento è quella di assumere il ruolo di malato (Krahn, Bostwick & Stonnington, 2008).

di promozione del sé riscontrabile negli individui ben adattati (Lazarus, 1998).

Recentemente, Mueller-Hanson, Heggestad, Thornton et al. (2006) hanno proposto un modello per spiegare i comportamenti di *faking*, integrando precedenti ricerche sull'argomento (McFarland & Ryan, 2006; Snell, Sydell & Lueke, 1999). Nella loro indagine, gli autori hanno inquadrato il fenomeno nell'ottica della teoria motivazionale di Vroom (1964). Quest'ultimo ha postulato l'esistenza di tre fattori che spingono gli esseri umani a porre in essere determinati corsi d'azione. In primo luogo, l'aspettativa che essi hanno di poter produrre un certo comportamento (*Expectancy*), in secondo luogo, la convinzione che tale comportamento sia collegato — e dunque conduca — al raggiungimento di un certo risultato (*Instrumentality*), infine il valore che essi attribuiscono al risultato medesimo (*Valence*). Per quanto riguarda il fattore *Expectancy*, il cosiddetto *faker* sarà tanto più spinto ad alterare le proprie risposte quanto più si sentirà in grado di farlo. Alcune risultanze empiriche hanno confermato questa ipotesi (e.g., Snell et al., 1999) anche se non mancano le ricusazioni (e.g., Weiner & Gibson, 2000). Relativamente al fattore *Instrumentality*, Mueller-Hanson et al. (2006) hanno elencato una serie di credenze soggettive e tratti di personalità (positivamente o negativamente relati) che inducono gli individui a giudicare i comportamenti di *faking* come strategicamente indispensabili: ad esempio, l'integrità morale, il machiavellismo, la manipolatorietà, la coscienziosità, lo stadio morale e la stabilità emotiva. Sono stati pubblicati diversi lavori che hanno confermato — pur con dimensioni dell'effetto moderate — il legame tra *faking* e i precedenti costrutti (e.g., McFarland & Ryan, 2006; Mueller-Hanson et al., 2006). In buona sostanza, le norme soggettive e l'organizzazione di personalità dell'individuo modellano la sua percezione del *faking*, eventualmente portandolo a ritenere che esso sia un comportamento necessario, o meglio necessariamente collegato al raggiungimento degli scopi che si pone. Infine, per quanto riguarda il fattore *Valence*, Leary e Kowalski (1990) hanno enfatizzato come gli individui siano più motivati ad alterare, anche disonestamente, le impressioni che suscitano negli altri quando tale condotta garantisca loro il raggiungimento di obiettivi personali giudicati importanti. Ellingson, Smith e Sackett (2001) hanno invece notato una relazione inversa tra la disponibilità di un obiettivo — inteso come risorsa — e il suo valore percepito, quasi in ossequio alla legge economica della domanda e dell'offerta. Date queste premesse, e considerato che spesso l'*assessment* psicologico tramite test avviene in contesti con forti incentivi esterni concessi a pochi beneficiari (e.g., reclutamento del personale per l'assegnazione di un posto di lavoro, corresponsione di indennizzi economici

per malattia), è legittimo ipotizzare la presenza di comportamenti di *faking* negli individui che si sentono capaci di mentire e che sono caratterialmente propensi a farlo. Heggestad (2012) ha ripreso il modello di Mueller-Hanson et al. (2006) individuando fattori disposizionali, situazionali e fattori legati agli atteggiamenti/opinioni che — in un dinamismo di influenze reciproche — contribuiscono a orientare i comportamenti di *faking* (tabella 1.1). In altri termini, gli autori hanno sottolineato la natura processuale del fenomeno, frutto di inclinazioni caratteriali ma anche di pressioni ambientali, suggerendo di considerare il *faking* come un costrutto avente una componente stabile di tratto e una componente di stato più variabile e situazione-specifica (Heggestad, 2012, p. 91).

Tabella 1.1: Elenco dei costrutti legati al *faking* (Heggestad, 2012).

Construct	Relationship	Empirical support
Dispositional factors		
Cognitive ability	+	Mixed
Conscientiousness	-	Yes
Emotional Stability	-	Yes
Integrity	-	Yes
Locus of control (internal)	-	Yes
Machiavellanism	+	Yes
Manipulativeness	+	Yes
Need for approval	+	No
Stage of moral development	-	No
Self-monitoring	+	Mixed
Attitudinal factors		
Others' frequency	+	No
Perceived fairness of test	-	No
Perceived behavioural control	+	Yes
Personal attitude towards faking	+	No
Subjective norms	+	Yes
Situational factors		
Importance of the outcome	+	No
Knowledge on what test measures	+	No
Konwledge on how test will be used	+	No

1.1.1 Effetti sui punteggi di scala/strutture fattoriali/validità

***Nota**: nelle sezioni che seguono, sono stati impiegati diversi acronimi. Consulta la lista delle abbreviazioni che si trova all'inizio della tesi.*

Effetti del *faking* sui punteggi di scala.

Viswesvaran e Ones (1999) hanno compiuto una meta-analisi in merito ai comportamenti di *faking good* nei questionari *Big Five* (i.e., Estroversione, Amicalità, Coscienziosità, Stabilita Emotiva, Apertura); gli autori hanno riportato incrementi dei punteggi pari a circa 0.5 deviazioni standard per le scale di personalità[3]. Analoghe variazioni (in direzione contraria) sono state osservate negli scenari *faking bad*. Dunnette, McCartney, Carlson e Kirchner (1962), Rosse, Stecher, Miller e Levin (1998) e Stark, Chernyshenko, Chan, Lee e Drasgow (2001) hanno documentato scostamenti fino a 0.6 deviazioni standard dei punteggi di personalità con campioni di candidati nelle fasi di reclutamento (rispetto a lavoratori con medesimo profilo professionale), mentre Butcher, Morfitt, Rouse e Holden (1997) hanno stimato valori leggermente inferiori nell'ambito delle selezioni per piloti di aviazione civile. Hough, Eaton, Dunnette, Kamp e McCloy (1990) hanno riportato aumenti fino a 0.5 deviazioni standard in un campione di aspiranti militari rispetto al personale di carriera. In generale dunque i dati sembrano indicare elevazioni dei punteggi di scala con una dimensione dell'effetto media. Occorre notare, che diversi studi hanno manipolato direttamente le condotte di *faking* e non è chiaro se e come le variazioni occorse nei punteggi siano generalizzabili a contesti naturalistici. Riprendendo le parole di Smith e McDaniel: *"La manipolazione diretta del faking negli studi sperimentali [...] non può dare un'indicazione precisa di come le persone si comporteranno nelle situazioni reali. Al limite, fornisce informazioni sull'estensione massima delle condotte distorsive quando gli individui decidono di mentire. [...] È proprio il problema della motivazione che viene rimosso in questo genere di studi e ciò limita la nostra comprensione del problema."* (Smith & McDaniel, 2012, p. 55); in altri termini, la manipolazione diretta del *faking* chiarisce quella che potrebbe essere la performance di picco (capacità massima) e non la performance effettivamente riscontrata nei *setting* naturali (capacità tipica).

[3] Topping e O'Gorman (1997) e Holden, Wood e Tomashewski (2001) hanno riscontrato che la scala dell'Apertura è più resistente ai fenomeni distorsivi, anche se non sono state chiarite le motivazioni di questo comportamento differenziale.

Effetti del *faking* sulle strutture fattoriali.

In relazione al *faking* direttamente manipolato, diverse ricerche sull'MMPI, BPI, NEO-FFI, HPSI hanno confermato un aumento significativo delle correlazioni tra le scale a prescindere dal fatto che il disegno sperimentale prevedesse la simulazione di un buon adattamento o di disagio psichico. Negli scenari di *faking* naturale, Sackett, Schmitt, Ellingson e Kabin (2001) hanno raccolto e analizzato i dati multi-studio di alcuni strumenti di personalità, tra cui il CPI, il 16PF e l'HPI-R. Gli autori non hanno trovato variazioni significative delle strutture fattoriali. In linea con questi risultati, Smith e Ellingson (2002) non hanno rilevato modificazioni dell'HPI-R tra studenti e aspiranti lavoratori. Nel loro studio sul NEO-PI-R, Marshall, De Fruyt, Rolland e Bagby (2005) sono giunti alle medesime conclusioni confrontando gruppi di individui ad alta desiderabilità sociale con i campioni normativi e di *counseling*. In contrasto con queste evidenze, Schmit e Ryan (1993) e Brown e Barrett (1999) hanno evidenziato alterazioni significative, benché modeste, del NEO-FFI e del 16PF in un campione misto di aspiranti lavoratori e studenti.

In sintesi, gli effetti del *faking* sulle strutture fattoriali dei questionari di personalità sembrano essere nulli o modesti. Ciononostante, in due recenti lavori su una misura dei *Big Five* tratta dalla banca dati *Open Source* IPIP[4]. Biderman (2014; 2004) ha implementato un modello bi-fattoriale composto dalle variabili latenti dei costrutti di personalità e da una variabile di metodo (ortogonale alle precedenti) che catturava la varianza comune a tutti gli item imputabile, almeno in parte, ai comportamenti di *faking*. Il modello proposto ha dimostrato *fit* superiori rispetto a una soluzione comprensiva delle sole dimensioni di personalità.

Effetti del *faking* sulla validità.

In riferimento alle manipolazioni dirette del *faking*, Douglas, McDaniel e Snell (1996), Holden e Jackson (1985), Holden et al. (2001), Topping e O'Gorman (1997) hanno riportato un decremento della validità (di criterio) delle misure di personalità impiegate, indipendentemente dallo stile distorsivo adottato dai partecipanti, positivo o negativo che fosse. I risultati delle indagini sul *faking* in setting naturali sono invece contrastanti. Hough et al. (1990), in uno studio su più di 8.000 militari, hanno documentato che la

[4] Disponibile all'indirizzo https://ipip.ori.org.

validità delle scale di personalità non risentiva dei diversi livelli di *faking*. Similmente, Barrick e Mount (1996) hanno evidenziato che la validità di una misura dei *Big Five* applicata a un campione di camionisti non migliorava se si usavano due indici di *faking* come covariate. Coerentemente con i precedenti studi, Piedmont, McCrae, Riemann e Angleitner (2000) non hanno rilevato alcuna compromissione della validità di criterio del NEO-PI-R in un campione di studenti universitari. Procedendo in direzione contraria, alcune indagini empiriche hanno osservato significative riduzioni della validità legate ai comportamenti di risposta dei partecipanti. Ad esempio, analizzando i dati del questionario ABLE delle forze armate statunitensi su un campione di migliaia di militari, White, Young e Rumsey hanno osservato che i livelli di desiderabilità sociale *"limitavano severamente la validità di criterio dello strumento rispetto ad altre misure del temperamento"* (White et al., 2001, p. 550). Anche Holden (2007) ha constatato alterazioni della validità di una misura dei *Big Five* somministrata a un gruppo di studenti universitari in un disegno di tipo *self/peer-rating*. Le analisi dei dati dell'autore hanno chiarito che che circa il 10%∼15% della varianza delle valutazioni dei pari veniva spiegata dai punteggi della scala di controllo del questionario *self-report*.

In un'altra serie di indagini, l'impatto del *faking* sulla validità (di costrutto) dei questionari è stata valutata nell'ottica dell'Item Response Theory (IRT), analizzando i fenomeni di DIF (*Differential Item Functioning*) e DTF (*Differential Test Functioning*): il primo si verifica quando un item ha una relazione con la variabile latente che varia tra gruppi differenti (e.g., maschi vs femmine, *faker* vs *non faker*); il secondo è l'estensione del primo fenomeno a livello di scala (Natali, 2008). In altre parole, la presenza di DIF/DTF indica che gli item di una scala o la scala nella sua interezza funzionano in modo diverso per alcuni rispondenti rispetto ad altri e ciò si riverbera negativamente sulla validità di costrutto. Stark et al. (2001) hanno documentato la presenza di DIF/DTF analizzando un campione di studenti che avevano risposto al 16PF in due diverse condizioni di somministrazione (*fake* vs *no fake*). Il funzionamento differenziale ha riguardato 13 delle 16 scale del questionario. Tuttavia, in un altro studio condotto da Henry e Raju (2006) con uguale disegno sperimentale non sono stati rilevati importanti fenomeni DIF/DTF.

In conclusione, la letteratura presentata in questa sezione è frastagliata e poco coerente. Non risulta possibile chiarire in via definitiva se e in che misura il *faking* rappresenti una minaccia alla validità dei questionari *self-report*. Ragionando in via del tutto ipote-

tica e assumendo per vera la non relazione, rimarrebbe comunque irrisolto il problema — altrettanto grave — della validità dei profili individuali di personalità. Christiansen, Goffin, Johnston e Rothstein (1994), in uno studio sul 16PF con candidati in *assessment center*, hanno dimostrato che le condotte di *faking* producevano una modificazione dell'ordine di rango nella graduatoria finale (a vantaggio degli *high faker*) per circa l'85% dei candidati scrutinati, benché la presenza di una componente distorsiva nei punteggi di scala non pregiudicasse la relazione del 16PF con i criteri esterni.

1.1.2 Misurazione del *faking*

Scale di controllo

La strategia più diffusa per la misurazione del *faking*, e anche quella più longeva, è rappresentata dalle scale di controllo o scale *Lie* (Paulhus, 1991). Come è stato già detto nella sezione 1.1 (p. 1), la prima edizione del MMPI (Hathaway & McKinley, 1943) ne includeva tre, a conferma di un'attenzione verso gli stili di risposta consustanziale alla valutazione dei costrutti psicologici.

Generalmente, le scale di controllo del *faking good* sono costituite da item indicanti virtù (e.g., *non dico mai bugie*) o difetti (e.g., *a volte sono stato invidioso della fortuna degli altri.*) che gli individui si attribuiscono o disconoscono per dare un'impressione di sé ingannevolmente positiva. L'intento conoscitivo di questi item (al di là del loro contenuto manifesto) può essere tanto palese quanto mascherato a seconda che si vogliano catturare distorsioni grossolane e naïf o strategie più sofisticate (Huber, 2017, pp. 9–10)[5]. Il principio costruttivo delle scale di controllo del *faking bad* è simile: tali scale comprendono sintomi plausibili ma difficilmente realizzabili nel loro insieme a meno di non voler ipotizzare uno stato di disagio fittizio, comunque non corrispondente ai costrutti misurati dal questionario (Huber, 2017).

La premessa concettuale alla base di questa strategia di detezione è che il *faking* sia un processo di natura lineare o quasi-lineare (i.e., più alto è il punteggio delle scale *Lie* più elevato sarà il livello di distorsione). Nel rispondere positivamente agli item di controllo,

[5] Un altro metodo denominato *over-claiming technique* (Paulhus, Harms, Bruce & Lysy, 2003) è basato su item che si riferiscono a realtà inesistenti, anche se non del tutto campate in aria: chi cade nella trappola di riconoscerle come vere dimostra uno stile di risposta non completamente onesto.

gli individui forniscono ripetute evidenze delle loro condotte distorsive. Quando la somma di queste evidenze (i.e., degli item) supera la soglia di attenzione prestabilita (*cutoff score* normativo), gli individui sono classificati come *faker*. In un certo senso, ciascun item si configura come prova indiziaria di *faking* e l'insieme di tali prove formano il castello probatorio.

La letteratura che ha indagato il funzionamento delle scale di controllo (con particolare riferimento al *faking good*) ha fornito un parziale supporto in merito alla loro efficacia. Diverse meta-analisi hanno riscontrato un'effettiva capacità delle scale *Lie* di separare gli *high faker* dai *low faker* con dimensioni dell'effetto medie (Baer & Miller, 2002; Nelson, Hoelzle, Sweet, Arbisi & Demakis, 2010; Rogers, Sewell, Martin & Vitacco, 2003). Nonostante questi risultati, vi sono motivi per guardare con sospetto a questo genere di scale (Huber, 2017). Il problema centrale è che i loro punteggi non sembrano essere idonei in senso predittivo. Ones e Viswesvaran (1998) hanno sostenuto che le scale *Lie* — per essere davvero utili — dovrebbero agire come predittori del criterio esterno (in aggiunta ai costrutti specifici del questionario) o eventualmente funzionare da variabili mediatrici, moderatrici, di soppressione. In altri termini, l'utilità delle scale di controllo dovrebbe scaturire dalla possibilità di impiegarle in almeno uno dei seguenti casi: (a) per prevedere direttamente un determinato criterio esterno; (b) per chiarire le relazioni tra gli altri predittori e il criterio esterno; (c) per spiegare perché la validità di criterio degli altri predittori risulta alterata in determinati gruppi di individui; (d) per rimuovere la quota di varianza irrilevante dai punteggi degli altri predittori. Ones e Viswesvaran (1998) hanno sottoposto a verifica le prime tre ipotesi trovando scarso supporto per ciascuna di esse. L'ultima possibilità è che le scale di controllo agiscano come variabili moderatrici. Ciò si verificherebbe se i coefficienti di validità fossero diversi per gli individui che presentano livelli elevati di *faking* rispetto a quanti esibiscono livelli meno marcati. Tuttavia, nessuna indagine empirica è riuscita a confermare questo scenario in modo chiaro e incontrovertibile. Ad esempio, uno studio su larga scala di Hough et al. (1990) ha esaminato gli effetti di moderazione delle scale di controllo utilizzando 10 indicatori temperamentali del questionario ABLE e tre misure di criterio esterne; gli autori hanno trovato conferme estremamente deboli e incoerenti.

Approcci IRT alla misurazione del *faking*

Alcuni studiosi hanno tentato di sviluppare nuove tecniche di misurazione del *faking* basate sull'*Item Response Theory* (IRT) e, più in particolare, sugli indici di *person fit*. Tali indici esprimono la probabilità di osservare un determinato *pattern* di risposte dato un certo modello IRT (Meijer & Sijtsma, 2001). In questo senso, gli indici di *person fit* fungono da rilevatori di anomalie. L'idea è che gli individui impegnati in condotte di *faking* — a causa degli effetti perturbativi del loro stile di risposta — producano *pattern* anomali/improbabili. Va osservato che alcuni di questi indici possono essere definiti *di congruenza* (e.g., Z_3) mentre altri *di discrepanza* (e.g., F_2); nel primo caso valori elevati indicano una più alta compatibilità del *pattern* di risposte con il modello IRT, nel secondo caso il ragionamento è invertito (Zickar & Sliter, 2012). Brown e Harvey (2003), Ferrando e Chico (2001), Zickar e Drasgow (1996) hanno impiegato tali indici con gruppi di individui cui era stato somministrato un questionario di personalità in due condizioni diverse: istruzioni standard (*honest group*) e istruzioni manipolate al fine di indurre comportamenti simulativi (*fake group*). I risultati si sono rivelati deludenti. Gli indici di *person fit* non hanno esibito performance superiori rispetto alle tradizionali scale di controllo.

Un'altra linea di ricerca ha riguardato l'uso di una tecnica statistica chiamata *Mixed Model - Item Response Theory* (MM-IRT; Zickar, Gibby e Robie, 2004); tale tecnica unisce la *Latent Class Analysis* (LCA) con i modelli IRT; concettualmente è simile alle analisi DIF/DTF con la differenza che i gruppi vengono individuati a posteriori invece che a priori. In una procedura DIF/DTF, il ricercatore (a) individua *ex ante* due o più gruppi di rispondenti nel campione in esame (e.g., maschi vs femmine); (b) e accerta l'eventuale funzionamento differenziale degli item rispetto a essi. Con la MM-IRT, la segmentazione del campione avviene a partire dai dati e secondo una strategia esplorativa che richiama alla mente la *cluster analysis*. I gruppi determinati in questo modo, e cioè *ex post*, condividono processi di risposta che risultano più omogenei. Utilizzando questa tecnica per analizzare un campione di individui che avevano risposto a un questionario di personalità proprietario, Zickar et al. (2004) hanno individuato tre classi di rispondenti, di cui due caratterizzate da comportamenti di *faking* rispettivamente moderati ed estremi. Gli autori hanno concluso che la MM-IRT rappresenta una tecnica promettente allo studio e alla misurazione delle condotte distorsive, ma hanno aggiunto che deve essere perfezionata prima di possibili applicazioni pratiche.

Faking e tempi di latenza.

Un ulteriore filone di indagini ha esaminato il *faking* in relazione alla latenze di risposta. Queste ultime riflettono il tempo che intercorre tra la presentazione di un item e l'atto di rispondere, atto che viene inteso come conclusione effettuale di una successione cognitiva (Hsu, Santelli & Hsu, 1989). Negli anni, sono state prospettate diverse ipotesi teoriche, tra loro contrastanti (Holden et al., 2001; Holtgraves, 2004; Vasilopoulos, Reilly & Leaman, 2000). Tali ipotesi muovono da un modello stadiale del processo di risposta che postula le seguenti fasi: (a) interpretazione dell'item; (b) recupero delle informazioni pertinenti; (c) formazione di un giudizio basato sui dati recuperati; (d) mappatura del giudizio sulla scala *likert* dell'item (McDaniel & Timm, 1990; Tourangeau & Rasinski, 1988).

Una prima prospettiva teorica suggerisce che il *faking* richieda tempo, portando a latenze di risposta più lunghe. In quest'ottica, la menzogna viene concettualizzata come un compito più oneroso della verità, con conseguente carico cognitivo maggiore (Vrij, Edward & Bull, 2001; Zuckerman, DePaulo & Rosenthal, 1981). Nella loro meta-analisi, DePaulo et al. (2003) hanno riferito che la menzogna è associata a latenze superiori quando le persone non possono preparare in anticipo le loro risposte. Tourangeau e Rasinski (1988) hanno argomentato che i tempi sono più lunghi poiché il processo di risposta subisce un'attività supplementare di *editing* basata sulla desiderabilità sociale. In aggiunta a ciò, McDaniel e Timm (1990) hanno suggerito che la simulazione introduca una componente dilatoria poiché impone una fase aggiuntiva di deliberazione che corrisponde alla decisione di mentire. Coerentemente con queste idee, McDaniel e Timm (1990) hanno scoperto che i tempi delle risposte di *faking* a un questionario anagrafico erano significativamente superiori se paragonati alle risposte oneste. Analogamente, in tre indagini sperimentali, Holtgraves (2004) ha rilevato che una crescente attenzione al contenuto degli item in rapporto alle norme sociali determinava un incremento temporale delle latenze.

Una seconda prospettiva teorica — diametralmente opposta alla prima — ipotizza che il *faking* causi latenze più brevi. Holtgraves (2004), Hsu et al. (1989) hanno congetturato che il processo di simulazione comporti un'elaborazione cognitiva ridotta/semplificata, suggerendo che i *faker* non si muovano attraverso tutte le fasi del processo di risposta. Più specificamente, Hsu et al. (1989) hanno affermato che un approccio basato sull'onestà obbliga il rispondente a una forma temporalmente dispendiosa di auto-riflessione (e.g., per valutare lo scarto tra la realtà descritta nell'item e il sé), mentre le risposte

di *faking* implicano un'interpretazione puramente semantica che richiede minori tempi di elaborazione. Holtgraves (2004) ha osservato che i mentitori non cercano di recuperare informazioni personali accurate, ma forniscono risposte basate esclusivamente sulla desiderabilità sociale degli item. Coerentemente con questa prospettiva, Hsu et al. (1989) hanno riferito latenze più brevi nei *faker*. Inoltre, Holden, Fekken e Cotton (1991) hanno trovato una correlazione negativa tra tempi di risposta e desiderabilità sociale. A sostegno dell'idea che i comportamenti di *faking* siano più veloci, Holden et al. (2001) hanno dimostrato che l'imposizione esterna di vincoli temporali non impediva alle persone di fingere se motivate a farlo.

Una terza prospettiva, descritta da Holden e Kroner (1992), sostiene che gli effetti del *faking* sulle latenze siano variabili e dipendano dalla congruità tra la strategia di simulazione adottata e la desiderabilità sociale dei singoli item. Ad esempio, quando gli individui impiegano uno stile di risposta di tipo *faking good*, gli item socialmente desiderabili — concettualmente allineati al tipo di finzione veicolata — sono più facili da affrontare, mentre gli item socialmente indesiderabili risultano cognitivamente più difficili: facilità e difficoltà elaborativa incidono proporzionalmente sui tempi di latenza. Diversi studi hanno confermato questo modello ibrido che sembra riassumere, integrandoli, i primi due (e.g., Brunetti, Schlottmann, Scott & Hollrah, 1998; Holden & Kroner, 1992).

1.2 Psicopatia

Le prime formulazioni del concetto di psicopatia fanno riavvolgere il nastro del tempo fino a due secoli fa, a testimonianza di un interesse quasi coevo alla nascita della psichiatria moderna.

Nel 1801, il medico francese Pinel riconobbe che diversi suoi pazienti esibivano comportamenti crudeli, violenti e impulsivi, caratterizzati da una furia che lo psichiatra non esitò a definire cieca (Millon, Simonsen & Birket-Smith, 1998). Ciononostante, questi pazienti non presentavano alcuna compromissione delle funzioni intellettive, rimanendo pienamente consapevoli della gratuità e distruttività delle azioni commesse. Per loro, Pinel usò la locuzione *manie sans délire*.

Qualche anno più tardi, nel 1812, lo psichiatra americano Rush scrisse un saggio su quella che definì *perversion of the moral faculties*. Come Pinel, Rush descrisse soggetti socialmente predatori e distruttivi, privi di rimorso, sensi di colpa o preoccupazioni per le conseguenze dei propri atti; l'autore americano sottolineò la natura profondamente antisociale di questi individui, connotando in senso morale l'osservazione clinica del disturbo. Nello stesso torno di tempo e prendendo le mosse da Rush, Prichard, medico britannico, ampliò la descrizione della psicopatia fino a includere tutte quelle condizioni che rendevano i soggetti incapaci di conformarsi alle convenzioni e alle norme sociali, attribuendo l'etichetta di *moral insanity* a un ventaglio di disturbi molto diversi tra loro; operazione concettuale che rese evanescenti i confini epistemologici del disturbo psicopatico. Nonostante la sua scarsa specificità, il concetto di *moral insanity* continuò ad alimentare l'interesse degli studiosi europei nei decenni successivi. Nel 1874, Maudsley ipotizzò l'esistenza di uno specifico centro cerebrale sede di tutti i sentimenti morali (evidentemente deficitario negli psicopatici). A questa nozione anatomica furono associate evidenze fisiognomiche in linea con l'allora dominante pensiero lombrosiano sulle stigmate degenerative del delinquente nato.

Lo psichiatra Koch fu tra i primi studiosi tedeschi a usare il termine psicopatia nella letteratura medica germanica alla fine del XIX secolo, più precisamente nel 1881. Contrariamente agli alienisti inglesi, Koch associò la psicopatia a tutte quelle condizioni di disagio che oggi verrebbero definite come disturbi di personalità. Egli coniò il termine *psychopathische minderwertigkeiten* e vi incluse tutte *"le irregolarità, congenite o acqui-*

site, che influenzano una persona nell'arco di vita e la fanno sembrare, anche nei casi più favorevoli, non pienamente in possesso di capacità mentali normali." (Millon et al., 1998, p. 8).

Con l'introduzione della coppia speculare sadismo/masochismo nella sua opera *Psychopathia Sexualis* del 1886, il tedesco Kraft-Ebing catturò alcune caratteristiche cliniche vicine alla psicopatia, quali la dominazione a scopi di sopraffazione e la crudeltà. L'autore postulò l'esistenza di un desiderio sadico innato (con il suo complementare masochistico) le cui radici erano da ricercarsi in un'esacerbazione degli impulsi sessuali di natura aggressiva. Secondo l'autore gli individui psicopatici mostravano una più spiccata propensione ad agire sotto l'imperio di questi impulsi.

A cavallo tra il XIX ed il XX secolo, le diverse edizioni del manuale *Psychiatrie: Ein Lehrbuch* di Kraepelin suscitarono un rinnovato interesse conoscitivo per il disturbo psicopatico. Nella seconda edizione del suo lavoro (1887), lo psichiatra tedesco identificò il malato di mente morale come un soggetto costituzionalmente incapace di inibire l'immediata e sfrenata gratificazione di desideri egoistici e aggressivi. Nella quinta edizione (1896), si riferì a tale condizione con la locuzione *psychopathische zustände*, mentre nella settima (1903) adoperò il termine *psychopathische persönlichkeit*, a indicare un disturbo congenito-degenerativo delle funzioni affettive. Nell'ottava edizione del suo manuale (1903-1904), Kraepelin descrisse gli psicopatici come carenti negli affetti e nella volizione, raggruppandoli in due categorie: gli psicopatici con disposizioni morbose (ossessivi, impulsivi e devianti sessuali) e gli psicopatici con personalità peculiare. Quest'ultimo gruppo fu ulteriormente suddiviso in sette sottotipi: l'eccitabile, l'instabile, l'impulsivo, l'eccentrico, il bugiardo/truffatore, l'antisociale e il litigioso.

Nel 1909, un altro psichiatra tedesco, Birnbaum, propose di usare il termine sociopatico per designare buona parte degli individui descritti nella letteratura clinica sulla psicopatia, enfatizzando l'origine psico-sociale del disturbo. Egli infatti riteneva che solo alcuni delinquenti appartenenti ai gruppi Kraepeliani sopra menzionati fossero inclini alla criminalità su base costituzionale; la devianza in tutti gli altri casi derivava, secondo l'autore, dalla mancata acquisizione di comportamenti pro-sociali, in un'ottica di condizionamento ambientale più che di predisposizione biologica.

Schneider fu tra i più importanti teorici della psichiatria germanica dopo la prima guerra mondiale a occuparsi di psicopatia. Nella sua opera *Psychopatische Persönlich-*

keiten (1933), lo psichiatra tedesco espresse la convinzione che molti criminali fossero delinquenti fin dalla gioventù e in gran parte incorreggibili, ma che altri potessero trovarsi nella società in generale, anche in posizioni di potere politico o materiale. Individuò dieci tipi di psicopatici: l'ipertimico, il depressivo, l'insicuro, il diffidente, il fanatico, colui che ricerca attenzioni, il labile, l'esplosivo, l'anaffettivo, il sottomesso/astenico. Per Schneider, quindi, la personalità psicopatica equivaleva grosso modo all'insieme dei disturbi di personalità. Tra i dieci tipi sopra descritti è quello anaffettivo ad avvicinarsi maggiormente al concetto odierno di psicopatia. Per dirla con Schneider, questo tipo: *"manca di compassione, vergogna, onore, rimorso e coscienza. Come risultato, la sua personalità è spesso sinistra, fredda e scontrosa e la condotta brutale e sfrenata. Oltre a coloro che formano il gruppo criminale nella nostra società, questo tipo può essere rintracciato anche nella società in generale. Sono persone incallite e fredde che a volte possono essere assolutamente spietate e in cui l'intelligenza, lungi dall'essere carente, è spesso notevolmente alta."* (Crowhurst & Coles, 1989, p. 239).

Il 1941 segnò lo spartiacque tra le prime concettualizzazioni — che trasformarono il costrutto della psicopatia in una disordinata *collectanea* di significati per un ampio ventaglio di sindromi cliniche e disturbi di personalità — e la concezione moderna molto più specifica. In quell'anno, lo psichiatra americano Cleckley pubblicò il libro *The Mask of Sanity*, notevole successo editoriale tanto da venire ristampato in 4 edizioni successive, l'ultima nel 1976. In questo libro, l'autore tracciò l'identikit ideale dell'individuo psicopatico sulla base della sua esperienza clinica con soggetti istituzionalizzati. Patrick (2006) ha riassunto le caratteristiche salienti dei casi presentati da Cleckley, notando in particolare che (1) la mancanza di ansia era evidente nella maggior parte di questi casi; (2) il comportamento ostile-aggressivo era una caratteristica importante solo in una minoranza di essi; (3) altri tipi di comportamento deviante (ad esempio, frodi, furti, falsificazioni, incendi, reati di droga, ubriachezza disordinata e molesta, atti vandalici, assenze dal lavoro ingiustificate, guida spericolata) erano presenti in quasi tutti i casi e contraddistinti da particolare futilità.

Nella tabella 1.2 sono riassunti i 16 criteri proposti da Cleckley per descrivere la personalità psicopatica, raggruppati in tre categorie (Patrick, 2006). La prima è costituita dalle caratteristiche cosiddette *di maschera* che distinguono gli psicopatici dagli altri pazienti psichiatrici: buona intelligenza e fascino sociale; assenza di nervosismo, di deli-

ri/irrazionalità; suicidio raramente commesso (tabella 1.2, parte superiore). Vale la pena di notare che Cleckley, nel presentare il concetto di maschera, fece riferimento non solo all'assenza di sintomi nevrotici/psicotici, ma anche alla presenza di un equilibrio sociale ed emotivo tanto forte quanto illusorio: *"In superficie, lo psicopatico [...] si presenta in modo normale e non dà alcun indizio di disagio interiore. Niente di lui suggerisce stranezze, inadeguatezza o fragilità morale. La sua maschera è quella di una sana condizione mentale"* (p. 383). Tuttavia, le mostre di questa salute mentale fittizia sono accompagnate da una persistente e grave devianza comportamentale: *"Lo psicopatico, per quanto imiti perfettamente l'uomo adattato — almeno quando parla di sé — fallisce miseramente nella vita reale. Il suo fallimento è così completo e drammatico che è difficile vedere come esso possa essere ottenuto da chiunque non sia meno che pazzo."* (p. 370). Questo aspetto di devianza del disturbo è rappresentato da una seconda serie di indicatori che comprendono atti antisociali impulsivi, scarso giudizio, irresponsabilità, promiscuità sessuale e assenza di un chiaro progetto esistenziale (tabella 1.2, parte centrale). Insieme alle caratteristiche di maschera e di devianza, i criteri di Cleckley includevano una terza serie di indicatori affettivi e interpersonali, tra cui incapacità di amare, mancanza di lealtà o reciprocità sociale, falsità, inautenticità, assenza di rimorsi (tab. 1.2, parte inferiore).

Il lavoro di Cleckley — sebbene illuminato da intuizioni sorprendenti e tuttora valide — non proponeva un metodo di misurazione della psicopatia; quest'ultima rimase confinata entro la sfera della speculazione/osservazione clinica fino alla comparsa della *Psychopathy Checklist* (PCL; Hare, 1980), strumento concepito dallo psicologo canadese nel tentativo di operazionalizzare i criteri di Cleckley. Egli sviluppò la PCL, oggi edita nella forma rivista a 20 item denominata PCL-R (Hare, 2003), come strumento per compiere una valutazione della psicopatia mediante l'utilizzo di informazioni raccolte a partire da un'intervista semi-strutturata. Nella tabella 1.3 sono riportati i venti item della PCL-R raggruppati in due fattori e quattro sfaccettature secondo quanto emerso da recenti analisi fattoriali esplorative e confermative (per una rassegna vedi Neumann, Hare & Pardini, 2015). Muovendo da questo modello, si può dire che lo psicopatico di Hare sia definibile sulla base delle sue caratteristiche affettivo-interpersonali (fattore 1), e sulla base del suo comportamento deviante (fattore 2). Il Fattore 1 descrive una costellazione di tratti caratterologici in cui prevalgono le condotte manipolatorie unite a un'affettività superficiale il cui dato più eclatante è la mancanza di partecipazione emotiva. Il fattore 2 invece si caratterizza per la presenza di comportamenti antisociali e di uno stile di

Tabella 1.2: I critieri di Cleckley (secondo Patrick, 2006).

Categoria	Criterio	Testo
Maschera	01	Fascino superficiale e buona "intelligenza"
	02	Assenza di deliri e di altri segni di pensiero irrazionale
	03	Assenza di "nervosismo" o di manifestazioni psiconevrotiche
	14	Suicidio raramente portato a termine
Comp. devianti	07	Motivazione inadeguata dei comportamenti antisociali
	08	Scarso giudizio e incapacità di apprendere dall'esperienza
	04	Inaffidabilità
	13	Comportamento bizzarro o sgradevole in stato di ebbrezza alcolica e talora indipendentemente da essa
	15	Vita sessuale impersonale, promiscua, scarsamente integrata
	16	Incapacità di seguire un progetto esistenziale
Superficialità e inganno	05	Falsità e inautenticità
	06	Mancanza di rimorso o vergogna
	10	Povertà complessiva nelle reazioni affettive più importanti
	09	Egocentricità patologica e incapacità di amare
	11	Mancanza specifica di insight
	12	Insensibilità nella generalità delle relazioni interpersonali

vita impulsivo, irresponsabile e parassitario. Entrambi i fattori compongono l'immagine di un predatore scaltro, incallito, senza scrupoli, impegnato in un'opera di sistematica spoliazione del mondo, in un contesto emotivo che si potrebbe definire sterile nelle sue componenti positive e fertile invece in quelle negative.

Tabella 1.3: Gli item della PCL-R (Hare, 2003).

Fattore	Sfaccettatura	Item	Testo
Fattore 1	Affetti	06	Assenza di rimorso o di senso di colpa
		07	Affettività superficiale
		08	Insensibilità / Mancanza di empatia
		16	Incapacità di accettare la responsabilità delle proprie azioni
	Interpersonale	01	Loquacità / Fascino superficiale
		02	Senso di sé grandioso
		04	Menzogna patologica
		05	Impostore / Manipolativo
Fattore 2	Comp. devianti	10	Deficit del controllo comportamentale
		12	Problematiche comportamentali precoci
		18	Delinquenza in età giovanile
		19	Revoca della libertà condizionale
		20	Versatilità criminale
	Stile di vita	03	Bisogno di stimoli / Propensione alla noia
		09	Stile di vita parassitario
		13	Mancanza di obiettivi realistici / a lungo termine
		14	Impulsività
		15	Irresponsabilità
N.a.	N.a.	11	Comportamento sessuale promiscuo
		17	Numerosi rapporti di coppia di breve durata

Una controversia importante nella concettualizzazione del disturbo psicopatico è il peso della componente antisociale (Lilienfeld, 1994). Molte ricerche hanno confermato come la psicopatia sia un fattore di rischio per la violenza e le condotte criminali recidivanti (Salekin, Rogers & Sewell, 1996), ma alcuni studiosi hanno messo in discussione la validità del legame ritenendolo gravato da una fallacia di natura tautologica, dato che il comportamento antisociale — incluso nella definizione del costrutto — viene anche usato

come criterio esterno di validazione (Andrade, 2008). Inoltre, altri ricercatori hanno sostenuto la necessità di attribuire all'antisocialità una qualità epifenomenica e quindi non specifica o patognomonica (Skeem & Cooke, 2010). Alla luce di quanto precede, Cooke, Hart, Logan e Michie (2012) hanno concepito un nuovo modello teorico della psicopatia, denominato *Comprehensive Assessment of Psychopathic Personality* (CAPP) e basato su una tecnica che ricorda l'approccio psico-lessicale proposto da Goldberg (1992) nello sviluppo dei *Big Five*. Il modello CAPP cerca di superare le limitazioni menzionate a inizio paragrafo concentrandosi meno sulle caratteristiche comportamentali e più sulle qualità dinamiche e processuali della personalità psicopatica. Secondo tale modello, il disturbo comprende sei domini sintomatici per un totale di 33 diversi indicatori (vedi tabella 1.4).

Tabella 1.4: Modello CAPP della psicopatia (Cooke et al., 2012).

Domain	Facet	Descriptors
Attachment	Detached	Remote, Distant, Cold
	Uncommitted	Unfaithful, Undevoted, Disloyal
	Unempathic	Uncompassionate, Cruel, Callous
	Uncaring	Inconsiderate, Thoughtless, Neglectful
Behavioural	Lacks perseverance	Idle, Undisciplined, Unconcientious
	Unreliable	Undependable, Untrustworthy, Irrespons.
	Reckless	Rush, Impetuous, Risk-taking
	Restless	Overactive, Fidgety, Energetic
	Disruptive	Disobedient, Unruly, Unmanageable
	Aggressive	Threatening, Violent, Bullying
Cognitive	Suspicious	Distrustful, Guarded, Hyper-vigilant
	Lacks concentration	Distractible, Inattentive, Unfocused
	Intolerant	Narrow-minded, Bigoted, Hypercritical
	Inflexible	Stubborn, Rigid, Uncompromising
	Lacks planfulness	Aimless, Unsystematic, Disorganized
Dominance	Antagonistic	Hostile, Disagreeable, Contemptuous
	Domineering	Arrogant, Overbearing, Controlling
	Deceitful	Dishonest, Deceptive, Duplicitous
	Manipulative	Devious, Exploitative, Calculating
	Insincere	Superficial, Slick, Evasive
	Garrulous	Glib, Verbose, Pretentious
Emotional	Lacks anxiety	Unconcerned, Unworried, Fearless
	Lacks pleasure	Pessimistic, Gloomy, Unenthusiastic
	Lacks emotional depth	Unemotional, Indifferent, Inexpressive
	Lacks emotional stab.	Temperamental, Moody, Irritable
	Lacks remorse	Unrepentant, Unapologetic, Unashamed
Self	Self-centered	Egocentric, Selfish, Self-absorbed
	Self-aggrandizing	Self-important, Conceited, Condescending
	Sense of uniqueness	Being extraordinary, Exceptional, Special
	Sense of entitlement	Demanding, Insistent, Being deserving
	Sense of invulnerability	Invincible, Indestructible, Unbeatable
	Self-justifying	Minimizing, Denying, Blaming
	Unstable self-concept	Labile, Incomplete, Chaotic sense of self

1.2.1 La rappresentazione della psicopatia nel DSM

Nella prima edizione del Manuale Diagnostico e Statistico dei Disturbi Mentali (DSM-I; APA, 1952), il costrutto della psicopatia richiamava — sebbene solo in modo narrativo e parziale — le descrizioni cliniche di Cleckley (Lewis, 2010). Per evitare confusione con termini simili, come psicotici, il comitato decise di usare la locuzione *disturbo sociopatico della personalità: reazione antisociale*.

Il DSM-II (APA, 1968) fu pubblicato 16 anni dopo. Secondo Bodholdt, Richards e Gacono (2000) questa edizione mirava a ridurre l'ambiguità diagnostica fornendo indicazioni più chiare. In particolare, il comitato ampliò la descrizione della psicopatia aggiungendo l'insensibilità, l'impulsività, l'egoismo, la mancanza di colpa, la bassa tolleranza alla frustrazione, l'esternalizzazione della colpa. Nel DSM-II, la psicopatia subì un ulteriore cambio di nome, passando da *disturbo sociopatico della personalità - reazione antisociale* a *personalità antisociale*.

Nel 1980 l'APA pubblicò il DSM-III. Questa edizione fu concepita con un approccio volto ad aumentare l'accessibilità, l'affidabilità e l'applicabilità del manuale. La psicopatia cambiò nuovamente nome da *personalità antisociale* a *disturbo antisociale della personalità* (ASPD) e perse molti dei suoi indicatori caratterologici, a favore di un'enfasi sui comportamenti antisociali manifesti. Per ricevere una diagnosi di ASPD, un individuo doveva avere avuto una storia di almeno tre comportamenti devianti prima dell'età di 15 anni e mostrare un modello pervasivo di criminalità (APA, 1980). Questa modifica fu apportata in quanto si riteneva che gli aspetti comportamentali fossero più facili da definire e valutare rispetto ai tratti di personalità (Bodholdt et al., 2000). Impulsività e Irresponsabilità furono gli unici elementi caratterologici a non essere espunti. Secondo Blackburn (2007), il DSM-III vide il ritorno larvato del concetto di *moral insanity*, dato che i criteri diagnostici dell'ASPD si concentravano prioritariamente sulle condotte trasgressive.

Per la realizzazione del DSM-IV nel 1994, l'APA effettuò numerosi studi sul campo ponendosi come scopo la verifica delle revisioni via via proposte (Gurley, 2009). Uno di questi esaminò le modifiche da apportare all'ASPD. Hare, Hart e Hempur (1991) hanno osservato che tale studio si prefiggeva lo scopo di valutare se il disturbo dovesse includere nuovamente i tratti di personalità associati alla psicopatia e se i suoi criteri potessero essere abbreviati in qualche modo. I risultati misero in luce da un parte la possibilità di

semplificare la diagnosi e dall'altra la necessità di incorporare alcuni elementi della PCL (Widiger et al. 1996). Le modifiche dell'APA furono però criticate perché ritenute troppo inclusive e dunque scarsamente valide in termini diagnostici e predittivi (Gurley, 2009). Infatti, esaminando il legame ASPD/Psicopatia, Hildebrand e de Ruiter (2004) hanno notato che i due costrutti erano asimmetricamente correlati: tutti i pazienti con diagnosi di psicopatia soddisfacevano i criteri dell'ASPD, mentre non era vero il contrario.

L'ultima edizione del manuale, il DSM-V (APA, 2013) non riconosce ancora la psicopatia come entità nosografica autonoma, continuando a includere molti dei suoi aspetti nella sezione *Caratteristiche associate a supporto della diagnosi* del disturbo ASPD, tuttora fortemente incentrato su indicatori di devianza comportamentale. Anche se il quadro diagnostico categoriale non ha subito cambiamenti rispetto al DSM-IV, è stato introdotto un approccio dimensionale alternativo basato sui tratti (*Modello alternativo del DSM-V per i disturbi di personalità*); nell'ambito di questo nuovo approccio, la descrizione dell'ASPD si riavvicina, per così dire, alla personalità psicopatica per come fu immaginata da Cleckley e perfezionata dagli studiosi che lo seguirono. Ciononostante, secondo Strickland, Drislane, Lucy, Krueger e Patrick (2013), l'ASPD nell'ottica dimensionale non comprende ancora tutti i criteri considerati prototipici della psicopatia.

1.2.2 Misurazione della psicopatia

Psychopathic Checklist - Revised (PCL-R)

Gli ultimi decenni hanno testimoniato importanti progressi nell'*assessment* della psicopatia; ciò si deve in larga parte allo sviluppo della *Psychopathic Checklist* (PCL; Hare, 1980) e alle sue successive revisioni culminanti nella PCL-R (Hare, 1991; Hare, 2003). Lo strumento è oggi considerato come l'*aurum vexillum* per la misurazione del disturbo (Cooke & Michie, 1997). È costituito da 20 item[6] e utilizza un'intervista semi-strutturata con l'aggiunta di informazioni collaterali (e.g., consultazione del casellario giudiziario). La scala *likert* (da 0 a 2) adottata per valutare gli item esprime in che misura il contenuto di ciascuno di essi si applica all'individuo e alla sua storia. Il punteggio totale varia da 0 a 40 e riflette il grado con cui l'esaminato corrisponde all'idealtipo psicopatico.

[6] Consulta tabella 1.3 (p. 18) per l'elenco completo degli item.

La PCL-R si è dimostrata affidabile e valida; negli anni è andata accumulandosi una corposa messe di articoli e capitoli di libri dedicati all'analisi delle sue proprietà psicometriche (per una rassegna vedi Häkkänen-Nyholm e Nyholm, 2012). La consistenza interna è elevata ($\alpha > 0.85$), la correlazione intraclasse supera in genere il valore di 0.85, mentre l'errore standard di misurazione del punteggio totale è di circa 2/3 punti (Hare, 2003). Sebbene siano state riscontrate differenze di genere, etniche e culturali (e.g., nel modo in cui si manifestano alcune caratteristiche della psicopatia), la PCL-R ha dimostrato una buona generalizzabilità (Malterer, Lilienfeld, Neumann & Newman, 2010). Ricerche recenti suggeriscono che il costrutto alla base dello strumento sia di natura dimensionale, ma il punteggio *cutoff* di 30 si è dimostrato utile come definizione operativa di psicopatia (Hare, 2003).

Alcuni ricercatori hanno espresso le loro preoccupazioni in merito all'instabilità della struttura fattoriale dello strumento. I primi studi hanno infatti messo in luce un modello a due fattori (*interpersonal/affective*, *social deviance*; Hare et al., 1990), mentre ricerche più recenti hanno proposto una soluzione a tre fattori *(interpersonal*, *affective*, *lifestyle*; Cooke, Kosson e Michie, 2001) e negli ultimi anni lo stesso Hare (2003) ha suggerito un'organizzazione quadripartita degli item (*interpersonal*, *affective*, *lifestyle*, *antisocial*). Neal e Sellbom (2012) hanno sottolineato che il dibattito sulla struttura della PCL-R rimane ancora aperto.

Dello strumento sono disponibili una versione di *screening* abbreviata a 12 item (Hart, Cox & Hare, 1995) e una versione per adolescenti (Forth, Kosson & Hare, 2003).

Questionari *Self-Report*

La procedura di somministrazione e di valutazione della PCL-R è dispendiosa in termini di tempo e di informazioni da processare e richiede una formazione più che adeguata. Per questi motivi, l'uso della PCL-R potrebbe risultare inappropriato in alcuni ambiti specifici (e.g., studi con popolazione non carceraria; Copestake, Gray e Snowden, 2011). Da qui, l'esigenza di sviluppare questionari *self-report*.

Secondo Lilienfeld, Fowler, Katherine e Patrick (2006), la valutazione della psicopatia tramite *self-report* è relativamente veloce, economica e può fornire interessanti dati supplementari (e.g., stili di risposta misurati attraverso le scale di controllo). Tuttavia, vi

sono degli svantaggi insiti in tale modalità, svantaggi che hanno spinto alcuni esperti a ritenere i questionari inadatti alla valutazione del disturbo. Occorre ad esempio osservare che gli individui psicopatici hanno la tendenza a mentire e a manipolare gli altri (per propri scopi o per futili motivi) e non sono in grado compiere un'esame accurato del loro universo interiore, sia per volontà di mascheramento che per scarso *insight* (Lilienfeld et al., 2006). Pur riconoscendo queste limitazioni, gli autori affermano che le risposte fornite dagli psicopatici, anche se non veridiche o solo parzialmente tali, rappresentano comunque degli importanti indicatori verbali dotati di intrinseca utilità diagnostica (Lilienfeld et al., 2006).

Forti di questa argomentazione, i ricercatori del settore hanno sviluppato diversi questionari *self-report* per la misurazione della psicopatia. Tra gli strumenti attualmente più usati (Lewis, 2010) figurano i seguenti (in ordine cronologico): *Levenson Self-Report Psychopathy Scale* (LRSP; Levenson, Kiehl e Fitzpatrick, 1995), *Psychopathic Personality Inventory - Revised* (PPI-R; Lilienfeld e Widows, 2005), *Elemental Psichopathy Assessment* (EPA; Lynam et al., 2011), *Self-Report Psychopathy Scale* (SRP-III; Paulhus e Neumann, 2013), *Triarchic Psychopathy Measure* (TRIPM; Brislin, Drislane, Smith, Edens e Patrick, 2015), *Comprehensive Assessment of Psychopathic Personality - Self Report* (CAPP-SR; Sellbom, Cooke e Shou, 2019). Nella sezione che segue, verrà descritto il PPI-R, dato che è stato impiegato nello studio 2 della presente ricerca.

Questionario PPI-R

La prima versione del PPI (Lilienfeld & Andrews, 1996) è stata concepita per rilevare i tratti chiave della personalità psicopatica. La scelta del *pool* iniziale di item è dipesa da una revisione completa della letteratura clinica e di ricerca, tra cui gli scritti fondamentali di Cleckley. Nel tentativo di distinguere la psicopatia da costrutti concettualmente vicini, ma separati (ad esempio, il disturbo antisociale), sono stati esclusi tutti gli item afferenti all'area della devianza e della criminalità. Il *pool* iniziale è stato progressivamente ridotto fino a 187 item mediante analisi fattoriali su tre campioni di studenti universitari ($N = 1.156$). In tutte e tre i campioni sono emersi otto fattori (scale di contenuto) che compongono il nucleo interpretativo del PPI[7]: (1) *Egocentrismo machiavellico (ME)* che

[7] Oltre alle scale di contenuto il PPI possiede tre scale di controllo: (1) IR Risposte inconsistenti; (2) VR Risposte virtuose; (3) DR Risposte devianti.

misura la propensione del soggetto a manipolare gli altri per conseguire obiettivi personali e ad avere una visione cinica, disillusa e strumentale della natura umana. (2) *Anticonformismo ribelle (RN)* che misura la tendenza verso la non convenzionalità, la presenza di atteggiamenti ostili verso l'autorità e la resistenza alle norme sociali. (3) *Esternalizzazione della colpa (BE)* che misura la percezione che il soggetto ha del mondo esterno, di quanto lo avverta ingiusto e ostile, di quanto lo reputi causa dei suoi problemi. (4) *Mancanza di pianificazione (CN)* che misura la predisposizione a porre in essere comportamenti impulsivi. (5) *Influenza sociale (SOI)* che misura l'attitudine del soggetto a essere affascinante, attraente e abile nell'influenzare gli altri. (6) *Mancanza di paura (F)* che misura l'assenza di ansia anticipatoria riguardo a danni fisici e l'interesse per attività rischiose. (7) *Immunità allo stress (STI)* che misura la capacità del soggetto di rimanere calmo/lucido di fronte a stimoli stressogeni. (8) *Freddezza emotiva (C)* che misura l'assenza di legami e di sentimenti profondi (come senso di colpa, empatia. ecc.), nonché l'incapacità di mantenere nel tempo relazioni significative con altre persone.

Il PPI è stato aggiornato nel 2005 (Lilienfeld & Widows, 2005) al fine di ridurne la lunghezza, di facilitarne la comprensione, di eliminare gli item psicometricamente deboli o dipendenti dal contesto culturale e di sviluppare norme specifiche per la popolazione generale e quella carceraria. La versione riveduta del test (PCL-R) è composta da 154 item suddivisi nelle stesse otto scale di contenuto e nelle tre scale di controllo del PPI originario. Lilienfeld e Widows (2005) hanno condotto una serie di analisi fattoriali esplorative e confermative ottenendo tre fattori gerarchicamente superiori alle scale di contenuto. Un fattore, chiamato *Dominanza priva di paura (FD)*, un altro denominato *Impulsività autocentrata (SCI)* e l'ultimo corrispondente alla scala di contenuto *Freddezza emotiva (C)* che non saturava nei primi due fattori.

L'affidabilità test-retest del punteggio totale PPIR è di 0.93, mentre quella delle scale di contenuto varia da 0.82 a 0.95. La coerenza interna del punteggio totale PPIR è > 0.80 (α di Cronbach), mentre per le scale di contenuto essa è compresa tra 0.71 e 0.87. Numerosi studi su campioni universitari e carcerari hanno confermato la validità di costrutto del PPI e del PPI-R. I punteggi totali di entrambi gli strumenti: (a) correlano in modo significativo con altre misure *self-report* di psicopatia; (b) correlano moderatamente con misure di disturbi della personalità che notoriamente si sovrappongono alla psicopatia, come il disturbo narcisistico, istrionico e borderline, ma debolmente con tutti gli altri; (c)

correlano positivamente con misure che riguardano i comportamenti criminali o di abuso di sostanze e misure di laboratorio sul discontrollo degli impulsi (Lilienfeld & Widows, 2005).

Psicopatia e *faking*

In questa sezione del capitolo presentiamo una revisione qualitativa della letteratura empirica esistente. Come premessa iniziale, si può dire che i ricercatori hanno approfondito due temi principali, riassumibili nelle seguenti domande: (a) gli psicopatici/ASPD mentono con più frequenza? (b) gli psicopatici/ASPD sono in grado di mentire meglio? I precedenti quesiti sono assolutamente legittimi se si considera che la psicopatia è contraddistinta da una *cluster* di tratti personologici che spinge gli individui a essere elusivi, tendenziosi, mendaci, e a violare — per i propri scopi, anche futili — quel principio di cooperazione che rende la comunicazione umana trasparente e soprattutto utile per entrambe le parti (Grice, Cole, Morgan et al., 1975).

Gli psicopatici mentono di più?

Diversi studi hanno trovato sostegno all'idea che la psicopatia sia correlata con una maggiore probabilità di mentire (Cima et al., 2003; Heinze e Vess, 2005; Kucharski, Duncan, Egan e Falkenbach, 2006; Delain, Stafford e Ben-Porath, 2003; Grillo, Brown, Hilsabeck, Price e Lees-Haley, 1994). Ad esempio, Cima et al. (2003) hanno esaminato l'associazione tra *faking bad* e psicopatia in un campione di 188 criminali maschi. Il *faking bad* è stato valutato con più misure, tra cui l'MMPI2 (Butcher, Dalstrom, Graham, Tellegen & Kraemmer, 1989), Il PAI (Morey, 1996), e l'Intervista Strutturata dei Sintomi Simulati (SIRS; Rogers, Gillis, Dickens e Bagby, 1991). È stata impiegata la PCL-R (Hare, 2003) come strumento per la misurazione dei tratti psicopatici. I punteggi PCL-R sono stati discretizzati in tre fasce: A - valori bassi (< 20), B - moderati ($20|29$) e C - elevati (> 29). Sono stati presi in considerazione sia il Fattore PCL-R 1, che copre i tratti interpersonali e affettivi (e.g., manipolazione e insensibilità), sia il Fattore PCL-R 2, che riflette la componente antisociale della psicopatia. Rispetto agli altri, gli individui del gruppo C presentavano un'elevazione significativa dei punteggi delle scale di controllo dell'MMPI2, del PAI e della SIRS. Inoltre, gli autori hanno scoperto che il Fattore PCL-R 1 era è

più correlato con il *faking bad* rispetto al Fattore 2. Questi risultati tenderebbero a confermare l'idea che gli psicopatici — a causa dei loro tratti manipolativi (Fattore PCL-R 1) — simulino più frequentemente e in modo più marcato. Tuttavia, gli autori hanno anche notato che una certa quota di individui psicopatici era caratterizzata da assenza di *faking*. Un'evidenza simile invita a usare grande cautela nel considerare la psicopatia come sinonimo di condotte distorsive. In un altro studio, Cima e van Oorsouw (2013) hanno esaminato il rapporto tra psicopatia e il *faking bad* in un campione di 131 detenuti. Gli autori hanno impiegato il PPI (Lilienfeld & Andrews, 1996) per valutare la psicopatia e il *Structured Inventory of Malingered Symptomatology* (SIMS; Smith e Burger, 1997) è stato somministrato per rilevare gli stili di risposta distorsivi. Al contrario di Kucharski et al. (2006), gli autori del presente studio hanno osservato che il *faking bad* era significativamente correlato con il Fattore PCL-R 2, ma non con il Fattore PCL-R 1. In sintesi, le precedenti ricerche — pur avendo trovato un legame tra psicopatia e *faking* — hanno mancato di restituire un quadro interpretativo coerente del fenomeno. Altre indagini empiriche non hanno rilevato alcuna associazione tra psicopatia e simulazione. Ad esempio, Pierson, Rosenfeld, Green e Belfi (2011) hanno somministrato la SCID-II (First, Gibbon, Spitzer, Williams & Benjamin, 1997) per valutare la presenza di ASPD/Psicopatia e la SIRS (Rogers et al., 1991) per misurare il *faking* in un campione di pazienti psichiatrici ristretti in contesto di massima sicurezza. Dei 71 pazienti esaminati, 28 soddisfacevano i criteri per l'ASPD. I risultati dell'indagine hanno evidenziato come i partecipanti con diagnosi di ASPD non differissero dagli altri nei loro punteggi SIRS. Ciò dimostra che la diagnosi ASPD/Psicopatia non implica l'automatica presenza di stili di risposta distorsivi. In un altro studio, Cima et al. (2003) hanno esaminato l'associazione tra psicopatia e *faking good* reclutando un gruppo di pazienti in ambito forense. I partecipanti hanno completato il PPI (Lilienfeld & Andrews, 1996) e la *Supernormality Scale-Revised* (SS-R; Cima et al., 2003) come indici di psicopatia e di *faking*. Gli autori hanno accertato che il punteggio totale PPI era inversamente correlato con quello della SS-R. In contrasto con questi risultati, Freeman e Samson (2012) non hanno rilevato alcuna associazione significativa tra dissimulazione e psicopatia, in seguito alla somministrazione del *Balanced Inventory of Desiderable Responding* (BIDR; Paulhus, 1988) e della *Self-Report Psychopathy Scale-III* (SRP-III; Paulhus e Neumann, 2013) a un campione di 300 detenuti.

Gli psicopatici mentono meglio?

Alcuni autori hanno sostenuto che gli psicopatici sanno mentire bene perché non si sentono in colpa quando lo fanno e dunque lo fanno spesso (Porter, ten Brinke & Wallace, 2012). Seguendo questa linea di pensiero tali individui dovrebbero essere più capaci degli altri di simulare quadri sintomatici o mascherarne la presenza se questo torna a loro vantaggio. Tuttavia, diverse indagini empiriche hanno disconfermato quest'ipotesi per carenza di risultati significativi. In altri termini, non è stato trovato sostegno alla *vulgata* secondo cui gli psicopatici sarebbero abili mentitori, capaci di eludere qualunque meccanismo di controllo (Boone et al., 1995; Edens, Buffington e Tomicic, 2000; Marion et al., 2012; Poythress, Edens e Watkins, 2001). Le poche ricerche a favore della presunta superiorità degli psicopatici nel mentire hanno raccolto evidenze deboli o solo parzialmente confermative. Ad esempio, Book, Holden, Starzyk, Wasylkiw e Edwards (2006) hanno indotto i partecipanti della loro ricerca ad alterare (in positivo o in negativo) le risposte all'*Holden Psychological Screening Inventory* (HPSI; Holden, 1996), creando quattro gruppi distinti: *faking good efficace/inefficace* e *faking bad efficace/inefficace* (a seconda che le condotte distorsive venissero rilevate dalle scale di controllo). La psicopatia è stata valutata utilizzando la LSRP (Levenson et al., 1995). Analizzando i risultati, gli autori hanno trovato che i partecipanti del gruppo *faking good inefficace* erano caratterizzati da punteggi più bassi sulle misure di psicopatia rispetto al gruppo *faking good efficace*. Tuttavia, non è stata riscontrata un'associazione simile nel caso del *faking bad*. Utilizzando il PPI piuttosto che la LSRP come misura della psicopatia, MacNeil e Holden (2006) hanno condotto una ricerca simile e confermato i risultati di Book et al. (2006). In altri termini, i soggetti con punteggi elevati nei tratti psicopatici erano più abili nell'adottare strategie distorsive. Tali risultati positivi hanno riguardato solo alcune scale del PPI (e.g., Egocentrismo Machiavellico, Esternalizzazione della colpa) e hanno presentato una dimensione dell'effetto comunque moderata.

In conclusione, non è tuttora chiaro (o comunque i risultati emersi necessitano di ulteriori conferme) se e in che misura gli psicopatici/ASPD siano portati a mentire più spesso degli altri. Similmente, le evidenze di una loro presunta superiorità nel farlo sono poche e controverse. Ciononostante, Anderson, Sellbom, Wygant e Edens (2013) hanno ribadito la necessità di usare *self-report* muniti di scale di controllo. Nella loro ricerca sul PPIR gli autori hanno dimostrato che gli individui — se motivati — sono in grado

di elevare/ridurre significativamente i punteggi delle scale del test. In assenza di un meccanismo di verifica dei fenomeni di *over-reporting* o *under-reporting*, le interpretazioni tratte dai protocolli PPIR (e più in generale da qualunque questionario sulla psicopatia) potrebbero risultare imprecise e fuorvianti.

1.3 *Machine Learning*

L'apprendimento automatico, o *machine learning* (ML), è la branca dell'intelligenza artificiale concernente lo studio e l'applicazione di algoritmi[8] che consentono ai sistemi informatici di eseguire compiti (e.g., di classificazione) senza ricevere istruzioni specifiche, attraverso un processo graduale di apprendimento induttivo (Marsland, 2011). Mitchell ha fornito la definizione più citata di *machine learning*: "*Si dice che un programma apprende dall'esperienza E con riferimento ad alcune classi di compiti T e con misurazione della performance P, se le sue performance nel compito T, come misurato da P, migliorano con l'esperienza E*" (Mitchell, 1997, p. 2).

Dall'analisi dell'enunciato di Mitchell emergono 4 aspetti fondamentali. In primo luogo, il concetto di *esperienza* che rappresenta l'accesso alle informazioni disponibili su un certo dominio di conoscenza. In secondo luogo, il termine *compito* che si traduce nella definizione operativa degli obiettivi di apprendimento. In terzo luogo, la nozione di *performance* che consente la valutazione oggettiva dei compiti eseguiti. Infine, il concetto di *apprendimento* inteso quale dinamica iterativa che trasforma — per affinamenti successivi — la materia bruta dei dati in schemi di conoscenza. In altri termini, gli algoritmi ML: (a) analizzano gli input che ricevono in ingresso; (b) estraggono da essi le informazioni necessarie alla risoluzione dei problemi che sono chiamati a risolvere; (c) producono un risultato in uscita; (d) migliorano con l'esperienza.

Nel 1989, in un lavoro che divenne celebre per le sue implicazioni epistemologiche, Cybenko dimostrò il seguente teorema, denominato *di approssimazione universale* e qui riportato nella sua succinta formulazione matematica: sia $\varphi : \mathbb{R} \to \mathbb{R}$ una funzione continua, non costante, limitata e monotona crescente. Sia I_m un ipercubo unitario $[0, 1]^m$. Allora, dati $\epsilon > 0$ e una qualunque funzione $f \in C(I_m)$ — essendo $C(I_m)$ l'insieme delle funzioni continue su I_m — esiste un intero n e costanti reali α_i, b_i, w_{ij} dove $i = 1, \ldots, n$ e $j = 1, \ldots, m$ tale che possiamo definire

$$F(x_1, \ldots, x_m) = \sum_{i=1}^{n} \alpha_i \cdot \varphi(\sum_{j=1}^{m} w_{ij} x_j + b_i) \tag{1.1}$$

[8] Un algoritmo è un procedimento che risolve un determinato problema attraverso un numero definito di passi elementari, chiari e non ambigui. Il termina deriva dalla trascrizione latina del nome del matematico persiano *al-Khawarizm*, vissuto nel IX secolo d.C., che è considerato uno dei primi autori ad aver fatto riferimento a questo concetto.

come una realizzazione approssimata della funzione f; ovvero

$$|F(x_1, \ldots, x_m) - f(x_1, \ldots, x_m)| < \epsilon \tag{1.2}$$

per ogni x che giace nello spazio degli input.

In altri termini, ciò che Cybenko fu in grado di dimostrare è che gli algoritmi ML[9] sono in grado di approssimare qualunque fenomeno la cui natura sia riconducibile a una funzione, per quanto complessa possa essere. Appare evidente la portata di questo teorema quando si ponga mente al fatto che i ricercatori in psicologia implementano spesso modelli di relazioni causative o associative tra l'insieme delle variabili indipendenti e le variabili dipendenti oggetto di studio. Relazioni che ricordano, per l'appunto, quelle intercorrenti tra dominio e codominio di una funzione.

Tradizionalmente, gli algoritmi ML[10] vengono raggruppati in tre macro-classi, a seconda dei dati impiegati nella fase di apprendimento e della natura del *feedback* disponibile (Ayodele, 2010):

- *Supervised learning* — Gli algoritmi di questa categoria ricevono un *dataset* iniziale di *training* costituito da casi risolti, per quali è già noto l'output desiderato e che per questo motivo fungono da *exempla*. Estrapolando la conoscenza insita negli *exempla* — e dunque secondo un processo di tipo induttivo — tali algoritmi generano una funzione che viene utilizzata per prevedere l'output associato alle istanze del *dataset* di *training* e a nuove istanze prive di output. I problemi che possono essere risolti con quest'approccio sono di natura classificatoria o regressiva. Nel primo scenario, gli algoritmi stimano output limitati a un insieme discreto di classi, nel secondo gli

[9] In particolare, una rete *feed-forward* con singolo *layer* nascosto e funzione di attivazione sigmoidale (con il termine *feed-forward* si intende un insieme di nodi-neuroni interconnessi senza che le connessioni formino cicli, per cui l'output è determinato solamente dall'input corrente). Integrando Cybenko, Hanin (2017) ha dimostrato che le reti *deep learning* con larghezza n+1, funzione di attivazione ReLU e libere di espandersi in profondità possono approssimare qualunque funzione Lebesgue-misurabile nello spazio n-dimensionale degli input.

[10] Non è possibile produrre un elenco esaustivo di tutti gli algoritmi ML disponibili in letteratura. Limitandosi all'apprendimento supervisionato e alla libreria *scikit-learn* (ver. 0.21.03) impiegata nei due studi della ricerca, il ricercatore è posto di fronte a una scelta tanto ampia quanto intimidatoria: *Generalized Linear Models, Linear and Quadratic Discriminant Analysis, Kernel ridge regression, Support Vector Machines, Stochastic Gradient Descent, Nearest Neighbors, Gaussian Processes, Cross decomposition, Naive Bayes, Decision Trees, Ensemble methods, Multiclass and multilabel algorithms, Feature selection, Semi-Supervised, Isotonic regression, Probability calibration, Neural network models (supervised).*

output possono assumere un valore numerico qualsiasi all'interno di un intervallo definito.

- *Unsupervised learning* — Gli algoritmi di questa categoria ricevono e analizzano vettori di input senza output associati. Il processo di apprendimento consiste nell'individuazione di *pattern* comuni/ricorrenti e, per certi versi, ricorda le strategie euristiche della *cluster analysis*. La conoscenza che se ne ricava non è dunque il riflesso di un'*expertise* pregressa (e consolidata negli *exempla*), ma rappresenta un'acquisizione *ex novo*.

- *Reinforcement learning* — Gli algoritmi di questa categoria costruiscono sistemi (definiti anche agenti a sottolinearne la natura attiva) che interagiscono con altri sistemi, tra cui l'essere umano, in un ambiente non totalmente deterministico (nel quale i risultati conseguibili sono parzialmente dipendenti dal caso) al fine di migliorare un set di comportamenti assunti come target dell'apprendimento e ciò in risposta a segnali di ricompensa o di punizione forniti dall'ambiente medesimo.

L'implementazione di un algoritmo ML prevede l'esecuzione di un numero predefinito di fasi (Marsland, 2011): (a) raccolta ed elaborazione dei dati; (b) *feature engineering*, ovvero individuazione/creazione di variabili chiave da incorporare nel modello predittivo; (c) scelta, addestramento e validazione dell'algoritmo; (d) applicazione operativa. Il ciclo inizia con la raccolta dei dati e con le operazioni di pulizia su di essi (e.g., normalizzazione, *missing imputation*, ecc.); continua con l'individuazione e/o creazione[11] delle variabili che fungeranno da predittori; la parte centrale del ciclo riguarda la scelta dell'algoritmo da implementare, il suo addestramento e la valutazione della generalizzabilità dei suoi risultati. Infine, verificatane la corretta implementazione, il modello predittivo generato dall'algoritmo viene impiegato operativamente.

1.3.1 Apprendimento supervisionato

Nel seguito della presente trattazione, il focus attentivo sarà rivolto agli algoritmi di apprendimento supervisionato per compiti classificatori. Tali algoritmi sono stati impiegati nei due studi del capitolo 2.

[11] Ad esempio, attraverso tecniche di *dimensionality reduction*.

Vengono presentate le seguenti definizioni formali utili per una migliore comprensione dei concetti illustrati nelle sezioni successive:

- **Dominio di conoscenza**. Un insieme $\mathcal{X}$ costituito da tutte le istanze che devono essere classificate. Ciascun elemento di $\mathcal{X}$ è rappresentato da un vettore di *features* (i.e., caratteristiche) x.

- **Classi**. Un insieme $\mathcal{Y}$ costituito da k classi, $\{y_1, \ldots, y_k\}$. Ogni elemento di X è associato in modo univoco a una classe di $\mathcal{Y}$.

- **Dati di addestramento**. $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ è una sequenza finita di istanze correttamente etichettate. Esse rappresentano gli input a cui l'algoritmo ML può accedere in fase di addestramento.

- **Modello generativo dei dati**. Si presume che i dati S siano stati generati da una distribuzione di probabilità $\mathcal{D}$ che opera su $\mathcal{X}$. Per quanto riguarda le classi, si ipotizza l'esistenza di una funzione non contraddittoria [12] che associa a ogni elemento di $\mathcal{X}$ un elemento di $\mathcal{Y}$ ($f : \mathcal{X} \to \mathcal{Y}$). Occorre sottolineare che l'algoritmo ML non è a conoscenza di $\mathcal{D}$ o di f.

- **Funzione ipotesi**. La funzione f rappresenta l'obiettivo di apprendimento. In altre parole, ogni coppia in S viene generata da una distribuzione di probabilità ignota $\mathcal{D}$ seguita da una funzione f, altrettanto ignota. Compito dell'algoritmo ML è di definire una funzione h ($h : \mathcal{X} \to \mathcal{Y}$, con h appartenente a una classe finita $\mathcal{H}$) che imiti il funzionamento di f su S e per estensione su $\mathcal{X}$. Tale funzione prende il nome di *Ipotesi*. L'algoritmo ML produce h stimando/aggiornando i suoi parametri interni attraverso metodi di ottimizzazione numerica[13]; bisogna infine precisare che l'algoritmo stesso possiede proprie impostazioni; queste ultime — anch'esse ottimizzabili con tecniche numeriche — prendono il nome di iper-parametri e possono essere concepite come meta-impostazioni che orientano funzionamento dell'algoritmo medesimo.

- **Performance**: viene assunto come indicatore di performance l'errore di classificazione che equivale alla probabilità di trovare un elemento $\mathcal{X}$, in accordo alla

[12] Con il sintagma "funzione non contraddittoria" si intende una funzione che assegna lo stesso output al medesimo input, in ogni condizione, in ogni tempo.

[13] I parametri interni di h assomigliano, per analogia impropria, ai coefficienti $\hat{\beta}$ di una regressione lineare.

distribuzione $\mathcal{D}$, tale per cui $h(x)$ è diverso $f(x)$

$$L_{\mathcal{D},f}(h) \stackrel{\text{def}}{=} \mathcal{D}(\{x : h(x) \neq f(x)\}) \tag{1.3}$$

Il pedice $\mathcal{D}, f$ di $L_{\mathcal{D},f}(h)$ indica che l'errore viene misurato rispetto alla distribuzione $\mathcal{D}$ e alla funzione f. $L_{\mathcal{D},f}(h)$ ha diversi sinonimi quali *errore di generalizzazione*, *rischio*, *errore vero* di h. L'obiettivo dell'algoritmo è di trovare una funzione h che minimizzi $L_{\mathcal{D},f}(h)$;

$$h \underset{\text{h}\in\mathcal{H}}{=} \arg\min\left(L_{\mathcal{D},f}(h)\right) \tag{1.4}$$

tuttavia, l'errore vero non è quantificabile direttamente poiché la distribuzione $\mathcal{D}$ e la funzione f sono ignote. Al posto di $L_{\mathcal{D},f}(h)$, viene calcolato l'errore in cui incorre l'algoritmo nel classificare gli elementi di S

$$L_S(h) \stackrel{\text{def}}{=} \frac{|\{i \in [n] : h(x_i) \neq y_i\}|}{n} \tag{1.5}$$

dove $n = card(S)$. Tale misura viene definita *errore empirico* o *rischio empirico*. Dal momento che il *dataset* di addestramento è (auspicabilmente) una rappresentazione accurata dello spazio $\mathcal{X}$, è ragionevole considerare $L_S(h) \approx L_{\mathcal{D},f}(h)$. Il processo di apprendimento basato sulla minimizzazione di $L_S(h)$ prende il nome di *minimizzazione del rischio empirico*, abbreviato in ERM. Dal punto di vista pratico, $L_S(h)$ può assumere diverse forme numeriche, ciascuna delle quali viene definita *funzione di costo* J.

Minimizzazione del rischio empirico

Come anticipato nella sezione 1.3.1 (p. 32), gli algoritmi ML apprendono h (i.e., i suoi parametri) attraverso la minimizzazione della funzione di costo J. Intuitivamente, si tratta di ridurre lo scarto tra le previsioni compiute da h e la realtà rappresentata da S. Una strategia simile viene impiegata nel metodo dei minimi quadrati (OLS) per la stima della retta di regressione lineare $y = \sum_{j=1}^{p} X_j\beta_j$, metodo che presenta la seguente soluzione analitica (Field, 2013, p. 200)

$$\hat{\beta} = \arg\min \left|\left| \sum_{j=1}^{p} X_j\beta_j - y \right|\right|_2 = (X^T X)^{-1} X^T y \tag{1.6}$$

Sfortunatamente, nel caso degli algoritmi ML, la minimizzazione di J non prevede soluzioni esatte, per cui i parametri di h vengono calcolati attraverso strategie di ottimizzazione numerica, tra cui, ad esempio, il metodo della discesa del gradiente (Cherkassky & Mulier, 2007). Nella figura 1.1, è rappresentata visivamente l'ottimizzazione numerica dei parametri di una regressione lineare usando tale metodo su un *dataset* sintetico[14].

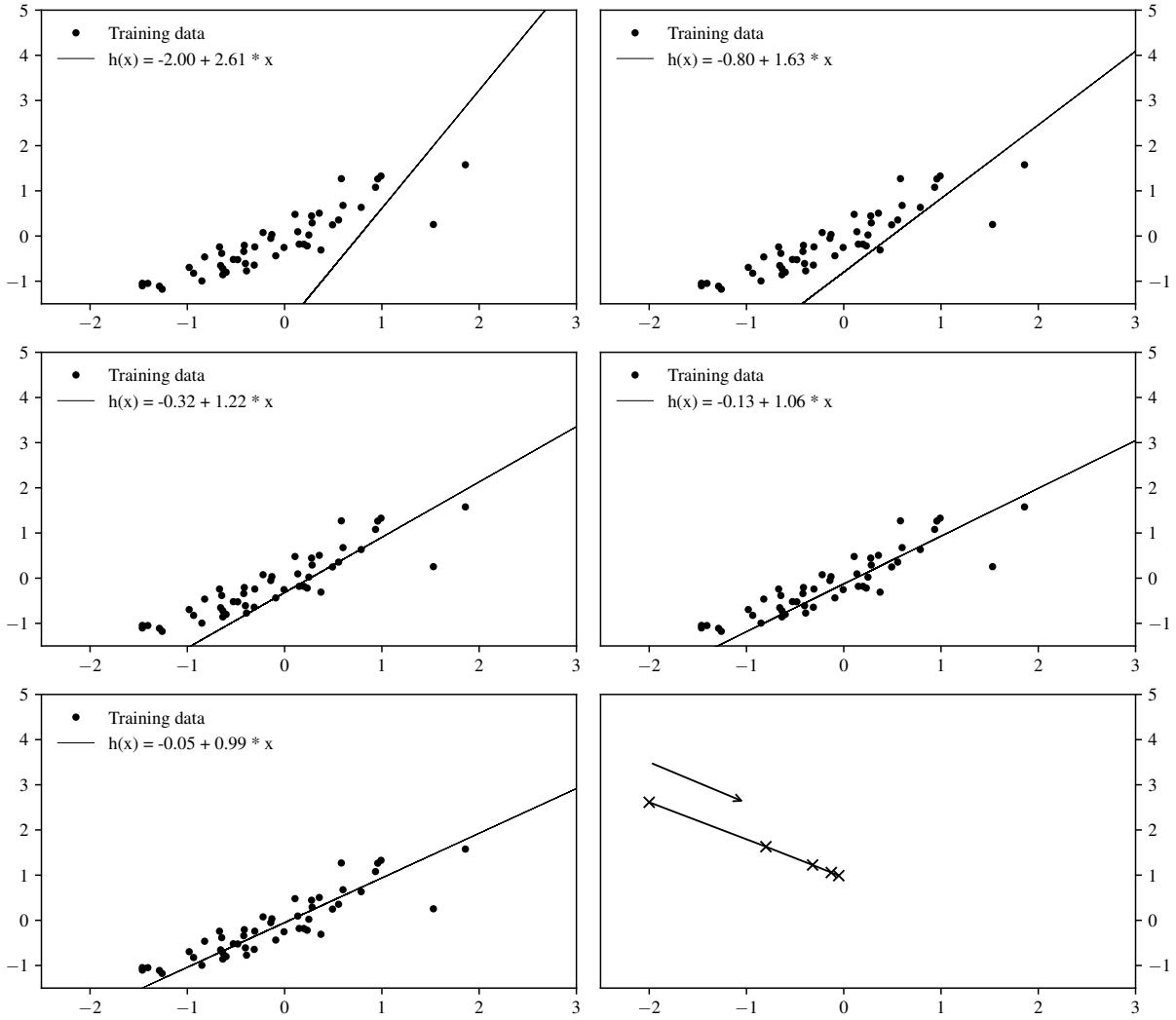

Figura 1.1
Ottimizzazione con il metodo della discesa del gradiente

Partendo dal grafico in alto a sinistra, si può notare il graduale aggiustamento della retta di regressione via via che la funzione di costo viene minimizzata. Nell'ultima figura in basso, sono rappresentate le diverse iterazioni dei parametri $\hat{\beta}$ che convergono verso la stima finale, idealmente corrispondente alla soluzione analitica (equazione 1.6).

[14] disponibile all'indirizzo *http://bit.ly/2lZrRF2*.

Generalmente, i metodi di ottimizzazione numerica terminano o con il raggiungimento del numero massimo di iterazioni consentite o quando la minimizzazione di J comporta un miglioramento inferiore a un determinato valore di *threshold* (Marsland, 2011). Va infine notato che le funzioni di costo J possono essere convesse — e cioè presentare un minimo globale — o non convesse. Le prime consentono agli algoritmi ML di conseguire una soluzione ottimale, le seconde non offrono tale garanzia per la presenza di punti critici (e.g., minimi/massimi locali, punti di sella) che — sotto certe condizioni — potrebbero impedire l'efficace minimizzazione di J (Jin, Ge, Netrapalli, Kakade & Jordan, 2017).

Complessità vs Parsimonia

La minimizzazione della funzione di costo J dipende dalla complessità di h (Domingos, 2000). Questo importante concetto viene illustrato nella figura 1.2; il grafico in alto a sinistra rappresenta un insieme di 20 punti generati da $y = x - 1.5(x^2) + 0.1(x^3)$ con l'aggiunta di una componente di errore gaussiano.

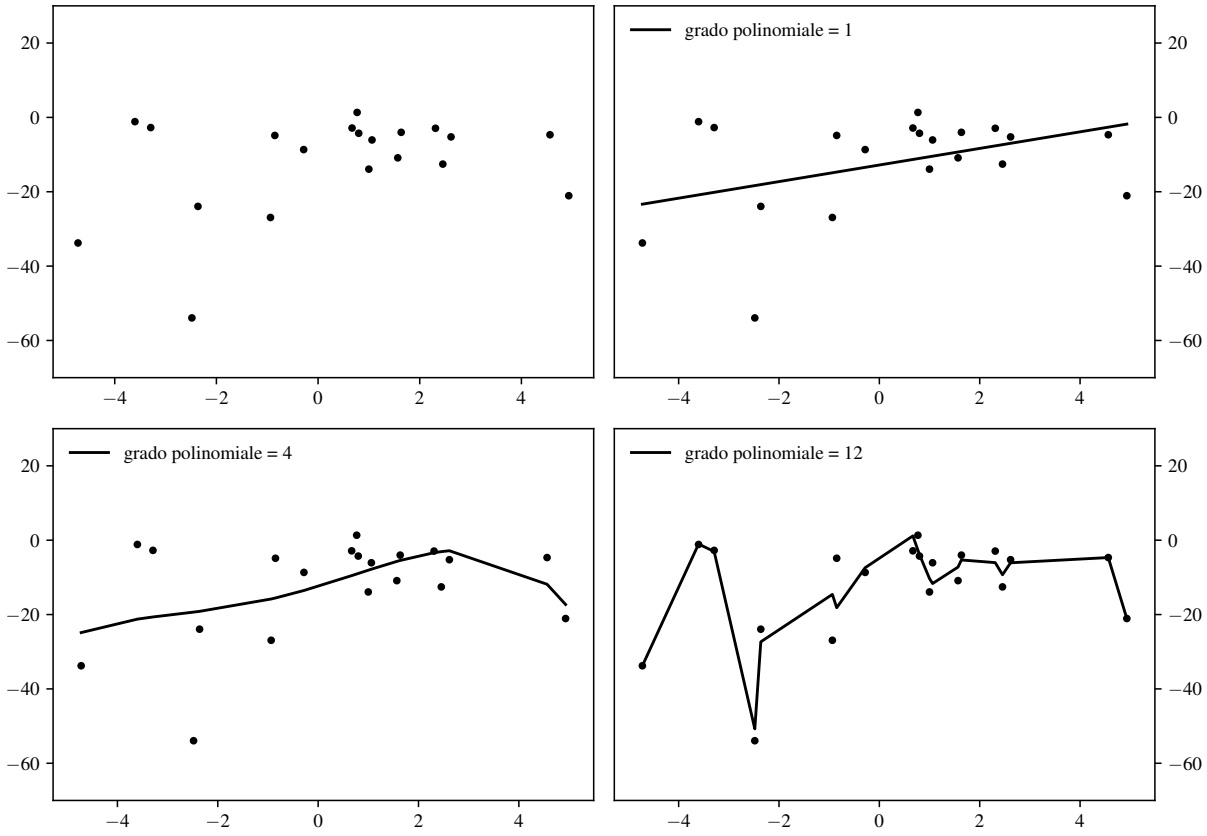

Figura 1.2
Esempio di complessità vs parsimonia

Nel riquadro in alto a destra, i dati sono stati stimati con una regressione di grado polinomiale 1, mentre nei riquadri in basso, il grado è stato portato prima a 4 poi a 12. Si può facilmente notare come l'incremento del grado polinomiale (i.e., della complessità) consenta al modello di adattarsi ai dati con maggiore precisione. In altri termini, la regressione polinomiale di grado 1 presenta un alto *bias* e cioè un basso potere predittivo (i.e., *under-fit*), mentre la regressione di grado 12 si adatta quasi perfettamente ai dati e presenta, di conseguenza, un basso valore di *bias*. Si potrebbe allora concludere che la migliore strategia per gli algoritmi ML sia quella di implementare funzioni ipotesi h molto complesse. Tuttavia, maggiore è la complessità di h, maggiore è la probabilità che essa incorpori informazioni irrilevanti presenti nei dati di addestramento[15]. In uno scenario simile, h presenterebbe una *variance* elevata (i.e., *over-fit*), poiché le sue stime — particolarmente efficaci con i dati di *training* — non riuscirebbero a generalizzare con altri *dataset* (Domingos, 2000). *Bias* e *variance* sono concetti legati a doppio filo: non è possibile migliorare la performance di h aumentandone arbitrariamente la complessità; il prezzo da pagare per la riduzione del *bias* è una difficoltà di generalizzazione e cioè un'alta variabilità (*variance*) nella performance. Una parte importante dell'ottimizzazione di un algoritmo ML è allora quella di tenere contemporaneamente sotto controllo i fenomeni di *bias* e *variance*, implementando funzioni ipotesi h col giusto grado di complessità (Alpaydin, 2009).

Misurazione della performance

In merito al tema della performance, occorre sottolineare che lo scopo di una funzione ipotesi h è di operare efficacemente oltre il *dataset* di *training*; non avrebbe senso sviluppare un modello predittivo che si adatta ai dati di addestramento ma che risulta inadeguato in tutti gli altri casi. Per questo motivo, la misurazione della performance di h viene effettuata su istanze non incluse nel campione di *training* (Dwyer, Falkai & Koutsouleris, 2018). Idealmente, se vi fossero abbastanza osservazioni, si potrebbe ricavare un set di validazione separato. Spesso però i dati non sono sufficienti. Al fine di risolvere questo problema, è stata sviluppata un tecnica nota come *k-fold cross-validation* (Friedman, Hastie & Tibshirani, 2001); tale tecnica prevede che i dati vengano suddivisi in K partizioni

[15] Vale la pena di ricordare che i dati della figura 1.2 contengono una componente di errore gaussiana che non dovrebbe essere modellizzata da h.

approssimativamente equivalenti[16]; h viene quindi addestrata su $k-1$ parti (combinate insieme), mentre la valutazione della sua performance viene effettuata sulla k-esima (separata dalle altre). La procedura viene ripetuta k volte, ricavando infine la media delle k stime delle prestazioni di h. Più in dettaglio sia $k : \{1, \ldots, N\} \longmapsto \{1, \ldots, K\}$ una funzione di indicizzazione casuale che assegna l'istanza i-esima alla partizione k-esima e sia h^{-k} la funzione ipotesi applicata alla k-esima parte (e addestrata sulle rimanenti $k-1$). Allora, la stima dell'errore di predizione di h sarà

$$CV(h) = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{1}{N_k} \sum_{i \in N_k} L(y_i, h^{-k}(x_i)) \right) \tag{1.7}$$

1.3.2 Algoritmi per problemi classificatori

Data la quantità di algoritmi ML disponibili, il primo compito del ricercatore è di scegliere quelli che possono garantire il migliore risultato rispetto ai dati che si intendono processare. Nel presente lavoro, è stata presa le decisione di impiegare le tecniche ML basate sugli alberi decisionali (descritte nelle sezioni successive) sulla base della seguente osservazione di Friedman et al. (2001):

> Un metodo *off-the-shelf* è un metodo che per essere applicato non richiede una notevole elaborazione preliminare dei dati o un'accurata messa a punto della procedura di apprendimento. Di tutti gli algoritmi ML noti, gli alberi decisionali si avvicinano di più al concetto di metodo *off-the-shelf* per il *data mining*. Sono veloci da costruire e producono modelli interpretabili (se gli alberi sono piccoli) [...] gestiscono naturalmente variabili numeriche, categoriali e i valori mancanti. Sono invarianti rispetto alle trasformazioni (strettamente monotone) dei singoli predittori. [...] sono immuni agli effetti degli *outlier*. Compiono automaticamente la selezione dei predittori più rilevanti da incorporare nel modello. Sono quindi resistenti, se non completamente immuni, all'inclusione di molte variabili irrilevanti. Queste proprietà sono il motivo per cui gli alberi decisionali rappresenatno il metodo di apprendimento più popolare per il *data mining*. (p. 352).

[16] Valori tipici di k sono 5 o 10. Il caso estremo di $K = N$ è noto come *leave-one-out cross-validation* (Friedman et al., 2001).

Prima di illustrare gli algoritmi ML (basati sugli alberi decisionali) adottati nei due studi della ricerca, occorre ricordare la definizione di problema classificatorio secondo il paradigma dell'apprendimento supervisionato. Considerato un *dataset* di *training* $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, con $x \in \mathbb{R}^m$ e y derivante da $\{y_i, \ldots, y_k\}$, l'obiettivo di un algoritmo ML è di generare una funzione ipotesi h in grado di assegnare correttamente a ogni elemento x di S la classe y che naturalmente gli appartiene.

Alberi decisionali CART

Gli alberi decisionali CART per problemi classificatori[17] sono costituiti da un insieme di decisioni binarie organizzate in una struttura gerarchica. Possono essere formalmente definiti come grafi aciclici diretti, composti da un insieme di nodi che codificano decisioni (Pauly, 2012); ciascun nodo è collegato a un elemento padre (livello superiore) ed eventualmente — se esiste un biforcazione discendente — a due elementi figli (livello inferiore). Il nodo principale da cui origina l'albero è chiamato radice, tutti gli altri sono nodi-interni oppure nodi-foglia, a seconda che abbiano sotto di sé ulteriori ramificazioni.

Le istanze x da classificare attraversano l'albero procedendo dall'altro verso il basso, seguendo un percorso determinato dalle decisioni prese in ciascun nodo, fino a quando non raggiungono un nodo-foglia (vedi figura 1.3). Ogni nodo non terminale è caratterizzato da una procedura di *splitting* il cui compito è di suddividere le istanze che processa in due partizioni mutualmente esclusive, i.e. $S = S_{left} \cup S_{right}$ e $S_{left} \cap S_{right} = \emptyset$. La procedura di *splitting* del nodo generico c è così definita

$$\text{per ogni } x \text{ in ingresso} \begin{cases} f_c(x) = x \cdot v_c \geq \tau_c \\ \text{Se } f_c(x) = 0 \text{ vai a sinistra} \\ \text{Se } f_c(x) = 1 \text{ vai a destra} \end{cases} \tag{1.8}$$

dove $dim(v_c) = dim(x)$ e $\tau_c \in \mathbb{R}$. Per quanto riguarda gli alberi CART, v_c è un vettore con un solo elemento diverso da 0 (corrispondente al predittore che viene testato dalla procedura di *splitting*).

[17] Nel presente lavoro, si farà riferimento agli alberi *Classification and Regression Tree* (CART) introdotti per la prima volta da Breiman, Friedman, Olshen e Stone (1984). Tuttavia sono stati proposti in letteratura altri tipi di alberi decisionali come, ad esempio, ID3, C4.5, CHAID, MARS, ecc. (Gupta, Rawat, Jain, Arora & Dhami, 2017).

Tutte le istanze che ricadono nei singoli nodi-foglia ricevono un valore γ (i.e., previsione della classe) calcolato nel seguente modo. In un nodo generico c, rappresentante una regione R_c con N_c istanze, sia

$$\hat{p}_{ck} = \frac{1}{N_c} \sum_{x_i \in R_c} I(y_i = k) \tag{1.9}$$

la proporzione di istanze appartenenti alla classe k. Allora, γ sarà

$$\arg\max_k \hat{p}_{ck} \tag{1.10}$$

Ossia, le istanze che arrivano al nodo c ricevono un valore γ pari alla classe predominante.

In sintesi, un albero decisionale suddivide lo spazio dei predittori in c regioni disgiunte R_c, ciascuna rappresentata da un nodo-foglia a cui viene associato un valore di classe γ. La sua struttura può essere espressa nel seguente modo

$$T(x; \Theta) = \sum_{c=1}^{C} \gamma_c I(x \in R_c) \tag{1.11}$$

con parametri $\Theta = \{R_c, \gamma_c\}_{c=1}^{C}$.

In tal senso, un albero è simile a un diagramma di flusso in cui ciascun nodo interno rappresenta un test sul valore di un predittore, ciascun nodo-foglia un'etichetta di classe, mentre i diversi percorsi — dalla radice alle foglie — le regole di classificazione.

In aggiunta a quanto già detto, ogni nodo è caratterizzato da un certo grado di purezza/impurezza legato alla presenza di istanze appartenenti a classi diverse (i.e., i nodi puri contengono istanze appartenenti a una singola classe, non così i nodi impuri che, dunque, possono essere ulteriormente partizionati). Durante la fase di addestramento, la creazione di nodi supplementari risponde alla necessità di ridurre il grado di impurità della struttura dell'albero.

Nell'algoritmo CART, l'indice di purezza usato è il coefficiente di Gini[18] (Friedman et al., 2001). Riprendendo l'equazione 1.9, il computo di tale coefficiente per un nodo

[18] Esistono altri coefficienti di impurezza, come ad esempio, *Cross-entropy*, o *Misclassification error* (Friedman et al., 2001, p. 309).

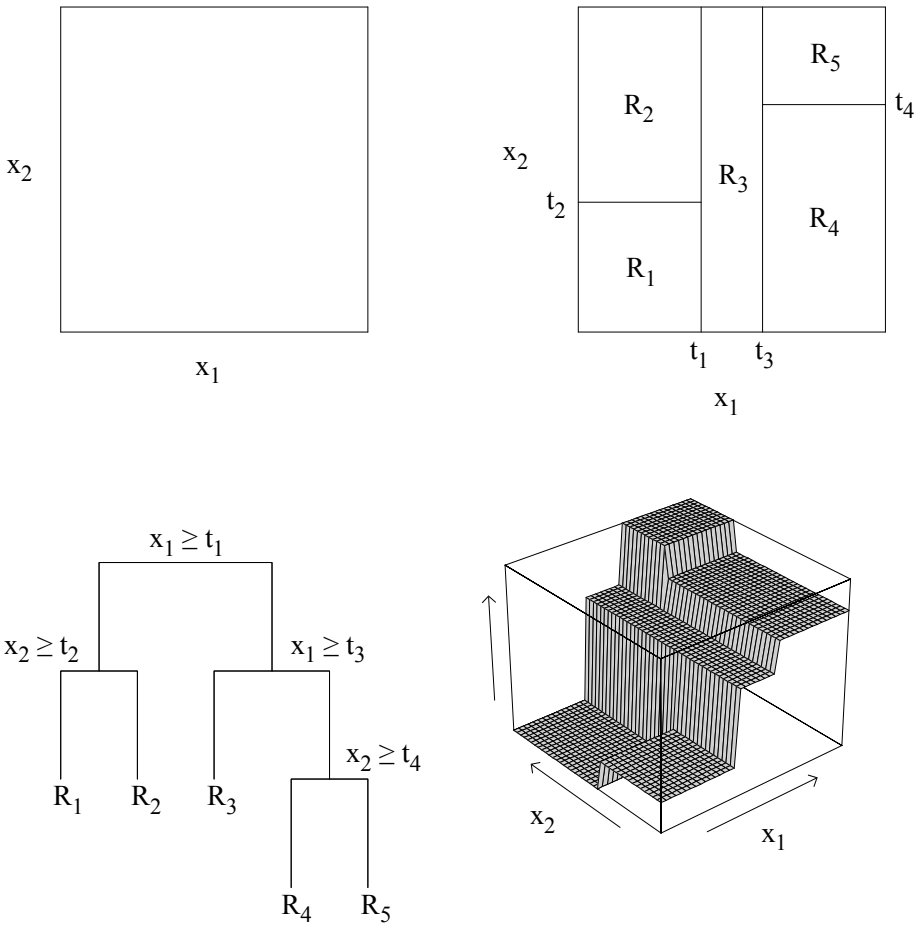

Figura 1.3
Alberi decisionali: partizionamento ricorsivo

generico c[19] viene effettuato nel seguente modo

$$I_c = \sum_{k=1}^{K} \hat{p}_{ck} \sum_{k' \neq k} \hat{p}_{ck'} = \sum_{k=1}^{K} \hat{p}_{ck}(1 - \hat{p}_{ck}) = 1 - \sum_{k=1}^{K} \hat{p}_{ck}^2 \tag{1.12}$$

Nel processo iterativo, l'algoritmo sceglierà il predittore (variabile di *split*) e il valore di *split* che garantiscono la maggiore riduzione dell'impurezza $\triangle I$

$$\triangle I = I_c - \underbrace{(P_{left} I_{left} + (1 - P_{left}) I_{right})}_{\text{a}} \tag{1.13}$$

essendo I_c l'impurità del nodo generico c e (a) le impurità pesate dei nodi figli. Breiman

[19] Dall'analisi dell'equazione 1.12 si evince che il coefficiente di Gini misura la probabilità di assegnare erroneamente a un'istanza di classe k (con probabilità di occorrenza pari a $\hat{p}_{ck}$) una classe $k' \neq k$ tra quelle disponibili nel nodo generico c.

et al. (1984) hanno dimostrato che la selezione dello *split* che massimizza il decremento di impurità definito in 1.13 è equivalente alla selezione dello *split* che minimizza l'impurità totale dell'albero. Ciò significa che il criterio di ottimizzazione locale di un nodo equivale alla ottimizzazione globale dell'albero.

È possibile definire tre criteri di arresto nella creazione di nodi supplementari (Pauly, 2012): (a) profondità massima dell'albero; (b) popolazione minima per foglia (c) variazione minima dell'impurità. Il primo criterio considera la struttura gerarchica dell'albero; una volta raggiunta una certa profondità, la divisione iterativa si interrompe. Il secondo criterio si basa sul numero di istanze contenute in un nodo; se esso è inferiore a una determinata soglia, la divisione non procede oltre. Per quanto riguarda l'ultimo criterio, se la riduzione dell'impurità è al di sotto di un valore di *threshold* specificato, il processo iterativo si conclude.

Quanto dovrebbe crescere un albero decisionale? Strutture molto ramificate possono presentare fenomeni di *over-fitting*, mentre strutture poco ramificate possono esibire il problema contrario e cioè non cogliere adeguatamente il segnale presente nei dati. Quindi, la dimensione di un albero è il parametro che regola la complessità del modello predittivo sotteso e per tale motivo dovrebbe essere ottimizzato. La strategia preferita (Friedman et al., 2001) consiste nel far crescere un albero T_0 arrestando il processo solo quando i nodi-foglia contengono un numero molto basso di istanze (e.g., 3). L'albero così ottenuto — presumibilmente *over-fitted* — viene potato usando un metodo definito *cost complexity pruning* che fissa il seguente criterio di complessità

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m I_m + \alpha|T| \tag{1.14}$$

con m riferito alla regione R_m, $N_m = \#\{x_i \in R_m\}$, $|T|$ uguale al numero di nodi-terminali e I_m indicante l'impurità del nodo m.

L'idea è di trovare la sotto-struttura $T_\alpha \subseteq T_0$ in grado di minimizzare $C_\alpha(T)$. Il parametro di ottimizzazione $\alpha \geq 0$ regola il compromesso tra le dimensioni dell'albero e la sua adattabilità ai dati. Valori elevati di α danno luogo ad alberi poco sviluppati e viceversa per valori di α più piccoli. Per ogni α si può dimostrare che esiste un'unica sotto-struttura T_α che minimizza $C_\alpha(T)$. Per trovarla, vengono via via trasformati in foglie i nodi-interni di T_0, partendo da quelli che producono il più piccolo incremento della

quantità $\sum_m N_m I_m$; il processo di *pruning* continua fino all'ottenimento di un albero a nodo singolo (radice). Questa procedura genera una sequenza finita di sotto-strutture $T \subseteq T_0$; Breiman et al. (1984) hanno dimostrato che tale sequenza contiene sicuramente T_α.

Tecniche di insieme basate sugli alberi decisionali

Nel suo trattato *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix* del 1785, il marchese di Condorcet propose e dimostrò un importante teorema sulla probabilità che un gruppo di individui giunga, per votazione, a una scelta corretta. Le premesse del teorema descrivono una situazione in cui una giuria è chiamata a deliberare su una decisione che prevede un esito binario e in cui ogni elettore ha una probabilità indipendente p di votare correttamente. Il teorema si chiede quanti individui cooptare per aumentare la probabilità che la giuria voti la scelta più idonea. Il risultato dipende dal fatto che p sia maggiore o minore di $1/2$: nel primo caso, l'aggiunta di più individui aumenta la probabilità che la decisione della maggioranza sia corretta (e si avvicina asintoticamente a 1 all'aumentare del numero degli elettori); nel secondo caso, la probabilità diminuisce. Dunque, affinché una votazione collettiva sia più efficace della decisione del singolo, devono sussistere due condizioni: (1) ogni membro del gruppo deve avere una probabilità $p = 1/2 + \epsilon$ con $\epsilon > 0$ di individuare la scelta corretta e (2) i membri devono essere indipendenti, ovvero le loro decisioni non correlate.

Le tecniche di insieme sfruttano il teorema di Condorcet per migliorare la performance degli algoritmi ML, combinando più modelli di base approssimativamente corretti[20] (cond. 1) in un unico comitato le cui stime risultano qualitativamente superiori alle stime dei modelli presi singolarmente. Un aspetto centrale di queste tecniche è l'impiego di strategie volte a garantire la massima de-correlazione possibile tra i modelli che confluiscono nel comitato (cond. 2).

[20] Definiti anche *weak learner* a segnalare la non necessità che essi abbiano performance elevate.

Foreste casuali

Una foresta casuale $\mathcal{F}$ è un insieme di B alberi decisionali $\mathcal{F} = \{F_1, \ldots, F_B\}$. Come dimostrato da Breiman (2001), sostituendo un singolo albero con un insieme di alberi de-correlati si ottiene una migliore performance del modello finale in termini di generalizzazione dei risultati. L'idea essenziale delle foreste casuali è di aggregare alberi rumorosi (i.e., con elevati valori di *variance*) ma approssimativamente *unbiased* al fine di ridurre la varianza totale del comitato. Gli alberi decisionali sono i candidati ideali per questo genere di approccio, poiché possono catturare interazioni complesse nei dati e dunque presentare bassi valori di *bias*. Al fine di garantire la de-correlazione degli alberi è necessario iniettare un certo quantitativo di casualità nella procedura di addestramento. Breiman et al. (1984) hanno introdotto il concetto di *bagging* che deriva dal fusione dei termini *bootstrap* e *aggregating*. Dato un *dataset* di *training* $\mathcal{S} = \{(x_n, y_n)\}_{n=1}^{N}$, una replica bootstrap è un sottoinsieme S_b di S in cui gli elementi sono stati campionati casualmente usando una distribuzione uniforme (con o senza sostituzione). Ogni albero F_b della foresta casuale viene quindi addestrato usando un set di dati *bootstrap* $\mathcal{S}_b$ leggermente diverso. Questa procedura conduce a migliori prestazioni della foresta perché ne viene diminuita la varianza senza aumentarne il *bias*. Ciò significa che, mentre le previsioni di un singolo albero sono molto sensibili al rumore nel *dataset* di *training*, la previsione media di molti alberi non lo è (purché questi ultimi non siano correlati). Un'altra strategia di de-correlazione interviene nel processo di crescita degli alberi attraverso la selezione casuale dei predittori nella procedura di *node splitting*.

In sintesi, i diversi alberi di una foresta casuale sono esposti — in fase di addestramento — a dati leggermente diversi e a set parzialmente diversi di predittori. Ciò, come è già stato ribadito, riduce la varianza del modello aggregato, salvaguardandone la capacità di produrre previsioni *unbiased*. Di seguito, viene illustrato in forma schematica quanto appena esposto (Friedman et al., 2001, p. 588).

---

1. Per $b = 1$ fino a B
   (a) Genera un campione *bootstrap* Z tratto dal *dataset* di addestramento
   (b) Sviluppa un albero decisionale T_b a partire dal campione *bootstrap*, ripetendo i seguenti passi per ogni nodo dell'albero finché ciascun nodo non abbia raggiunto

la dimensione minima prestabilita

i. Seleziona casualmente m predittori tra tutti i predittori disponibili

ii. Scegli la migliore combinazione variabile/punto di split

iii. Dividi il nodo corrente in due nodi discendenti usando la variabile/punto di split prescelti

2. Restituisci l'albero completamente sviluppato $\{T_b\}_{b=1}^{B}$

Una volta costituitasi la foresta di alberi decisionali essa potrà essere usata per predire la classe di un'istanza generica x. Sia $\hat{C}_b(x)$ la classe predetta dal b-esimo albero. Allora

$$\hat{C}_{rf}^{B}(x) = \text{voto a maggioranza}\,\{\hat{C}_b(x)\}_{b=1}^{B} \tag{1.15}$$

XGBoost

L'algoritmo ML XGBoost (*Extreme Gradient Boosting*) rientra tra le tecniche di insieme che sfruttano una strategia denominata *boosting* (Chen & Guestrin, 2016). In senso molto generale, si può affermare che lo scopo del *boosting* sia quello di applicare sequenzialmente un algoritmo per la generazione di *weak learner* ciascuno dei quali prova a ridurre gli errori previsionali dei propri predecessori. In questo modo, via via che le iterazioni di addestramento si succedono, le istanze più difficili da classificare assumono un'importanza crescente e i classificatori candidati a entrare nel comitato sono spinti a concentrarsi su di esse.

Nella sua implementazione più tipica, XGBoost genera un comitato previsionale costituito da una sequenza di alberi CART. Matematicamente

$$\hat{y}_i = \phi(x) = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F} \tag{1.16}$$

dove $dim(x) = m$, K è il numero totale dei *weak learner* e $\mathcal{F}$ è lo spazio che include tutti gli alberi CART così definiti

$$f(x) = w_{q(x)}, q : \mathbb{R}^m \to T, w \in \mathbb{R}^T \tag{1.17}$$

In altri termini, ogni albero $f(x)$ viene rappresentato da una struttura fissa q con w a rappresentare il vettore dei valori predetti associati ai nodi-foglia T; $\phi(x)$ si pone l'obiettivo di minimizzare la seguente funzione

$$\mathcal{L}(\phi) = \underbrace{\sum_{i}^{n} L(y_i, \hat{y}_i)}_{\text{a}} + \underbrace{\sum_{k=1}^{K} \Omega(f_k)}_{\text{b}} \tag{1.18}$$

$\mathcal{L}(\phi)$ è costituita da un termine (a) che rappresenta la nota funzione di costo degli algoritmi ML (i.e., lo scarto tra previsione e realtà) e un termine (b) — spiegato più oltre — che mira a ridurre l'*over-fit* del *weak learner*, penalizzando i modelli eccessivamente complessi.

Come anticipato più sopra, l'adesione dei *weak learner* al comitato avviene in modo sequenziale

$$\begin{aligned}
\hat{y}_i^{(0)} &= 0 \\
\hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\
\hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\
&\dots \\
\hat{y}_i^{(k)} &= \sum_{k=1}^{K} f_k(x_i) = \hat{y}_i^{(k-1)} + f_k(x_i)
\end{aligned} \tag{1.19}$$

Naturalmente, gli alberi iterativamente aggiunti devono minimizzare la funzione obiettivo definita in 1.18 e qui adattata alla t-esima iterazione.

$$\begin{aligned}
\mathcal{L}(\phi)^{(t)} &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) \\
&= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)
\end{aligned} \tag{1.20}$$

Utilizzando l'errore quadratico medio (MSE) come funzione di costo, $\mathcal{L}(\phi)^{(t)}$ diventa

$$\begin{aligned}
\mathcal{L}(\phi)^{(t)} &= \sum_{i=1}^{n} (y_i - (\hat{y}_i^{(t-1)} + f_t(x_i)))^2 + \sum_{i=1}^{t} \Omega(f_i) \\
&= \sum_{i=1}^{n} \underbrace{[2 \underbrace{(\hat{y}_i^{(t-1)} - y_i)}_{a.1} f_t(x_i)}_{a} + \underbrace{f_t(x_i)^2}_{b}] + \Omega(f_t)
\end{aligned} \tag{1.21}$$

Così definita, $\mathcal{L}(\phi)^{(t)}$ ha un un termine di primo grado (a) — spesso chiamato residuo per via di a.1 — e un termine quadratico (b). La funzione obiettivo (differenziabile) $\mathcal{L}(\phi)^{(t)}$ viene approssimata mediante un polinomio di Taylor del secondo ordine a

$$\widetilde{\mathcal{L}}(\phi)^{(t)} = \sum_{i=1}^{n}[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)] + \Omega(f_t) \tag{1.22}$$

con g_i e h_i definite come

$$\begin{aligned} g_i &= \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \\ h_i &= \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \end{aligned} \tag{1.23}$$

Dopo la rimozione di tutte le costanti, $\widetilde{\mathcal{L}}(\phi)^{(t)}$ diventa

$$\widetilde{\mathcal{L}}(\phi)^{(t)} = \sum_{i=1}^{n}[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)] + \Omega(f_t) \tag{1.24}$$

e rappresenta l'obiettivo di ottimizzazione all'iterazione t-esima.

Riferendosi alle equazioni 1.17 e 1.18, Chen e Guestrin (2016) hanno definito il termine di regolarizzazione $\Omega(f)$ nel seguente modo

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2, \text{ con} \gamma \text{ e } \lambda \text{ costanti} \tag{1.25}$$

Utilizzando $\Omega(f)$ per come definito in 1.25, l'equazione 1.24 diventa

$$\widetilde{\mathcal{L}}(\phi)^{(t)} = \sum_{i=1}^{n}[g_i w_{q(x_i)} + \frac{1}{2}h_i w_{q(x_i)}^2] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{1.26}$$

Partendo dal presupposto che tutte le istanze che giungono a un nodo-foglia j-esimo ricevono lo stesso valore (in termini di predizione) w_j, l'equazione 1.26 può essere riscritta nel seguente modo

$$\widetilde{\mathcal{L}}(\phi)^{(t)} = \sum_{j=1}^{T}[(\sum_{i \in I_j} g_i)w_j + \frac{1}{2}(\sum_{i \in I_j} h_i + \lambda)w_j^2] + \gamma T \tag{1.27}$$

dove $I_j = \{i \,|\, q(x_i) = j\}$ è il set degli indici delle istanze assegnate al nodo-foglia j-esimo.

La precedente equazione può essere ulteriormente semplificata in

$$\widetilde{\mathcal{L}}(\phi)^{(t)} = \sum_{j=1}^{T}[G_j w_j + \frac{1}{2}(H_j + \lambda)w_j^2] + \gamma T \tag{1.28}$$

definendo $G_j = \sum_{\in I_j} g_i$ e $H_j = \sum_{i \in I_j} h_i$

Considerato che il principio di ottimizzazione numerica di XGBoost è la discesa del gradiente e che $G_j w_j + \frac{1}{2}(H_j + \lambda)w_j^2$ ha una forma quadratica, i migliori w_j per una data struttura $q(x)$ sono

$$w_{j\,best}^{(*)} = -\frac{G_j}{H_j + \lambda} \tag{1.29}$$

Di conseguenza, la migliore riduzione della funzione obiettivo è

$$\widetilde{\mathcal{L}}(\phi)_{best}^{(*)} = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T \tag{1.30}$$

Quest'ultima equazione può essere intesa come misura della bontà strutturale di $q(x)$. A somiglianza di quanto espresso nell'equazione 1.13, lo sviluppo degli alberi prevede che i nodi vengano generati basandosi sul seguente coefficiente di guadagno strutturale

$$Gain = \frac{1}{2}\left[\underbrace{\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda}}_{1} - \underbrace{\frac{(G_L + G_R)^2}{H_L + H_R + \lambda}}_{2}\right] - \gamma \tag{1.31}$$

Dove (1) rappresenta i coefficienti dei nodi discendenti, (2) è il coefficiente del nodo-foglia prima dello *split* e γ è il termine di regolarizzazione. Vale la pena di osservare che lo *split* non viene portato a termine se il guadagno strutturale è inferiore a γ; ciò ricorda il concetto di *pruning* espresso nella sezione 1.3.2 (p. 39).

Capitolo 2

Studi empirici

2.1 Obiettivi generali

Come è stato già riferito nel primo capitolo (sezione 1.1.2, p. 8) la strategia più diffusa per la misurazione del *faking* è rappresentata dalle scale di controllo o scale *Lie* (Paulhus, 1991). La premessa concettuale alla base di questa strategia di detezione è che il *faking* sia un processo di natura lineare o quasi-lineare. Nel rispondere positivamente agli item di controllo, gli individui forniscono ripetute evidenze delle loro condotte distorsive. Quando la somma di queste evidenze (i.e., degli item) supera la soglia di attenzione prestabilita (*cutoff* normativo), gli individui sono classificati come *faker*.

In contrasto con quanto appena esposto, Kuncel e Tellegen (2001) hanno proposto di riconsiderare il problema del *faking* collocandolo nel contesto dei singoli item. Gli autori hanno presentato prove che la desiderabilità/indesiderabilità sociale spesso non è correlata linearmente alle distribuzioni dei tratti di personalità; alcuni item vengono percepiti come massimamente desiderabili a livelli medi (i.e., curva di desiderabilità a "u" inversa), mentre altri a livelli superiori ma comunque non estremi (i.e., curva di desiderabilità ascendente con plateau ed eventuale fase discendente). Per via della loro natura aggregativa, i punteggi di scala tendono a oscurare questi fenomeni item-specifici e ciò a causa dell'effetto livellante della somma algebrica.[21] In una recente indagine

[21] Per illustrare il concetto con un esempio banale, è facile osservare come $1+9$ e $5+5$ assommino ugualmente a 10, ma è altrettanto facile constatare come tali addizioni veicolino significati profondamente diversi, la prima di un forte contrasto, la seconda di perfetta medietà.

empirica, Kuncel e Borneman (2007) hanno portato conferme supplementari. Attraverso un disegno sperimentale di manipolazione diretta del *faking*, gli autori hanno sottoposto a un campione di 206 studenti una lista di descrittori aggettivali di personalità con scala likert a 9 punti, chiedendo a una parte dei partecipanti di descriversi onestamente e a un'altra parte di fornire le proprie valutazioni adottando una strategia di auto-promozione del sé (immaginando di trovarsi in uno scenario fittizio di selezione del personale). Nella figura 2.1 sono presentate — a titolo esemplificativo — le distribuzioni di 3 descrittori (*Careful*, *Imperturbable*, *Unenvious*) per condizione di somministrazione (*Honest - Ho* vs *Faking - Fk*).

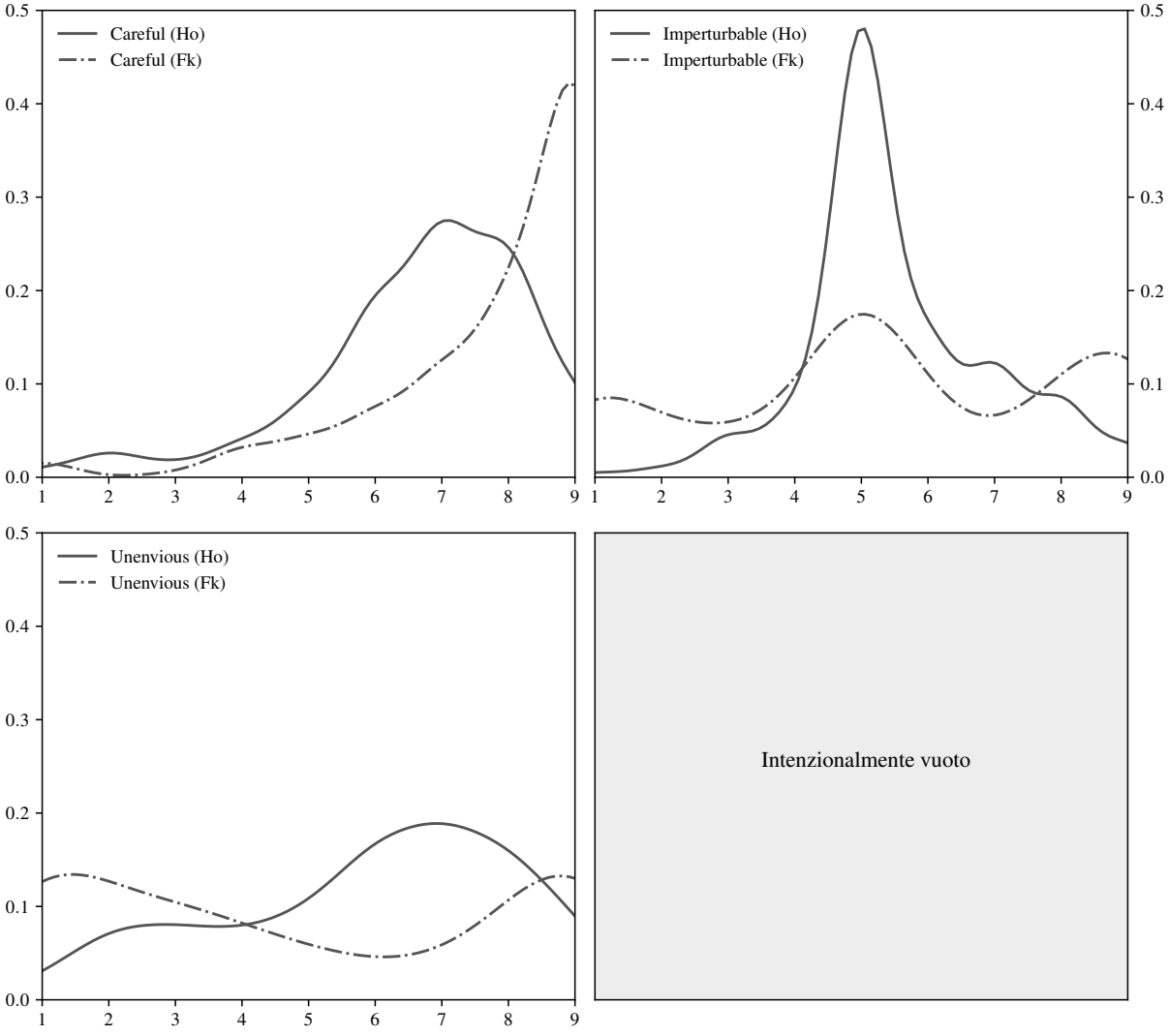

Figura 2.1
Esempi di *idiosyncratic item responses*

L'ispezione visiva dei *kernel density plot* dà un'idea immediata di quanto siano di-

verse le modalità di risposta degli individui che decidono di presentarsi sotto un luce particolarmente favorevole rispetto a quanti si rappresentano onestamente. Kuncel e Borneman (2007) hanno introdotto il concetto di *idiosyncratic item responses* per descrivere le perturbazioni dei comportamenti di risposta che sono indotte dal *faking* e implicitamente hanno suggerito di indagare tali perturbazioni onde intercettare le strategie manipolatorie dei soggetti che mentono.

La presente ricerca raccoglie due diversi studi condotti al fine di sviluppare una tecnica innovativa di classificazione del *faking* basata sull'analisi dei *pattern* di risposta agli item, mediante l'uso di algoritmi di *machine learning* (ML). L'impostazione generale di entrambi i lavori (vedi figura 2.2) ha previsto le seguenti fasi:

1. Manipolazione diretta del *faking* con lo scopo di ottenere un *dataset* di profili di personalità sia attendibili che distorti (*honest* vs *fake*).
2. Implementazione degli algoritmi ML:
   (a) addestramento di due o più classificatori ML in grado di rilevare la presenza di distorsioni o nei punteggi di scala o nei *pattern* di risposta;
   (b) scelta del classificatore ML più efficace in termini di performance.
3. Comparazione del miglior classificatore ML (2.a) con il classificatore di riferimento (CBC) basato sui punteggi delle scale *Lie* e relativi *cutoff* normativi.

Nel primo studio è stato impiegato il BFQ2 (Caprara et al., 2007) con un campione di studenti universitari, nel secondo l'inventario *Psychopathic Personality Inventory - Revised* (PPIR; Lilienfeld e Widows, 2005) con un campione di studenti universitari e uno di pazienti psichiatrici. Relativamente al secondo lavoro, la decisione di somministrare un questionario sulla psicopatia è scaturita dalla constatazione — suffragata dalla letteratura[22] — che gli individui con tratti di personalità psicopatici possono mettere in atto condotte distorsive e manipolatorie e dunque gli strumenti di misurazione sviluppati per questa classe di pazienti dovrebbero essere muniti di tecniche efficaci di rilevamento del *faking*.

[22] Consulta sezione 1.2.2, p. 26.

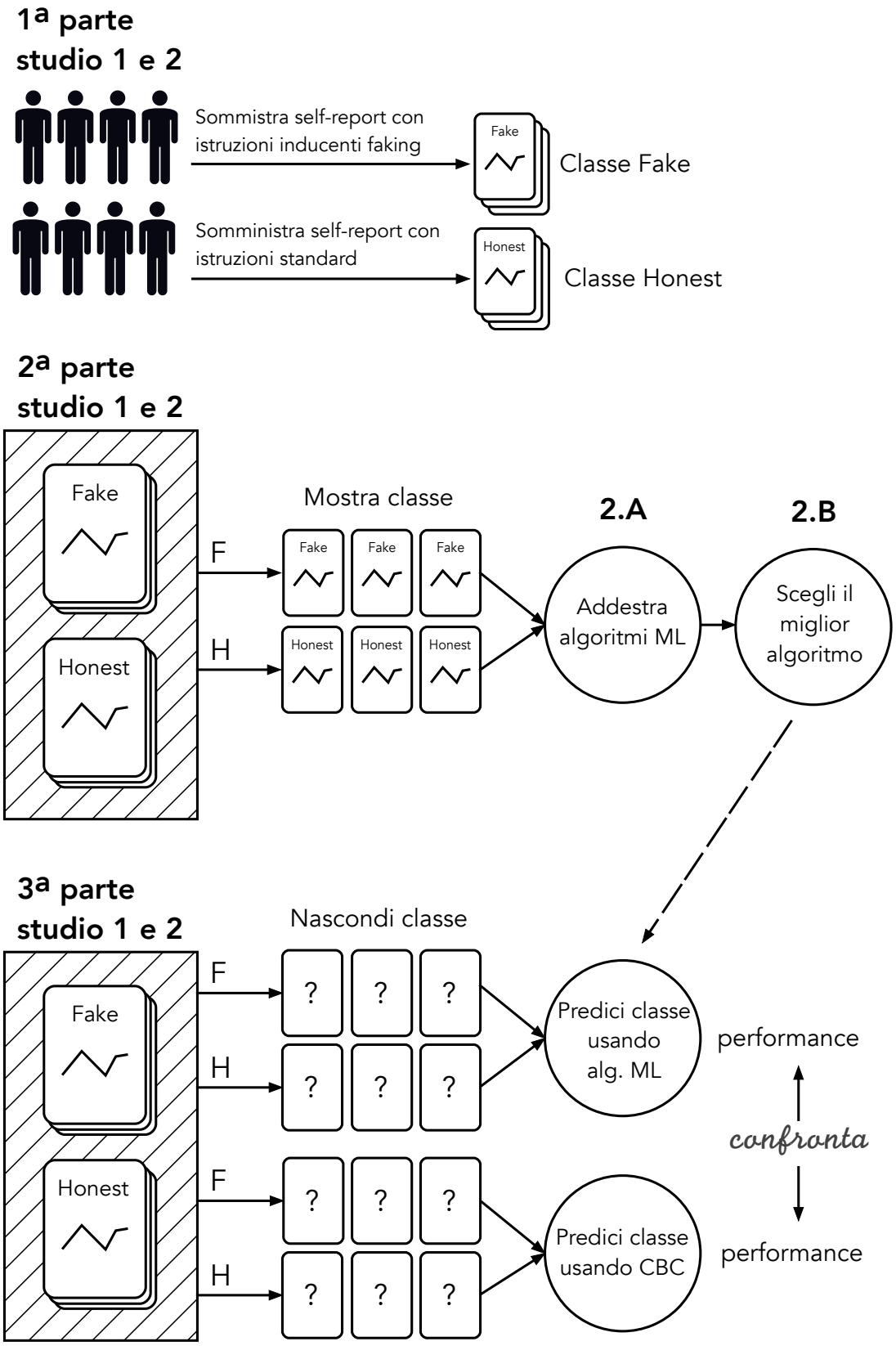

Figura 2.2
Studi 1 e 2 - Impostazione del disegno sperimentale

2.2 Studio BFQ2

2.2.1 Obiettivi

Come già anticipato nella sezione 2.1 (p. 49), il presente studio è stato realizzato al fine di sviluppare una tecnica di *machine learning* (ML) per la rilevazione dei fenomeni di *faking good* in un questionario *self-report* di personalità, tecnica basata sull'analisi delle matrici di risposta agli item e proposta in alternativa alle scale *Lie*.

2.2.2 Metodi

Partecipanti

I partecipanti sono stati 548 studenti universitari di psicologia dell'Università Sapienza di Roma e Gabriele D'Annunzio di Chieti-Pescara (età: $M = 22.11$, $SD = 3.45$) suddivisi in 3 gruppi di somministrazione *Uni-Ho*, *Uni-Fg(t)* e *Uni-Fg(f)*[23]. Oltre al campione dei partecipanti, è stato raccolto un *dataset* di profili reali[24] costituito da 4000 casi (età: $M = 32.27$, $DS = 8.95$), 2000 dei quali classificati come *High Faker* (Scala *Lie*: $M = 3.30$, $DS = 0.39$) e il resto come *Low Faker* (Scala *Lie*: $M = 2.29$, $DS = 0.33$). Le tabelle 2.1 e 2.2 illustrano schematicamente la composizione dei due campioni denominati rispettivamente *dataset Uni* e *dataset Olr*.

Tabella 2.1: BFQ2 - Composizione *dataset Uni*

Pop-Cond	F	M	All
Uni-Ho	220	74	294
Uni-Fg(t)	92	45	137
Uni-Fg(f)	95	22	117
All	407	141	548

[a] Uni-Ho = rispondenti onesti, Uni-Fg(t) = rispondenti *faker* (insegnanti), Uni-Fg(f) = rispondenti *faker* (vigili del fuoco)

[23] Consulta la sezione relativa alle procedure per ulteriori chiarimenti in merito ai gruppi di somministrazione.

[24] Ricavato da un repositorio privato della società *Giunti Psychometrics*.

Tabella 2.2: BFQ2 - Composizione *dataset Olr*

Pop-Cond	F	M	All
Olr-Ho	816	1184	2000
Olr-Fg	830	1170	2000
All	1646	2354	4000

[a] Olr-Ho = *Low faker*, Olr-Fg = *High faker*

Strumenti

Sono stati fissati due requisiti principali per il questionario di personalità *self-report* da impiegare nello studio: (1) doveva essere uno strumento di valutazione comunemente usato dai professionisti italiani delle risorse umane e (2) doveva avere una scala *Lie* per il rilevamento del *faking*. Alla luce di queste considerazioni, è stato scelto il *Big Five Questionnaire 2* (BFQ2; Caprara et al., 2007) un *self-report* di 134 item con likert a 5 punti e scale di dominio mutuate dal modello dei *Big Five*: E - Energia, A - Amicalità, C - Coscienziosità, S - Stabilità Emotiva, M - Apertura Mentale.

Il BFQ2 ha una scala *Lie* per rilevare le strategie di auto-miglioramento/ distorsione. È composta da 14 item che veicolano l'immagine irrealistica di un individuo capace, competente, brillante, coraggioso, particolarmente affabile, rispettoso degli altri e attento alle norme sociali (Caprara et al., 2007). La scala *Lie* ha un punteggio di *cutoff* normativo genere-specifico di 55 T, valore al di sopra del quale i comportamenti di *faking* sono considerati significativi.

Procedura

La prima parte della procedura ha riguardato l'addestramento di 6 classificatori ML con il *dataset Olr* (tabella 2.3). Tre di essi sono stato adattati ai punteggi delle scale di personalità (E, A, C, S, M) e i rimanenti tre alle matrici di risposta (tabella 2.3); ciò per separare l'effetto dei due set di predittori (scale vs item) dall'effetto dei diversi algoritmi ML implementati.

Nella seconda parte della procedura, è stato confrontato il miglior classificatore ML con il classificatore di riferimento (CBC), basato sulla scala *Lie* e cutoff normativi specifici per genere. Entrambi i classificatori sono stati applicati a 1000 repliche *bootstrap* del *dataset*

Tabella 2.3: BFQ2 - Elenco classificatori ML implementati (*dataset Olr*)

Predittori	Algoritmo ML	Nome
Scale di dominio	Regressione Logistica	LR-S
	Foresta Casuale	RF-S
	XGBoost	XGB-S
Matrice degli item	Regressione Logistica	LR-I
	Foresta Casuale	RF-I
	XGBoost	XGB-I

Uni che è stato raccolto somministrando il BFQ2 in 2 diverse condizioni: (1) il gruppo dei rispondenti onesti (*Uni-Ho*) che ha completato il questionario dopo aver ricevuto istruzioni standard; (2) il gruppo *fake good* (*Uni-Fg*) ai cui membri è stato chiesto di completare l'inventario immaginando di candidarsi per un posto di lavoro o come insegnante di scuola superiore *Uni-Fg(t)* o come vigile del fuoco *Uni-Fg(f)* e di rispondere in modo da superare il percorso selettivo[25].

Analisi

Le analisi statistiche sono state effettuate i) con STATA (ver. 14.2); ii) con Python (ver. 3.7.3) e le seguenti librerie: Scipy (ver. 1.3.0), Numpy (ver. 1.16.4), Pandas (ver. 0.24.2), Matplotlib (ver. 3.1.0), Scikit-learn (ver. 0.21.2), XGBoost (ver. 0.90).

2.2.3 Risultati

I risultati sono stati suddivisi in 3 sezioni: (1) addestramento degli algoritmi ML con il *dataset Olr* e scelta del miglior classificatore; (2) statistiche descrittive e *manipulation check* del *dataset* ricavato dal campione di ricerca (3) comparazione del miglior classificatore ML con il classificatore di riferimento basato sulla scala *Lie* (CBC).

[25] La scelta di proporre due profili professionali diversi (e, per certi versi, antitetici) è stata dettata dalla necessità di rendere gli algoritmi ML più resistenti alle "menzogne" professione-specifiche.

Addestramento degli algoritmi ML e scelta del miglior classificatore

Come già detto, sono stati addestrati diversi classificatori ML con il *dataset Olr*, adattandoli ai punteggi delle scale di personalità BFQ2 o ai *pattern* di risposta agli item. È importante sottolineare che le fasi di addestramento (*training*), messa a punto (*hyper-parameters tuning*) e validazione (*testing*) dei modelli ML sono state condotte con dati indipendenti; derogare a questa precauzione avrebbe comportato la sovrastima delle prestazioni reali dei classificatori a causa di un fenomeno che nella letteratura ML prende il nome di *data leakage* (Luo et al., 2016). Al fine di evitare simili problemi, è stata adottata una procedura di convalida incrociata su più livelli (Cawley & Talbot, 2010). Più precisamente, è stato effettuato (a) un ciclo interno di *5-fold cross validation* per l'ottimizzazione dei parametri degli algoritmi ML tramite *random grid search* con F_1 come metrica di riferimento (b) un ciclo esterno di *10-fold cross validation* per la stima *unbiased* della performance dei classificatori implementati. Per ciascuno di essi, la tabella 2.4 riporta gli indici di prestazione più comuni (Sokolova, Japkowicz & Szpakowicz, 2006).

Tabella 2.4: BFQ2 - Performance classif. ML implementati (*dataset Olr*)

Predittori	Algoritmo	Nome	Acc	Prec	Rec	F_1	AUC
Punt. scala	Regr. Logistica	LR-S	0.66	0.66	0.67	0.67	0.72
	Foresta Casuale	RF-S	0.62	0.59	0.66	0.65	0.67
	XGBoost	XGB-S	0.65	0.65	0.64	0.65	0.70
Matr. item	Regr. Logistica	LR-I	0.76	0.75	0.76	0.76	0.83
	Foresta Casuale	RF-I	0.74	0.76	0.72	0.74	0.82
	XGBoost	XGB-I	0.76	0.76	0.77	0.77	0.84

[a] Acc=Accuracy, Prec=Precision, Rec=Recall, F_1=F_1 score, AUC=Area under AUC curve

Dall'analisi della tabella, Si può notare come gli algoritmi addestrati sulle matrici di risposta agli item abbiano garantito performance costantemente superiori rispetto a quelli addestrati sui punteggi di scala. Il miglior classificatore si è rivelato essere XGB-I, anche se il secondo in ordine di accuratezza, LR-I, ha esibito differenze pressoché insignificanti; ciononostante, nel campo del ML, è consuetudine eleggere il *best model* anche in presenza di minimi incrementi prestazionali (Alpaydin, 2009).

Per ricavare una prima impressione della meccanica interna di XGB-I, sono stati considerati i suoi predittori elencati in ordine di "guadagno" (i.e., del contributo fornito al

miglioramento del modello). Dei primi 10 predittori maggiormente influenti, il 50% è risultato appartenere alla dimensione della Stabilità Emotiva (S), mentre il resto si è distribuito in modo non uniforme tra Amicalità (20%), Coscienziosità (10%), Energia (10%) e Apertura (10%).

Statistiche descrittive e *manipulation check* del *dataset Uni*

L'obiettivo dell'ultima fase dell'analisi è stato quello di confrontare XGB-I con il classificatore di riferimento CBC, basato sulla scala *Lie*. Per fare ciò, il BFQ2 è stato somministrato ai partecipanti dello studio fornendo loro diversi set di istruzioni[26], ottenendo così il *dataset Uni*. La tabella 2.5 mostra le statistiche descrittive delle scale BFQ2 nelle tre condizioni di somministrazione, mentre la figura 2.3 mostra i profili medi associati a ciascuna condizione. Entrambi i gruppi *Uni-Fg* hanno ottenuto punteggi più alti rispetto ai partecipanti *Uni-Ho*, sebbene il sottogruppo *Uni-Fg(f) vigili del fuoco* abbia prodotto profili meno distorti rispetto al sottogruppo *Uni-Fg(t) insegnanti*. Un'analisi MANOVA ha consentito di accertare la significatività degli incrementi dei punteggi di scala nei diversi gruppi di somministrazione ($F_{(2,545)} = 34.44$, $p < 0.001$; Traccia di Pillai $= 0.55$).

A seguire, sono state condotte delle ANOVA di follow-up con computo degli effetti e dei contrasti (tabella 2.6). Il test omnibus F è risultato significativo per tutte le scale, mentre la dimensione dell'effetto è stata[27]: (a) grande per le scale Stabilità Emotiva, Apertura Mentale, Energia e *Lie* e (b) media per le rimanenti scale. L'analisi dei contrasti ha confermato le differenze pronunciate tra gli intervistati onesti e le due condizioni di *faking*. Inoltre, sono emerse differenze significative tra il gruppo *Uni-Fg(t) insegnanti* e il gruppo *Uni-Fg(f) vigili del fuoco* su tutte le dimensioni di personalità tranne la scala *Lie*.

Oltre alle precedenti indagini, si è proceduto a valutare in che misura le distribuzioni dei punteggi delle scale BFQ2 si sovrapponessero nei gruppi di somministrazione (i.e., *Uni-Ho* vs *Uni-Fg(t)* e *Uni-Fg(f)* combinati o presi singolarmente). La figura 2.4 fornisce un indicazione visiva della questione. Le scale hanno mostrato vari livelli di sovrapposizione, tutti comunque cospicui. È evidente che qualsiasi tentativo di separare i gruppi con i

[26] Consulta la sezione relativa alle procedure per ulteriori chiarimenti in merito ai gruppi di somministrazione.

[27] Consulta (Field, 2013) per l'interpretazione della dimensione dell'effetto nelle 3 fasce piccolo/medio/grande.

Tabella 2.5: BFQ2 - Descrittive punteggi scale (*dataset Uni*)

Pop-Cond		E	A	C	S	M	L
	mean	3.18	4.04	3.75	2.75	3.78	2.63
	std	0.55	0.38	0.48	0.71	0.48	0.53
	min	1.71	2.42	2.46	1.00	2.42	1.14
Uni-Ho	25%	2.79	3.79	3.42	2.25	3.47	2.29
	50%	3.25	4.06	3.75	2.75	3.83	2.57
	75%	3.57	4.29	4.04	3.24	4.11	3.00
	max	5.00	4.88	4.96	4.46	4.79	4.29
	mean	3.67	4.34	4.14	3.95	4.26	3.34
	std	0.40	0.36	0.36	0.41	0.38	0.53
	min	2.54	2.08	3.17	3.08	3.04	2.07
Uni-Fg(t)	25%	3.42	4.12	3.92	3.67	4.04	3.00
	50%	3.67	4.38	4.12	4.00	4.25	3.36
	75%	3.92	4.58	4.46	4.21	4.54	3.71
	max	4.83	4.92	4.83	4.88	4.96	4.64
	mean	3.37	4.22	4.01	3.68	3.94	3.29
	std	0.47	0.40	0.40	0.50	0.42	0.54
	min	1.96	2.42	3.00	3.00	2.58	1.64
Uni-Fg(f)	25%	3.12	4.00	3.75	3.25	3.62	2.93
	50%	3.42	4.25	4.04	3.67	3.96	3.29
	75%	3.67	4.50	4.29	4.04	4.25	3.64
	max	4.46	4.96	4.79	4.96	4.92	4.50

[a] E = Energia, A = Amicalità, C = Coscienziosità, S = Stabilità Emotiva, M = Apertura Mentale, L = Scala *Lie*

Tabella 2.6: BFQ2 - ANOVA, effetti e contrasti punteggi scale (*dataset Uni*)

			Contrasti	
Scale BFQ	$F_{(2,545)}$	ω^2	C1	C2
E Energia	45.18 ***	0.14	−0.34 ***	0.30 ***
A Amicalità	32.25 ***	0.10	−0.24 ***	0.13 **
C Coscienziosità	43.27 ***	0.13	−0.33 ***	0.14 **
S Stabilità Emotiva	219.81 ***	0.44	−1.06 ***	0.26 ***
M Apertura	53.54 ***	0.16	−0.31 ***	0.32 ***
L Lie	114.58 ***	0.29	−0.68 ***	0.06

[a] *** p < 0.001, ** p < 0.01, * p < 0.05
[b] C1 = *Uni-Ho* vs *Uni-Fg(all)*, C2 = *Uni-Fg(t)* vs *Uni-Fg(f)*

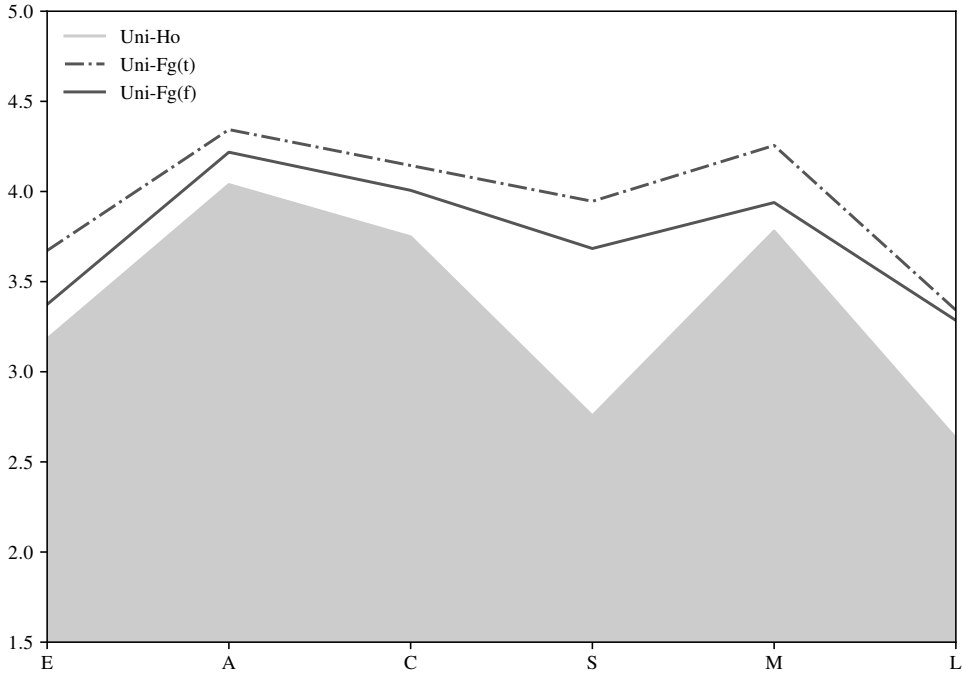

Figura 2.3
BFQ2 - Valori medi scale

punteggi di scala (inclusa la scala *Lie*) avrebbe prodotto un numero rilevante di falsi positivi/negativi[28] e ciò nonostante le differenze significative nei punteggi di scala tra le diverse condizioni di somministrazione.

Comparazione del classificatore ML con CBC

Nell'ultima parte della analisi, XGB-I e CBC sono stati applicati a 1000 repliche *bootstrap* del *dataset Uni* al fine di separare, classificandoli, i rispondenti onesti dai *faker*. Va precisato che il modello CBC, basato sulla scala *Lie*, aveva un *cutoff* genere-specifico di 55T, valore al di sopra del quale i comportamenti distorsivi sono considerati significativi e spaziano da livelli moderati a marcati (Caprara et al., 2007). I risultati hanno fornito le seguenti importanti indicazioni: (1) sia CBC che XGB-I hanno ottenuto performance migliori rispetto al modello nullo (i.e., che prevede la classe più frequente). Nel caso del *dataset Uni*, tale modello avrebbe avuto un'accuratezza del 54%, valore inferiore a CBC (72%) e XGB-I (82%); (2) XGB-I ha sistematicamente superato CBC rispetto a qualunque indice di performance considerato (tabella 2.7).

[28] Da qui la necessità di sviluppare un metodo alternativo di rilevazione del *faking*.

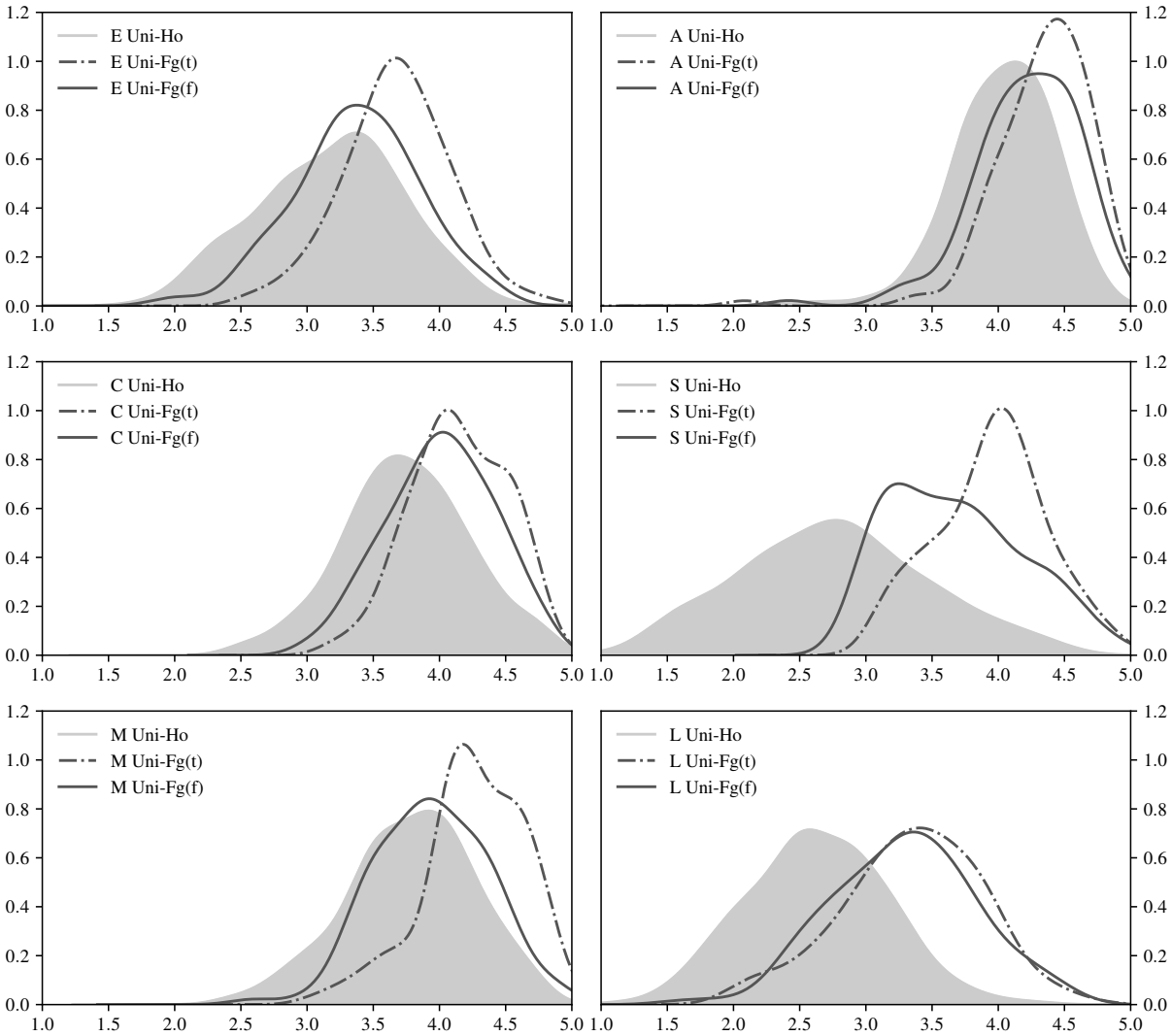

Figura 2.4
BFQ2 - Distribuzioni scale x condizione

Tabella 2.7: BFQ2 - Performance classificatori CBC, XGB-I (*dataset Uni*)

Repliche bootstrap	Classificatore	Acc	Prec	Rec	F_1	AUC
Uni-Ho vs *Uni-F(all)*	CBC	0.72	0.72	0.65	0.68	0.72
	XGB-I	0.82	0.82	0.79	0.81	0.90
Uni-Ho vs *Uni-F(t)*	CBC	0.68	0.75	0.66	0.70	0.69
	XGB-I	0.83	0.86	0.83	0.85	0.91
Uni-Ho vs *Uni-F(f)*	CBC	0.75	0.68	0.63	0.66	0.73
	XGB-I	0.82	0.77	0.76	0.76	0.88

[a] Acc=Accuracy, Prec=Precision, Rec=Recall, F_1=F_1 score, AUC=Area under ROC curve

Più formalmente, sono stati confrontati gli indici dei due classificatori eseguendo una serie di ANOVA. XGB-I ha ottenuto prestazioni migliori rispetto al CBC con dimensioni dell'effetto grandi (tabella 2.8).

Tabella 2.8: BFQ2 - ANOVA, Perf. classificatori CBC, XGB-I (*dataset Uni*)

Indici di Performance	$F_{(1,1999)}$	ω^2
Accuracy	17 592.26 ***	0.90
Precision	6778.66 ***	0.77
Recall	14 718.33 ***	0.88
F_1	16 687.82 ***	0.89
AUC	62 021.46 ***	0.97

[a] *** p < 0.001, ** p < 0.01, * p < 0.05

Al fine di ottenere informazioni dettagliate sugli errori predittivi di entrambi i modelli, sono stati confrontati i profili medi dei casi misclassificati da XGB-I e CBC con il profilo medio della classe di appartenenza. Questo confronto è stato effettuato mediante ispezione visiva (figura 2.5) e calcolando la distanza di Mahalanobis come misura di prossimità[29] (tabella 2.9).

Tabella 2.9: BFQ2 - Distanza di Mahalanobis casi misclassificati (*dataset Uni*)

	MD da classe reale		
	CBC		XGB-I
Uni-Ho misclassificati come *faker*	0.72	→	1.38
Uni-Fg(all) misclassificati come onesti	0.64	→	1.02

[a] MD = Distanza di Mahalanobis
[b] La direzione della freccia indica valori migliori

I risultati hanno mostrato che gli individui classificati erroneamente da XGB-I avevano un profilo di personalità più distante da quello della classe di appartenenza, rispetto al profilo dei casi erroneamente classificati da CBC. Dei 97 Falsi Positivi/Negativi segnalati da XGB-I, 68 casi (rappresentanti circa il 70% del totale) sono stati misclassificati anche da CBC.

[29] La distanza di Mahalanobis (o *generalized squared interpoint distance*) è definita come una misura di dissimilarità tra due vettori aleatori $\vec{x}$ e $\vec{y}$ con stessa funzione di densità di probabilità e con matrice di covarianza S: $D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}$ (De Maesschalck, Jouan-Rimbaud & Massart, 2000).

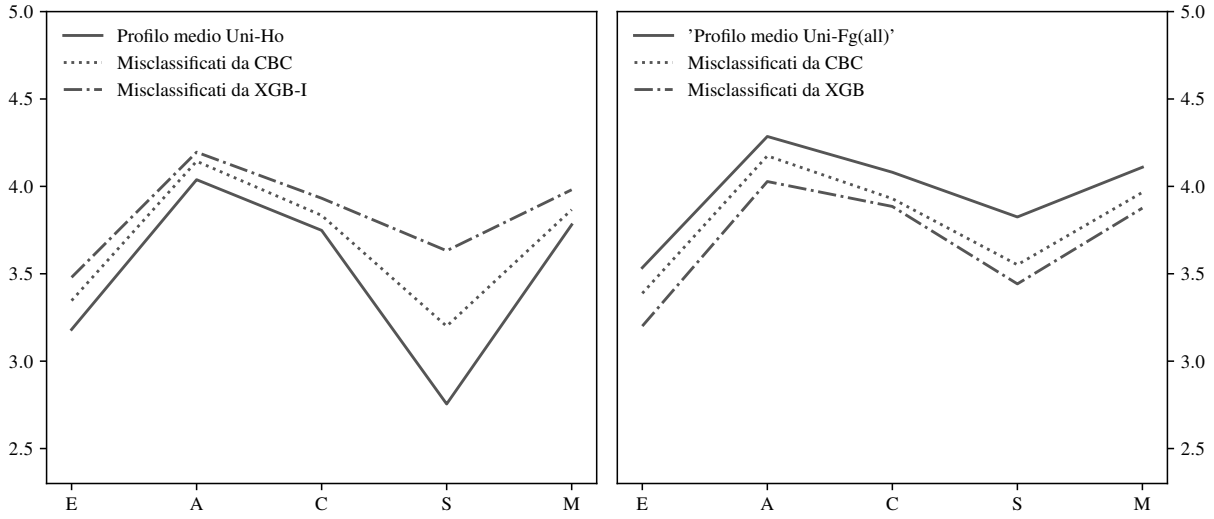

Figura 2.5
BFQ2 - Profili medi casi misclassificati (*dataset Uni*)

2.2.4 Discussione

Il presente studio ha esplorato la possibilità di impiegare gli algoritmi ML per esaminare le matrici di risposta degli item di un questionario *self-report* di personalità (BFQ2; Caprara et al., 2007) allo scopo di rilevare la presenza di *faking*. L'idea è derivata dai lavori di Kuncel e Borneman (2007) e Kuncel e Tellegen (2009), i quali hanno suggerito di considerare i cosiddetti *idiosyncratic item patterns*[30] come *marker* numerici dei comportamenti di risposta distorsivi.

Per questo motivo, sono stati implementati 3 algoritmi ML — una regressione logistica (LR), una foresta casuale (RF) e una macchina XGBoost (XGB) — per generare un totale di 6 classificatori ML. In particolare, 3 classificatori (LR-S, RF-S, XGB-S) sono stati addestrati sui punteggi di scala e altri 3 (LR-I, RF-I, XGB-I) sui *pattern* di risposta di 4000 profili reali BFQ2 (*dataset Olr*), metà dei quali etichettati come *high faker* e l'altra metà come *low faker*. I risultati hanno messo in luce la superiorità dei classificatori ML basati sui *pattern* di risposta agli item. Intuitivamente, le sottili relazioni intercorrenti tra le risposte contengono informazioni sul *faking* che non sono sfruttate dalle strategie di detezione basate esclusivamente su punteggi di scala.

Tra i classificatori ML implementati, XGB-I si è rivelato il migliore. L'ispezione dell'elenco dei predittori maggiormente influenti ha rivelato che il gruppo più numeroso era

[30] Consulta la sezione 2.1 (p. 49) per ulteriori approfondimenti.

costituito da item appartenenti alla Stabilità Emotiva (50%), seguiti da item afferenti all'area della Amicalità (20%), Coscienziosità (10%), Energia (10%) e Apertura (10%). Il secondo classificatore più efficace è stato LR-I. Le differenze tra LR-I e XGB-I si sono rivelate trascurabili. Questi risultati suggeriscono che il problema di separare i rispondenti onesti dai cosiddetti *faker* può essere affrontato usando un approccio lineare (cioè LR-I) con una perdita minima nella precisione delle previsioni, sebbene gli algoritmi basati sugli alberi decisionali (e.g., XGB-I) siano più capaci di rilevare qualsiasi non linearità presente nelle risposte agli item (Friedman et al., 2001).

Poiché in genere gli psicologi professionisti valutano il *faking* usando le scale di validità degli strumenti, sono state confrontate le prestazioni di XGB-I con CBC, un classificatore basato sulla scala *Lie* del BFQ2. Per effettuare tale comparazione sono stati raccolti nuovi profili di personalità da tre gruppi di studenti universitari (*dataset Uni*) le cui risposte sono state manipolate fornendo diversi set di istruzioni prima della somministrazione del questionario[31]. In linea con la letteratura (e.g., Birkeland, Manson, Kisamore, Brannick e Smith, 2006), i gruppi *Uni-Fg* hanno prodotto elevazioni significative nei profili — con dimensioni dell'effetto medio-grandi — rispetto ai partecipanti nella condizione *Uni-Ho*. Ciò nonostante, i gruppi sono risultati sovrapposti su quasi tutte le scale, suggerendo che poteva essere difficile separarli usando un punteggio di *cutoff* a livello di scala. XGB-I e CBC sono stati applicati a 1000 repliche *bootstrap* del *dataset Uni*. I risultati hanno mostrato che XGB-I era costantemente più preciso di CBC nell'individuare la presenza di *faking*. Quando le prestazioni sono state valutate in termini di previsioni errate, XGB-I si è rivelato, ancora un volta, migliore di CBC. Gli intervistati onesti e i *faker* classificati erroneamente da XGB-I avevano configurazioni di punteggi più dissimili dal profilo prototipico del gruppo al quale appartenevano rispetto agli intervistati classificati erroneamente da CBC. Quest'evidenza suggerisce che il tipo di errori di classificazione commessi da XGB-I sono stati più sottili, e forse più scusabili, di quelli di CBC.

[31] Istruzioni standard vs istruzioni manipolate per indurre comportamenti distorsivi. Consulta la sezione delle procedure per approfondimenti

2.3 Studio PPIR

2.3.1 Obiettivi

Come già anticipato nella sezione 2.1 (p. 49), il presente studio è stato realizzato al fine di sviluppare una tecnica di *machine learning* (ML) per la rilevazione dei fenomeni di *faking good & bad* in un questionario *self-report* di valutazione della psicopatia, tecnica basata sull'analisi delle matrici di risposta agli item e proposta in alternativa alle scale *Lie*.

2.3.2 Metodi

Partecipanti

Il campione dei partecipanti (tabella 2.10) è stato ottenuto aggregando 3 differenti banche dati — due delle quali provenienti da ricerche internazionali condotte in Australia e in Olanda — consistente in profili PPIR di studenti universitari (*Gen-Ho*, *Gen-Fg*, *Gen-Fb*; $N = 1363$, età $M = 24.12$, $DS = 9.17$) e di pazienti psichiatrici[32] (Clin-Ho, $N = 132$, età $M = 38.59$, $DS = 9.14$). Limitatamente alla componente italiana, i dati sono stati raccolti all'Università Sapienza di Roma e all'Università Gabriele D'Annunzio di Chieti-Pescara.

A causa dell'esiguità numerica totale e differentemente dallo studio 1 (sezione 2.2, p. 53) non è stato possibile generare *dataset* indipendenti per l'addestramento dei classificatori ML e per la comparazione finale con il classificatore di riferimento CBC. Le analisi sono state condotte con il campione aggregato, adottando però le necessarie precauzioni onde contrastare i fenomeni di *over-fitting* e *data leakage* (Smialowski, Frishman & Kramer, 2009).

[32] Il sotto-campione *Clin-Ho* è stato raccolto con dati provenienti da pazienti di due diverse strutture sanitarie; un centro psichiatrico e una clinica per il trattamento delle tossicodipendenze (van Dongen, Drislane, Nijman, Soe-Agnie & van Marle, 2017). La maggior parte dei pazienti del centro psichiatrico aveva una diagnosi primaria di schizofrenia e/o altri disturbi psicotici (35,3%) o ASPD (20,7%), mentre tutti i pazienti della clinica avevano una diagnosi di abuso/dipendenza da sostanze.

Tabella 2.10: PPIR - Composizione *dataset* partecipanti

Pop-Cond	F	M	All
Gen-Ho (*honest*)	240	245	485
Gen-Fg (*fake good*)	293	159	452
Gen-Fb (*fake bad*)	289	137	426
Clin-Ho (*honest*)	21	111	132

Strumenti

Per la valutazione dei tratti psicopatici è stato scelto lo *Psychopathic Personality Inventory - Revised* (PPIR; Lilienfeld e Widows, 2005)[33], un questionario *self-report* di 154 item con 8 scale di contenuto (ME Egocentrismo machiavellico, RN Anticonformismo ribelle, BE Esternalizzazione della colpa, CN Mancanza di pianificazione, SOI Influenza sociale, F Mancanza di paura, STI Immunità allo stress, C Freddezza emotiva), tre scale fattoriali (FD Dominanza priva di paura, SCI Impulsività auto-centrata e C Freddezza emotiva) e il punteggio totale (Total). Una *likert* a quattro valori consente ai rispondenti di esprimere le proprie valutazioni in merito al contenuto degli item. Il PPIR è inoltre munito di tre scale di controllo (IR Risposte inconsistenti, VR Risposte virtuose, DR Risposte devianti) per la valutazione della coerenza e dell'accuratezza dei profili.

Procedura

La prima parte della procedura ha riguardato l'addestramento di 12 classificatori ML, 6 legati alla detezione del *faking good* (tabella 2.11) e i rimanenti 6 concepiti per la rilevazione del *faking bad* (tabella 2.12). In entrambi gli scenari, il disegno sperimentale ha previsto l'induzione diretta dei comportamenti distorsivi attraverso la somministrazione del PPIR con istruzioni manipolate. In particolare, i sotto-gruppi *Gen-Ho* e *Clin-Ho* hanno ricevuto le consegne standard previste dal manuale, mentre i partecipanti dei sottogruppi *Gen-Fg* e *Gen-Fb* hanno completato il questionario dopo essere stati sollecitati a rispondere in modo da promuovere un'immagine di sé rispettivamente positiva e negativa.

Per ogni scenario di simulazione, 3 classificatori ML sono stato adattati ai punteggi delle scale di contenuto (ME, RN, BE, CN, SOI, F, STI, C) e tre alle matrici di risposta

[33] Consulta la sezione 1.2.2 (p. 24) per un approfondimento sullo strumento.

agli item; ciò per separare l'effetto dei due set di predittori (scale vs item) dall'effetto dei diversi algoritmi ML utilizzati.

Nella seconda parte della procedura, sono state confrontate le migliori implementazioni ML dei due scenari con i classificatori di riferimento: CBC-FG per il *faking good*, basato sulla scala VR e CBC-FB per il *faking bad*, basato sulla scala DR.

Tabella 2.11: PPIR - Elenco classificatori ML implementati (*faking good*)

Predittori	Algoritmo ML	Nome
Scale di contenuto	Regressione Logistica	LR-FG-S
	Foresta Casuale	RF-FG-S
	XGBoost	XGB-FG-S
Matrice degli item	Regressione Logistica	LR-FG-I
	Foresta Casuale	RF-FG-I
	XGBoost	XGB-FG-I

Tabella 2.12: PPIR - Elenco classificatori ML implementati (*faking bad*)

Predittori	Algoritmo ML	Nome
Scale di contenuto	Regressione Logistica	LR-FB-S
	Foresta Casuale	RF-FB-S
	XGBoost	XGB-FB-S
Matrice degli item	Regressione Logistica	LR-FB-I
	Foresta Casuale	RF-FB-I
	XGBoost	XGB-FB-I

Analisi

Le analisi statistiche sono state effettuate i) con STATA (ver. 14.2); ii) con Python (ver. 3.7.3) e le seguenti librerie: Scipy (ver. 1.3.0), Numpy (ver. 1.16.4), Pandas (ver. 0.24.2), Matplotlib (ver. 3.1.0), Scikit-learn (ver. 0.21.2), XGBoost (ver. 0.90).

2.3.3 Risultati

I risultati sono stati suddivisi in 3 sezioni: (1) statistiche descrittive e *manipulation check* del *dataset* ricavato dal campione di ricerca; (2) addestramento degli algoritmi ML e scelta del miglior classificatore per lo scenario *faking good* e *faking bad*; (3) comparazione dei migliori classificatori ML con i classificatori di riferimento basati sulle scale di controllo del PPIR.

Statistiche descrittive e *manipulation check*

La tabella 2.13 mostra le statistiche descrittive delle scale PPIR nelle quattro condizioni di somministrazione (*Gen-Ho*, *Gen-Fg*, *Gen-Fb*, *Clin-Ho*), mentre la figura 2.6 mostra i profili medi associati a ciascuna condizione. Il gruppo clinico e i gruppi di *faker* hanno ottenuto punteggi diversi rispetto ai partecipanti *Gen-Ho*. Un'analisi MANOVA ha consentito di accertare la significatività di tali differenze ($F_{(3,1491)} = 48.02$, $p < 0.001$; Traccia di Pillai = 0.67)

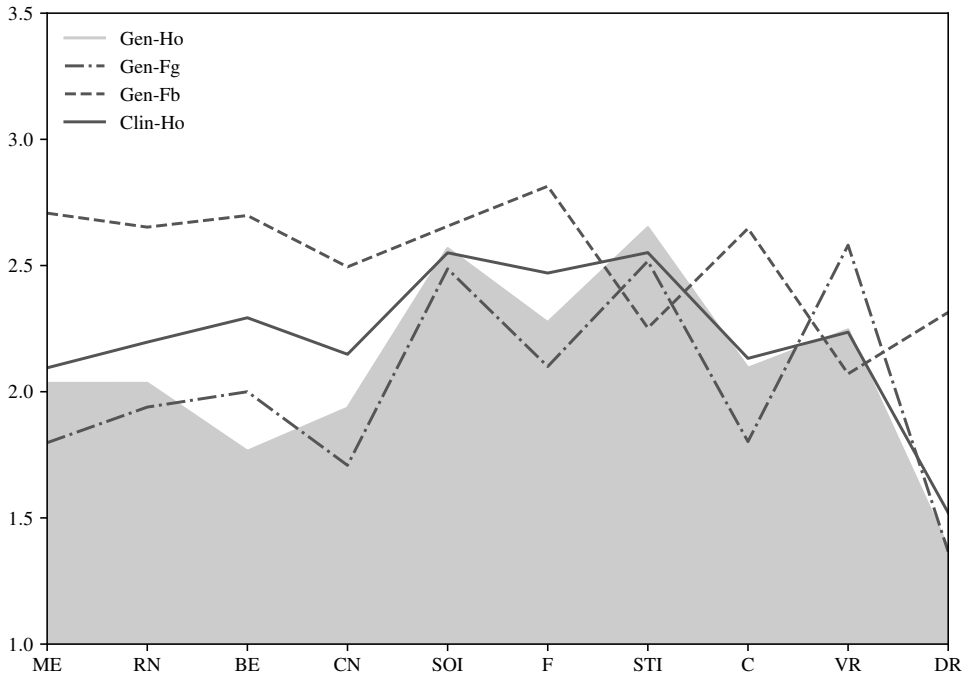

Figura 2.6
PPIR - Valori medi scale di contenuto/controllo

A seguire, sono state condotte delle ANOVA di follow-up con computo degli effetti e

Tabella 2.13: PPIR - Descrittive scale di contentuo/controllo

Pop-Cond		Me	Rn	Be	Cn	Soi	F	Sti	C	Vr	Dr
	mean	2.03	2.03	1.76	1.93	2.57	2.27	2.65	2.09	2.25	1.38
	std	0.44	0.52	0.50	0.41	0.48	0.66	0.56	0.42	0.41	0.39
	min	1.15	1.00	0.94	1.05	1.22	1.00	1.08	1.00	1.23	1.00
Gen-Ho	25%	1.70	1.62	1.38	1.63	2.28	1.79	2.31	1.81	2.00	1.10
	50%	2.00	2.00	1.75	1.89	2.61	2.21	2.69	2.12	2.23	1.30
	75%	2.35	2.38	2.12	2.21	2.89	2.71	3.08	2.38	2.54	1.50
	max	3.70	4.00	3.62	3.37	3.78	4.00	3.92	3.56	3.54	2.80
	mean	1.80	1.94	2.00	1.71	2.49	2.10	2.52	1.80	2.58	1.37
	std	0.46	0.44	0.53	0.38	0.46	0.59	0.50	0.40	0.50	0.31
	min	1.00	1.00	0.94	1.00	1.28	1.00	1.15	1.00	1.46	1.00
Gen-Fg	25%	1.45	1.62	1.56	1.47	2.17	1.64	2.15	1.50	2.23	1.17
	50%	1.75	1.94	2.00	1.68	2.50	2.07	2.54	1.81	2.54	1.30
	75%	2.10	2.25	2.38	1.95	2.83	2.50	2.85	2.00	2.92	1.50
	max	3.35	3.44	3.56	3.53	3.61	3.86	3.85	3.62	4.00	3.60
	mean	2.71	2.65	2.70	2.49	2.66	2.81	2.25	2.65	2.07	2.31
	std	0.75	0.70	0.70	0.69	0.72	0.85	0.63	0.79	0.55	0.75
	min	1.00	1.25	1.06	1.16	1.00	1.00	1.00	1.12	1.00	1.00
Gen-Fb	25%	2.05	2.06	2.19	1.95	2.17	2.14	1.77	2.00	1.69	1.70
	50%	2.75	2.75	2.81	2.47	2.67	2.93	2.19	2.62	2.08	2.40
	75%	3.35	3.25	3.25	2.95	3.17	3.57	2.69	3.31	2.46	2.80
	max	4.00	3.94	3.75	4.00	4.00	4.00	4.00	4.00	3.62	4.00
	mean	2.09	2.20	2.29	2.15	2.55	2.47	2.55	2.13	2.24	1.52
	std	0.48	0.59	0.56	0.49	0.54	0.75	0.58	0.50	0.40	0.38
	min	1.15	1.00	1.12	1.11	1.11	1.00	1.23	1.00	1.38	1.00
Clin-Ho	25%	1.80	1.75	1.88	1.84	2.22	2.05	2.13	1.81	1.92	1.30
	50%	2.05	2.25	2.28	2.11	2.56	2.50	2.54	2.06	2.23	1.40
	75%	2.40	2.58	2.62	2.43	2.90	3.07	2.92	2.50	2.46	1.70
	max	3.40	3.50	3.56	3.79	3.78	4.00	3.85	3.62	3.46	2.70

[a] Me = Machiavellismo egocentrico, Rn = Anticonformismo ribelle, Be = Esternalizzazione della colpa, Cn = Mancanza di pianificazione, Soi = Influenza sociale, F = Mancanza di paura, Sti = Immunità allo stress, C = Freddezza emotiva, Vr = Risposte virtuose, Dr = Risposte devianti

dei contrasti (tabella 2.14). Il test omnibus F è risultato significativo per tutte le scale, mentre la dimensione dell'effetto è stata[34]: (a) grande per la scale ME Egocentrismo machiavellico, RN Anticonformismo ribelle, BE Esternalizzazione della colpa, CN Mancanza di pianificazione, F Mancanza di paura, C Freddezza emotiva, VR Risposte virtuose, DR Risposte devianti; (b) media per la scala STI Immunità allo stress (c) e piccola per le scala SOI Influenza sociale. L'analisi dei contrasti ha confermato le differenze pronunciate tra gli intervistati onesti, le due condizioni di *faking* e il gruppo clinico.

Tabella 2.14: PPIR - ANOVA, effetti e contrasti punteggi scale

Scale PPIR	$F_{(3,1491)}$	ω^2	C1	C2	C3
ME Egocentrismo machiavellico	211.64 ***	0.30	−0.24 ***	0.67 ***	0.06
RN Anticonformismo ribelle	139.59 ***	0.22	−0.09 *	0.62 ***	0.16 **
BE Esternalizzazione colpa	213.16 ***	0.30	0.24 ***	0.93 ***	0.53 ***
CN Mancanza pianificazione	190.25 ***	0.28	−0.23 ***	0.56 ***	0.21 ***
SOI Influenza sociale	6.76 ***	0.01	−0.08 *	0.09 *	−0.02
F Mancanza di paura	81.01 ***	0.14	−0.17 ***	0.54 ***	0.20 **
STI Immunità stress	38.57 ***	0.07	−0.13 **	−0.40 ***	0.10 **
C Freddezza emotiva	174.96 ***	0.26	−0.29 ***	0.55 ***	0.04
VR Risposte virtuose	87.60 ***	0.15	0.34 ***	−0.17 ***	−0.01
DR Risposte devianti	349.32 ***	0.41	−0.01	0.94 ***	0.14 **

[a] C1 = Gen-Ho vs Gen-Fg, C2 = Gen-Ho vs Gen-Fb, C3 = Gen-Ho vs Clin-Ho;
[b] *** p < 0.001, ** p < 0.01, * p < 0.05

Oltre alle precedenti indagini, si è proceduto a valutare in che misura le distribuzioni dei punteggi delle scale PPIR si sovrapponessero nei gruppi di somministrazione (i.e., *Gen-Ho*, *Gen-Fg*, *Gen-Fb* e *Clin-Ho*). La figura 2.7 fornisce un indicazione visiva della questione. Le scale hanno mostrato vari livelli di sovrapposizione, tutti comunque cospicui. In altri termini, nonostante le accertate significatività statistiche delle differenze dei punteggi di scala tra le diverse condizioni di somministrazione, qualsiasi tentativo di separare i gruppi mediante tali punteggi (incluse le scale di controllo VR e DR) avrebbe prodotto un numero rilevante di falsi positivi/negativi [35].

[34] Consulta (Field, 2013) per l'interpretazione della dimensione dell'effetto nelle 3 fasce piccolo/medio/grande.
[35] Da qui la necessità di sviluppare un metodo alternativo di rilevazione del *faking*.

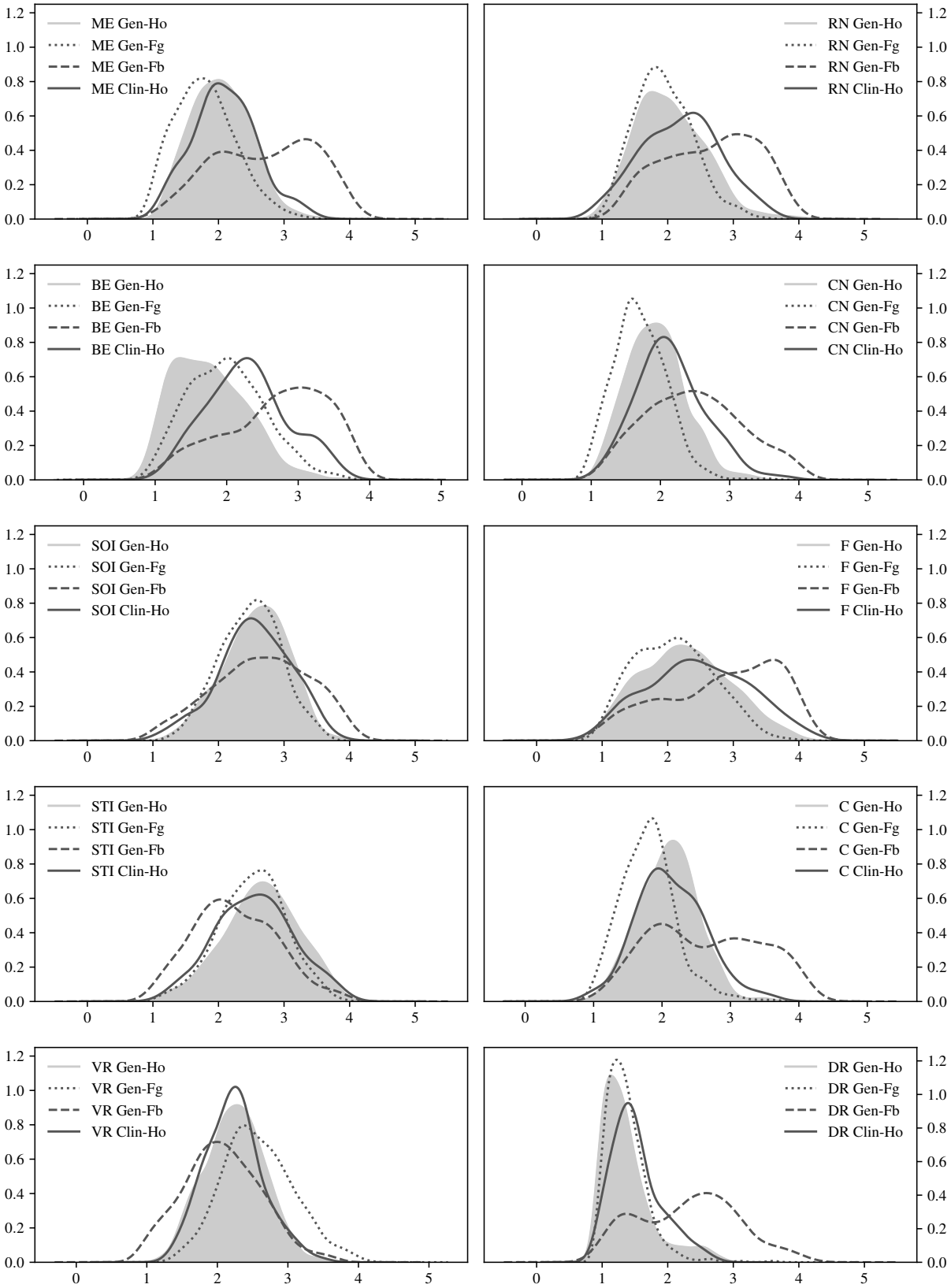

Figura 2.7
PPIR - Distribuzioni scale di contenuto/controllo x condizione

Addestramento dei classificatori ML (*Fake good & bad*)

Come già detto, sono stati addestrati diversi classificatori ML con, adattandoli ai punteggi della scala di PPIR o ai *pattern* di risposta agli item nei due scenari di *faking good* e *faking bad*. È importante sottolineare che le fasi di addestramento (*training*), messa a punto (*hyper-parameters tuning*) e validazione (*testing*) dei modelli ML sono state condotte con dati indipendenti; derogare a questa precauzione avrebbe comportato la sovrastima delle prestazioni reali dei classificatori a causa di un fenomeno che nella letteratura ML prende il nome di *data leakage* (Luo et al., 2016). Al fine di evitare simili problemi, è stata adottata una procedura di convalida incrociata su più livelli (Cawley & Talbot, 2010). Più precisamente, è stato effettuato (a) un ciclo interno di *5-fold cross validation* per l'ottimizzazione dei parametri degli algoritmi ML tramite *random grid search* con F_1 come metrica di riferimento (b) un ciclo esterno di *10-fold cross validation* per la stima *unbiased* della performance dei classificatori implementati. La tabella 2.15 riporta gli indici di prestazione dei classificatori ML *faking good*, mentre la tabella 2.15 è relativi ai classificatori *faking bad*.

Tabella 2.15: PPIR - Performance classificatori ML implementati (*faking good*)

Predittori	Algoritmo	Nome	Acc	Prec	Rec	F_1	AUC
Punt. scala	Regr. Logistica	LR-FG-S	0.71	0.72	0.67	0.69	0.76
	Foresta Casuale	RF-FG-S	0.72	0.71	0.69	0.69	0.78
	XGBoost	XGB-FG-S	0.70	0.68	0.69	0.69	0.76
Matr. item	Regr. Logistica	LR-FG-I	0.85	0.85	0.83	0.84	0.92
	Foresta Casuale	RF-FG-I	0.86	0.88	0.83	0.84	0.94
	XGBoost	XGB-FG-I	0.88	0.88	0.87	0.88	0.94

[a] Acc=Accuracy, Prec=Precision, Rec=Recall, F_1=F_1 score, AUC=Area under AUC curve

Dall'analisi di entrambe le tabelle, si può notare come gli algoritmi addestrati sulle matrici di risposta agli item abbiano garantito performance costantemente superiori rispetto a quelli addestrati sui punteggi di scala. In entrambi gli scenari, il miglior classificatore si è rivelato essere XGB-I (i.e., XGB-FG-I e XGB-FB-I).

Per ricavare una prima impressione della meccanica interna di XGB-FG-I (*faking good*), sono stati considerati i suoi predittori elencati in ordine di "guadagno" (i.e., del contributo fornito al miglioramento del modello). Dei primi 10 predittori maggiormente

Tabella 2.16: PPIR - Performance classificatori ML implementati (*faking bad*)

Predittori	Algoritmo	Nome	Acc	Prec	Rec	F_1	AUC
Punt. scala	Regr. Logistica	LR-FB-S	0.80	0.81	0.75	0.78	0.87
	Foresta Casuale	RF-FB-S	0.81	0.86	0.72	0.79	0.87
	XGBoost	XGB-FB-S	0.80	0.85	0.70	0.77	0.87
Matr. item	Regr. Logistica	LR-FB-I	0.86	0.85	0.85	0.85	0.91
	Foresta Casuale	RF-FB-I	0.88	0.89	0.83	0.86	0.93
	XGBoost	XGB-FB-I	0.90	0.90	0.87	0.89	0.94

[a] Acc=Accuracy, Prec=Precision, Rec=Recall, F_1=F_1 score, AUC=Area under AUC curve

influenti, circa il 25% è risultato derivare dalla scala CN Mancanza di pianificazione, mentre il resto si è distribuito in modo non uniforme tra RN Anticonformismo ribelle (25%), C Freddezza Emotiva (25%), ME Egocentrismo machiavellico (13%) e BE Esternalizzazione della colpa (13%).

Analoga ispezione è stata condotta per XG-FB-I (*faking bad*). Dei primi 10 predittori più rilevanti selezionati, circa il 29% è risultato appartenere alla scala Freddezza emotiva e un altro 29% alla scala BE Esternalizzazione della colpa, mentre il resto si è uniformemente diviso tra CN Mancanza di pianificazione (15%), F Mancanza di paura (15%) e ME Egocentrismo machiavellico (15%).

Comparazione dei migliori classificatori ML (*Fake good & bad*) con CBC

Nell'ultima parte della analisi, i classificatori XGB-FG-I e XGB-FB-I sono stati confrontati rispettivamente con CBC-FG e CBC-FB, i cui *cutoff* normativi sono stati ricavati dal lavoro di Anderson et al. (2013). In particolare, per la scala VR (CBC-FG) è stato usato il punteggio grezzo di 38, mentre per la scala DR (CBC-FB) il punteggio di 25. I riscontri numerici (tabella 2.17) hanno fornito le seguenti importanti indicazioni:

- per lo scenario *faking good*:
  - XGB-FG-I e CBC-FG hanno ottenuto performance superiori rispetto al modello nullo (i.e., che prevede la classe più frequente). Tale modello avrebbe avuto un'accuratezza del 51%, valore inferiore a XGB-FG-I (88%) e a CBC-FG (62%).

  - XGB-FG-I ha superato CBC-FG rispetto a tutti gli indici di performance considerati.

- per lo scenario *faking bad*:
  - XGB-FB-I e CBC-FB hanno ottenuto risultati migliori rispetto al modello nullo che nello scenario in esame avrebbe avuto un'accuratezza del 53%, valore inferiore a XGB-FB-I (90%) e a CBC-FB (73%).
  - XGB-FB-I ha complessivamente superato CBC-FB rispetto agli indici di performance considerati.

La figura 2.8 mostra le distribuzioni di F_1 dei classificatori sopra citati, calcolate a partire da 10 subset del campione di ricerca (*Gen-Ho + Gen-Fg + Gen-Fb)* con una procedura di *10-fold cross-validation.* In entrambi gli scenari di *faking*, i classificatori ML hanno esibito prestazioni superiori.

Tabella 2.17: PPIR - Performance classificatori CBC, XGB-I (*faking good & bad*)

Dataset di ricerca	Classificatore	Acc	Prec	Rec	F_1	AUC
Gen-Ho vs *Gen-Fg*	CBC-FG	0.62	0.82	0.27	0.40	0.61
	XGB-FG-I	0.88	0.88	0.87	0.88	0.94
Gen-Ho vs *Gen-Fb*	CBC-FB	0.73	0.91	0.46	0.61	0.71
	XGB-FB-I	0.90	0.90	0.87	0.89	0.94

[a] Acc=Accuracy, Prec=Precision, Rec=Recall, F_1=F_1 score, AUC= Area under ROC curve

Al fine di ottenere informazioni dettagliate sugli errori predittivi dei classificatori ML e dei classificatori di riferimento (CBC-FG e CBC-FB), sono stati confrontati i profili dei casi misclassificati da XGB-FG-I e CBC-FG con il profilo della classe di appartenenza; analogamente per XGB-FB-I e CBC-FB. I confronti sono stati effettuati mediante ispezione visiva (figura 2.9) e calcolando la distanza di Mahalanobis come misura di prossimità[36] (tabella 2.18). Complessivamente, i risultati hanno mostrato che gli individui classificati erroneamente da XGB-FG-I e XGB-FB-I avevano un profilo di personalità più distante da quello della classe di appartenenza, rispetto al profilo dei casi erroneamente classificati da CBC-FG e CBC-FB.

[36] La distanza di Mahalanobis (o *generalized squared interpoint distance*) è definita come una misura di dissimilarità tra due vettori aleatori $\vec{x}$ e $\vec{y}$ con stessa funzione di densità di probabilità e con matrice di covarianza S: $D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}$ (De Maesschalck et al., 2000).

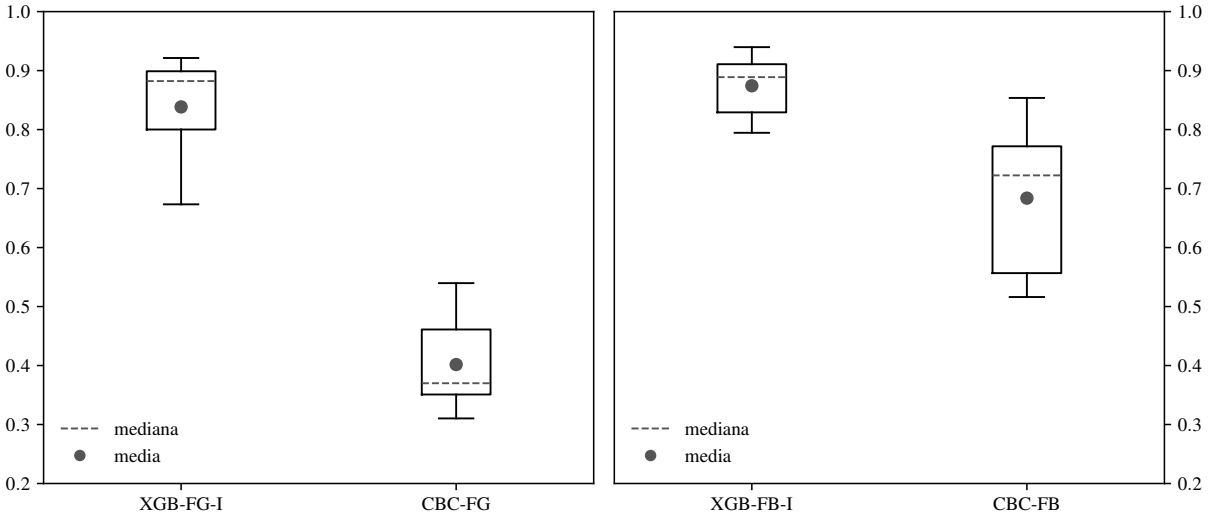

Figura 2.8
PPIR - Distribuzioni F_1 classificatori ML e CBC (pop. gen.)

Tabella 2.18: PPIR - Distanza di Mahalanobis casi misclassificati (pop. gen.)

	MD da classe reale		
	CBC		XGB-I
Gen-Ho misclassificati come *Fake good*	1.50	→	1.98
Fake good misclassificati come *Gen-Ho*	0.47	→	1.62
Gen-Ho misclassificati come *Fake bad*	1.17	→	3.38
Fake bad misclassificati come *Gen-Ho*	1.63	→	5.24

[a] MD = Distanza di Mahalanobis
[b] Prime due righe: CBC-FG vs XGB-FG-I, rimanenti righe: CBC-FB vs XGB-FB-I
[c] La direzione della freccia indica valori migliori

Nell'ultima parte delle analisi, l'attenzione è stata rivolta al gruppo clinico *Clin-Ho*, ricavato da un *dataset* olandese (van Dongen et al., 2017). Non è stato possibile ottenere informazioni sull'attendibilità dei profili PPIR dei pazienti psichiatrici del gruppo. In mancanza di una simile informazione, l'ipotesi di lavoro più conservativa è che tali pazienti abbiano risposto in modo onesto.

Si è dunque proceduto a calcolare il tasso di falsi positivi (FPR) — unico indicatore ottenibile con i dati a disposizione — dei classificatori ML, di CBC-FG e CBC-FB. Nel caso del *faking good*, XGB-FG-I ha esibito un basso FPR (7%) pari a quello di CBC-FG (7%), mentre nel caso del *faking bad*, l'FPR di XGB-FB-I è stato superiore a quello di CBC-FB (13% vs 6%).

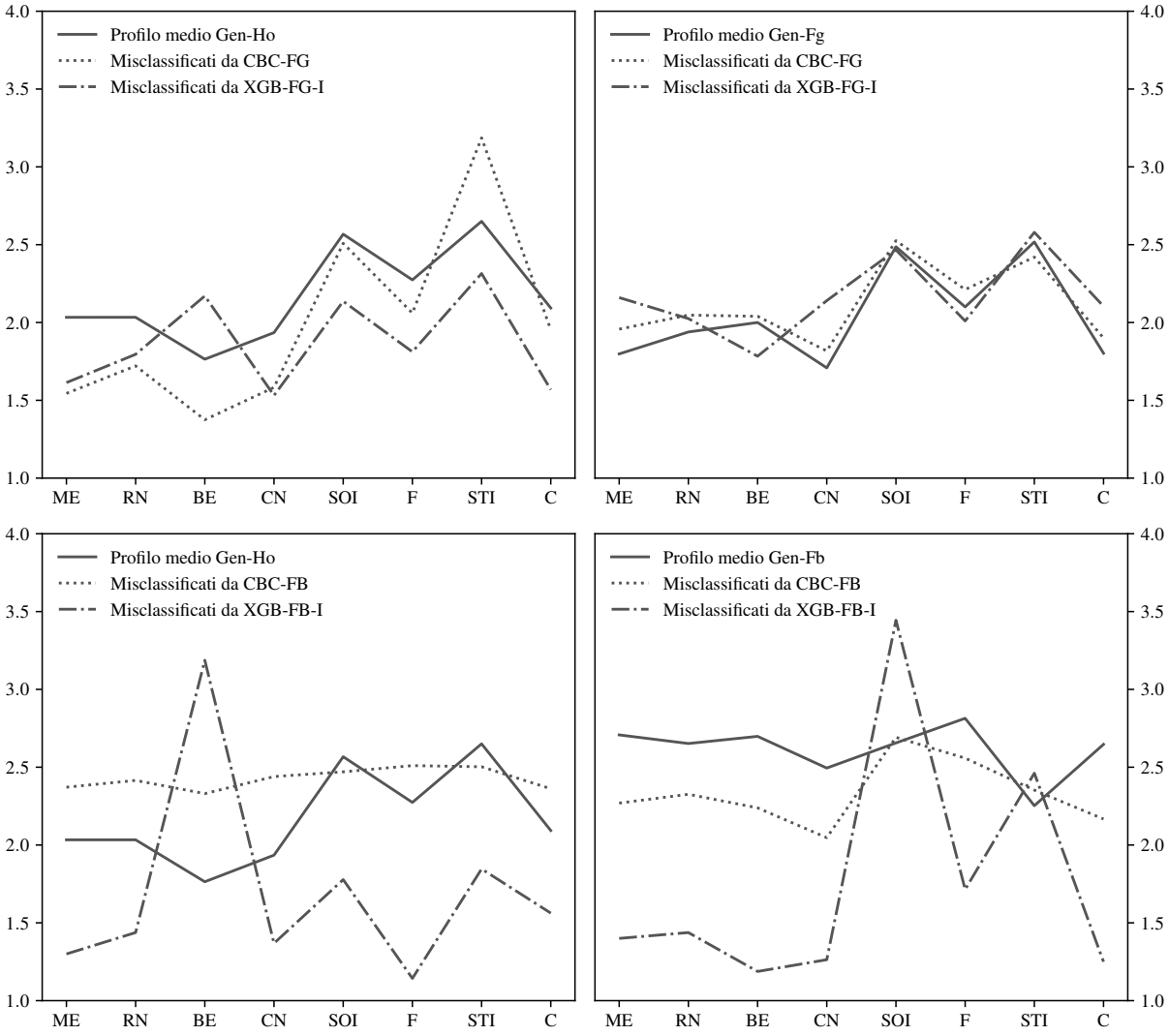

Figura 2.9
PPIR - Profili medi casi misclassificati (pop. gen.)

Infine, l'indagine degli errori predittivi (figura 2.10 e tabella 2.19) ha evidenziato che gli individui classificati erroneamente dai CBC-FG e CBC-FB avevano un profilo di personalità più distante da quello della classe di appartenenza rispetto ai profili dei casi erroneamente classificati da XGB-FG-I e XGB-FB-I. Questi ultimi però — e l'ispezione visiva lo conferma — hanno mostrato un grado di dissimilarità comunque alto, tale da giustificare la mancata attribuzione della classe corretta.

Tabella 2.19: PPIR - Distanza di Mahalanobis casi misclassificati (pop. clin.)

	MD da classe reale		
	CBC		XGB-I
Clin-Ho misclassificati come *Fake good*	1.77	←	1.50
Clin-Ho misclassificati come *Fake bad*	1.81	←	0.82

[a] MD = Distanza di Mahalanobis
[b] Prima riga: CBC-FG vs XGB-FG-I, nella seconda: CBC-FB vs XGB-FB-I
[c] La direzione della freccia indica valori migliori

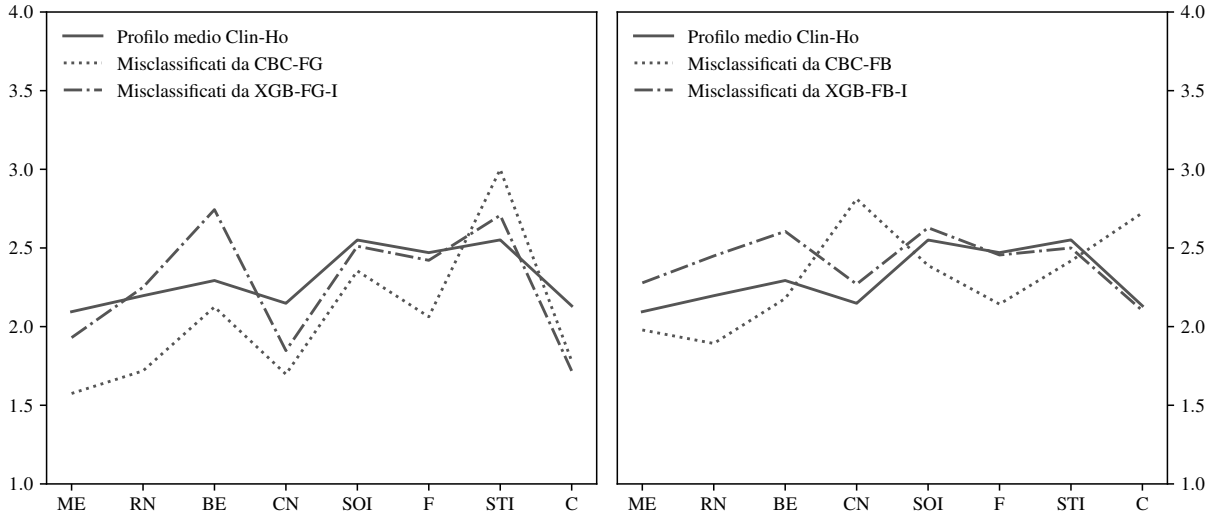

Figura 2.10
PPIR - Profili medi casi misclassificati (pop. clin.)

2.3.4 Discussione

Il presente lavoro ha esplorato la possibilità di impiegare gli algoritmi ML per esaminare le matrici di risposta degli item di un questionario *self-report* sulla psicopatia (PPIR; Lilienfeld e Widows, 2005) allo scopo di rilevare la presenza di *faking* (sia positivo che negativo). L'idea è derivata dai lavori di Kuncel e Borneman (2007) e Kuncel e Tellegen (2009), i quali hanno suggerito di considerare i cosiddetti *idiosyncratic item patterns*[37] come *marker* numerici dei comportamenti di risposta distorsivi.

Il campione di ricerca è stato ottenuto aggregando 3 differenti banche dati — due delle quali provenienti da studi internazionali condotte in Australia e in Olanda — consistente in profili PPIR di studenti universitari (*Gen-Ho*, *Gen-Fg*, *Gen-Fb*) e di pazienti psichia-

[37] Consulta la sezione 2.1 a p. 49 per ulteriori approfindimenti.

trici. Per gli scenari di *faking*, il disegno sperimentale ha previsto l'induzione diretta dei comportamenti distorsivi attraverso la somministrazione del PPIR con istruzioni manipolate. Più in particolare, i sotto-gruppi *Gen-Ho* e *Clin-Ho* hanno ricevuto le consegne standard previste dal manuale, mentre i partecipanti dei sottogruppi *Gen-Fg* e *Gen-Fb* hanno completato il questionario dopo essere stati sollecitati a rispondere in modo da promuovere un'immagine di sé rispettivamente positiva o negativa. Dall'analisi dei punteggi medi delle scale di contenuto e di controllo è emerso che i diversi sotto-gruppi *Gen-Fg* e *Gen-Fb* hanno prodotto profili significativamente diversi — con dimensioni dell'effetto medio-grandi — rispetto ai partecipanti nella condizione *Uni-Ho* e *Clin-Ho*. Ciò nonostante, le distribuzioni dei punteggi di scala sono risultate ampiamente sovrapposte, incluso le scale di controllo VR e DR.

Nella seconda parte delle analisi, sono stati implementati 3 algoritmi ML — una regressione logistica (LR), una foresta casuale (RF) e una macchina XGBoost (XGB) — al fine di generare 12 classificatori ML, 6 dei quali deputati alla rilevazione dei comportamenti di *faking good* e i rimanenti al *faking bad*. Considerando entrambi gli scenari distorsivi (i.e., positivo e negativo), 6 classificatori (LR-FG-S, LR-FB-S, LR-FG-S, LR-FB-S, XGB-FG-S, XGB-FB-S) sono stati addestrati sui punteggi di scala e altri 6 (LR-FG-I, LR-FB-I, RF-FG-I, RF-FB-i, XGB-FG-I, XGB-FB-I) sui *pattern* di risposta dei profili PPIR. I risultati hanno messo in luce la superiorità dei classificatori ML basati sui *pattern* di risposta rispetto agli altri. Intuitivamente, le sottili relazioni intercorrenti tra le risposte contengono informazioni sul *faking* che non sono sfruttate dalle strategie di detezione basate esclusivamente su punteggi di scala.

Tra i classificatori ML implementati, XGB-FG-I e XGB-FB-I si sono rivelati i migliori in termini prestazionali. La ricognizione dell'elenco dei predittori maggiormente influenti ha rivelato che in entrambi i casi sono risultati decisivi alcuni item delle scale CN Mancanza di pianificazione, RN Anticonformismo ribelle, C Freddezza emotiva, F mancanza di paura, BE Esternalizzazione della colpa. Va sottolineato che le relazioni tra tali predittori e i due classificatori ML si sono orientate in direzioni divergenti: nel caso di XGB-FG-I lungo una direttiva di significati tale da veicolare un'immagine di salute mentale, benessere affettivo e pro-socialità; nel caso di XFG-FB-I lungo la direttrice opposta e cioè verso la rappresentazione di un disagio psichico contraddistinto da impulsività, diffidenza/risentimento, senso di abbandono, comportamenti devianti e dis-regolazione

emotiva.

Poiché in genere gli psicologi professionisti valutano il *faking* usando le scale di validità degli strumenti, sono state confrontate le prestazioni di XGB-FG-I e XFG-FB-I con i classificatori di riferimento CBC-FG (basato sulla scala VR Risposte vistuose) e CBC-FB (basato sulla scala DR Risposte devianti). I risultati hanno mostrato che i modelli ML erano costantemente più precisi di CBC-FG e CBC-FB nell'individuare la presenza di contraffazione delle risposte.

L'indagine dei casi misclassificati ha confermato la superiorità di XGB-FG-I e XGB-FB-I. Complessivamente, gli individui classificati erroneamente da tali modelli avevano un profilo di personalità più distante da quello della classe di appartenenza, rispetto al profilo dei casi erroneamente classificati da CBC-FG e CBC-FB. Quest'evidenza suggerisce che il tipo di errori commessi da XGB-FG-I e XGB-FB-I sono stati più sottili, e forse più scusabili, di quelli di CBC-FG e CBC-FB.

Nell'ultima parte delle analisi, l'attenzione è stata rivolta al gruppo clinico *Clin-Ho*, ricavato da una banca dati olandese (van Dongen et al., 2017). Non è stato possibile ottenere informazioni sull'attendibilità dei profili PPIR dei pazienti psichiatrici del gruppo. In mancanza di una simile informazione, l'ipotesi di lavoro più conservativa è che tali pazienti abbiano risposto in modo onesto. Si è dunque proceduto a calcolare il tasso di falsi positivi (FPR) dei classificatori ML, di CBC-FG e CBC-FB. Nel caso del *faking good*, XGB-FG-I ha esibito un basso FPR (7%) pari a quello di CBC-FG (7%), mentre nel caso del *faking bad*, XGB-FB-I ha ottenuto un FPR superiore a quello di CBC-FB (13% vs 6%), producendo più falsi positivi. Fermo restando che tale risultato merita ulteriori approfondimenti empirici, nella valutazione del *faking* è preferibile disporre di strategie detettive ad elevata sensibilità, anche a discapito della specificità (i.e., alti valori di FPR). Infatti, un profilo attendibile classificato come inattendibile (i.e., falso positivo) spingerebbe il clinico a un supplemento di indagine a scopo dirimente (il cui esito potrebbe anche disconfermare l'*early warning* della strategia detettiva). Di converso, un profilo inattendibile classificato come attendibile (i.e., falso negativo) non riceverebbe ulteriori attenzioni da parte del clinico in merito alla sua presunta veridicità.

Capitolo 3

Conclusioni

Gli esperti anglosassoni hanno introdotto il termine *faking* per indicare l'atto cosciente di falsificare le risposte ai questionari *self-report* di personalità. Sulla scorta delle riflessioni di Ziegler et al (2012) è possibile fornire la seguente definizione formale: *il faking è una strategia adottata dall'individuo per fornire descrizioni di sé inaccurate al fine di conseguire un vantaggio personale. Il faking genera differenze sistematiche nei punteggi che non sono dovute al costrutto in esame* (p. 8). Esplicitando meglio la precedente definizione, è possibile evidenziare i seguenti punti. Il *faking*: i) è un comportamento intenzionale; ii) può avere una natura omissiva (tacere le informazioni) e/o commissiva (alterare le informazioni); iii) può portare a una descrizione del sé positiva (*faking good*) o negativa (*faking bad*); iv) è volto al raggiungimento di scopi personali di natura estrinseca[38].

Recentemente, Mueller-Hanson et al. (2006) hanno proposto un modello atto a spiegare i comportamenti di *faking*, integrando precedenti ricerche sull'argomento (McFarland & Ryan, 2006; Snell et al., 1999). Nella loro indagine, gli autori hanno inquadrato il fenomeno nell'ottica della teoria motivazionale di Vroom (1964). Quest'ulitmo ha postulato l'esistenza di tre fattori che spingono gli esseri umani a porre in essere determinati corsi d'azione. In primo luogo, l'aspettativa che essi hanno di poter produrre un certo comportamento (*Expectancy*), in secondo luogo, la convinzione che tale comportamento sia collegato (e dunque conduca) al raggiungimento di un certo risultato (*Instrumnetality*),

[38] Questa precisazione si rende necessaria per distinguere il *faking* dai disturbi fittizi, dove la falsificazione di sintomi fisici o psicologici avviene senza un chiaro incentivo esterno; la motivazione di questo comportamento è quella di assumere il ruolo di malato (Krahn et al., 2008).

infine il valore che essi attribuiscono al risultato medesimo (*Valence*). Per quanto riguarda il fattore *Expectancy*, il cosiddetto *faker* sarà tanto più spinto ad alterare le proprie risposte quanto più si sentirà in grado di farlo. Alcune risultanze empiriche hanno confermato questa ipotesi (e.g., Snell et al., 1999) anche se non mancano gli studi di segno contrario (e.g., Weiner & Gibson, 2000). Relativamente al fattore *Instrumentality*, Mueller-Hanson et al. (2006) hanno elencato una serie di credenze soggettive e tratti di personalità (positivamente o negativamente relati) che inducono gli individui a giudicare i comportamenti di *faking* come strategicamente indispensabili: ad esempio, l'integrità morale, il machiavellismo, la manipolatorietà, la coscienziosità, lo stadio morale e la stabilità emotiva. In buona sostanza, le norme soggettive e l'organizzazione di personalità dell'individuo modellano la sua percezione del *faking*, eventualmente portandolo a ritenere che esso sia un comportamento necessario, o meglio necessariamente collegato al raggiungimento degli scopi che si pone. Infine, per quanto riguarda il fattore *Valence*, Leary e Kowalski (1990) hanno enfatizzato come gli individui siano più motivati ad alterare, anche disonestamente, le impressioni che suscitano negli altri quando tale condotta garantisca loro il raggiungimento di obiettivi personali giudicati importanti. Ellingson et al. (2001) hanno invece notato una relazione inversa tra la disponibilità di un obiettivo — inteso come risorsa — e il suo valore percepito, quasi in ossequio alla legge economica della domanda e dell'offerta. Date queste premesse, e considerato che spesso l'*assessment* psicologico tramite test avviene in contesti con forti incentivi esterni concessi a pochi beneficiari (e.g., reclutamento del personale per l'assegnazione di un posto di lavoro, corresponsione di indennizzi economici per malattia), è legittimo ipotizzare la presenza di comportamenti di *faking* negli individui che si sentono capaci di mentire e che sono caratterialmente propensi a farlo. Heggestad (2012) ha ripreso il modello di Mueller-Hanson et al. (2006) individuando fattori disposizionali, situazionali e fattori legati agli atteggiamenti/opinioni che — in un dinamismo di influenze reciproche — contribuiscono a orientare la tendenza degli individui a contraffare le risposte ai test. In altri termini, gli autori hanno sottolineato la natura processuale del fenomeno, frutto di inclinazioni caratteriali ma anche di pressioni ambientali, suggerendo dunque di considerare il *faking* come un costrutto avente una componente stabile di tratto e una componente di stato più variabile e situazione-specifica (Heggestad, 2012, p. 91).

Dall'analisi della letteratura effettuata, non risulta possibile chiarire in via definitiva se e in che misura il *faking* rappresenti una minaccia alla validità dei questionari. Ragionando in via del tutto ipotetica e assumendo per vera la non relazione, rimarrebbe

comunque irrisolto il problema — altrettanto grave — della validità dei profili individuali di personalità. Christiansen et al. (1994), in uno studio sul 16PF con candidati in *assessment center*, hanno dimostrato che le condotte di *faking* producevano una modificazione dell'ordine di rango nella graduatoria finale (a vantaggio degli *high faker*) per circa l'85% dei candidati scrutinati, benché la presenza di una componente distorsiva nei punteggi di scala non pregiudicasse la relazione del 16PF con i criteri esterni.

Quanto precede è particolarmente vero nella valutazione di alcuni quadri psicopatologici, tra i quali spicca la psicopatia. La definizione moderna del disturbo risale alle intuizioni folgoranti di Cleckley (1976), successivamente sistematizzate (e ampliate) da Hare con l'introduzione della *Psychopathic Checklist* (PCL; 1980, 1991, 2003), strumento concepito dallo psicologo canadese nel tentativo di operazionalizzre i criteri di Cleckley. Egli sviluppò la PCL, oggi edita nella forma rivista a 20 item denominata PCL-R (Hare, 2003), come strumento per compiere una valutazione della psicopatia mediante l'utilizzo di informazioni raccolte a partire da un'intervista semi-strutturata. Muovendo dal modello teorico alla base della PCL-R, si può dire che lo psicopatico di Hare sia definibile sulla base delle sue caratteristiche affettivo-interpersonali (fattore 1), e del suo comportamento deviante (fattore 2). Il Fattore 1 descrive una costellazione di tratti caratterologici in cui prevalgono le condotte manipolatorie unite a un'affettività superficiale il cui dato più eclatante è la mancanza di partecipazione emotiva. Il fattore 2 invece si caratterizza per la presenza di comportamenti antisociali e per uno stile di vita impulsivo, irresponsabile e parassitario. Entrambi i fattori compongono l'immagine di un predatore scaltro, incallito, senza scrupoli, impegnato in un'opera di sistematica spoliazione del mondo, in un contesto emotivo che si potrebbe definire sterile nelle sue componenti positive e fertile invece in quelle negative.

Non è tuttora chiaro (o comunque i risultati emersi necessitano di ulteriori conferme) se e in che misura gli psicopatici siano portati a mentire più spesso degli altri. Similmente, le evidenze di una loro presunta superiorità nel farlo sono poche e controverse. Ciononostante, Anderson et al. (2013) hanno ribadito la necessità di usare *self-report* muniti di scale di controllo. Nella loro ricerca sul PPIR gli autori hanno dimostrato che gli individui — se motivati — sono in grado di elevare/ridurre significativamente i punteggi delle scale del test. In assenza di un meccanismo di verifica dei fenomeni di *over-reporting* o *under-reporting*, le interpretazioni tratte dai protocolli PPIR (e più in generale da qualunque

questionario sulla psicopatia) potrebbero risultare imprecise e fuorvianti.

La strategia più diffusa per la misurazione del *faking*, e anche quella più longeva, è rappresentata dalle scale di controllo o scale *Lie* (Paulhus, 1991). Come è stato già detto nella sezione 1.1, la prima edizione del MMPI (Hathaway & McKinley, 1943) ne includeva tre, a conferma di un'attenzione verso gli stili di risposta consustanziale alla valutazione dei costrutti psicologici. La premessa concettuale alla base di questa strategia di detezione è che il *faking* sia un processo di natura lineare o quasi-lineare (i.e., più alto è il livello di distorsione più alto sarà il punteggio delle scale *Lie*). Nel rispondere positivamente agli item di controllo, gli individui forniscono ripetute evidenze delle loro condotte distorsive. Quando la somma di queste evidenze (i.e., degli item) supera la soglia di attenzione prestabilita (*cutoff score* normativo), gli individui sono classificati come *faker*.

In contrasto con quanto appena esposto, Kuncel e Tellegen (2001) hanno proposto di riconsiderare il problema del *faking* collocandolo nel contesto dei singoli item. Gli autori hanno presentato prove che la desiderabilità/indesiderabilità sociale spesso non è correlata linearmente alle distribuzioni dei tratti di personalità; alcuni item vengono percepiti come massimamente desiderabili a livelli medi (i.e., curva di desiderabilità a "u" inversa), mentre altri a livelli superiori ma comunque non estremi (i.e., curva di desiderabilità ascendente con plateau ed eventuale fase discendente). Per via della loro natura aggregativa, i punteggi di scala tendono a oscurare questi fenomeni item-specifici e ciò a causa dell'effetto livellante della somma algebrica. Kuncel e Borneman (2007) hanno introdotto il concetto di *idiosyncratic item responses* per descrivere le perturbazioni dei comportamenti di risposta che sono indotte dal *faking* e implicitamente hanno suggerito di indagare tali perturbazioni onde intercettare le strategie manipolatorie dei soggetti che mentono.

Alla luce di quanto precede, abbiamo condotto due diversi studi volti a sviluppare una tecnica innovativa di classificazione del *faking* basata sull'analisi dei *pattern* di risposta agli item, mediante l'uso di algoritmi di *machine learning* (ML). L'impostazione generale di entrambi i lavori ha previsto le seguenti fasi: (1) manipolazione diretta del *faking* al fine di ottenere un *dataset* di profili di personalità sia attendibili che distorti (*honest* vs *fake*); (2) implementazione di due o più algoritmi ML in grado di rilevare la presenza di *faking* o nei punteggi di scala o nei *pattern* di risposta; (3) comparazione del miglior algoritmo ML (scelto tra quelli implementati al punto precedente) con il classificatore di

riferimento (CBC) basato sui punteggi delle scale *Lie* e relativi *cutoff* normativi.

Nello studio 1, è stata esplorata la possibilità di impiegare gli algoritmi di *machine learning* ML per esaminare le matrici di risposta degli item di un questionario *self-report* di personalità, il *Big Five Questionnaire 2* (BFQ2; Caprara et al., 2007) allo scopo di rilevare la presenza di *faking*.

I partecipanti sono stati 548 studenti universitari di psicologia dell'Università Sapienza di Roma e Gabriele D'Annunzio di Chieti-Pescara suddivisi in due 2 diverse condizioni: (1) il gruppo dei rispondenti onesti (*Uni-Ho*) che ha completato il questionario dopo aver ricevuto istruzioni standard; (2) il gruppo *fake good* (*Uni-Fg*) ai cui membri è stato chiesto di completare l'inventario immaginando di candidarsi per un posto di lavoro o come insegnante di scuola superiore *Uni-Fg(t)* o come vigile del fuoco *Uni-Fg(f)* e di rispondere in modo da aumentare le *chance* di superare il percorso selettivo. Oltre al campione dei partecipanti, è stato raccolto un *dataset* di profili reali costituito da 4000 casi, 2000 dei quali classificati come *High Faker* e il resto come *Low Faker*.

Sono stati implementati 3 algoritmi ML — una regressione logistica (LR), una foresta casuale (RF) e una macchina XGBoost (XGB) — per generare un totale di 6 classificatori. Tre di essi (LRC-S, RF-S, XGB-S) sono stato adattati ai punteggi delle scale di personalità (E, A, C, S, M) e i rimanenti tre (LRC-I, RF-I, XGB-I) alle matrici di risposta; ciò per separare l'effetto dei due set di predittori (scale vs item) dall'effetto dei diversi algoritmi ML implementati.

Tra i classificatori ML implementati, XGB-I si è rivelato il migliore. L'ispezione dell'elenco dei predittori maggiormente influenti ha rivelato che il gruppo più numeroso era costituito da item appartenenti alla Stabilità Emotiva, seguiti da item afferenti all'area della Coscienziosità, dell'Amicalità e dell'Energia. Il secondo classificatore più efficace è stato LR-I. Le differenze tra LR-I e XGB-I si sono rivelate trascurabili. Questi risultati suggeriscono che il problema di separare i rispondenti onesti dai cosiddetti *faker* può essere affrontato usando un approccio lineare (cioè LR-I) con una perdita minima nella precisione delle previsioni, sebbene gli algoritmi basati sugli alberi decisionali (XGB-I) sia più capaci di rilevare qualsiasi non linearità nelle risposte agli item (Friedman et al., 2001).

Poiché in genere gli psicologi professionisti valutano il *faking* usando le scale di validità

degli strumenti, sono state confrontate le prestazioni di XGB-I con CBC, un classificatore basato sulla scala *Lie* del BFQ2. Per effettuare tale comparazione è stato usato il *dataset Uni*, composto da profili di personalità da tre gruppi di studenti universitari, *Uni-Ho*, *Uni-Fg(t), Uni-Fg(f).* In linea con la letteratura (e.g., Birkeland et al., 2006), i gruppi *Uni-Fg* hanno prodotto elevazioni significative nei profili — con dimensioni dell'effetto medio-grandi — rispetto ai partecipanti nella condizione *Uni-Ho*. Ciò nonostante, i gruppi sono risultati sovrapposti su quasi tutte le scale, suggerendo che poteva essere difficile separarli usando un punteggio di *cutoff* a livello di scala. XGB-I e CBC sono stati applicati a 1000 repliche bootstrap del *dataset Uni*. I risultati hanno mostrato che XGB-I era costantemente più preciso di CBC nel individuare la presenza di *faking*. Quando le prestazioni sono state valutate in termini di previsioni errate, XGB-I si è rivelato, ancora un volta, migliore di CBC. Gli intervistati onesti e i *faker* classificati erroneamente da XGB-I avevano configurazioni di punteggi più dissimili dal profilo prototipico del gruppo al quale appartenevano rispetto agli intervistati classificati erroneamente da CBC. In altri termini, il tipo di errori commessi da XGB-I sono stati più sottili, e forse più scusabili, di quelli di CBC.

Nel secondo studio — non mera replica, ma estensione del primo — è stato impiegato un questionario per la valutazione della psicopatia, il *Psychopathic Personality Inventory - Revised* (PPIR; Lilienfeld e Widows, 2005) con un campione di studenti universitari e con uno di pazienti psichiatrici. È stata esplorata la possibilità di impiegare gli algoritmi di *machine learning* per esaminare le matrici di risposta degli item allo scopo di rilevare la presenza di *faking* sia positivo che negativo. La decisione di somministrare un questionario sulla psicopatia è scaturita dalla constatazione che gli individui con tratti di personalità psicopatica possono esibire condotte distorsive e manipolatorie e dunque gli strumenti di misurazione sviluppati per questa classe di pazienti dovrebbero essere muniti di tecniche efficaci di rilevamento del *faking*.

Il campione dei partecipanti è stato ottenuto aggregando 3 differenti banche dati — due delle quali provenienti da studi internazionali condotte in Australia e in Olanda — consistente in profili PPIR di studenti universitari (*Gen-Ho*, *Gen-Fg*, *Gen-Fb*) e di pazienti psichiatrici. Per gli scenari di *faking*, il disegno sperimentale ha previsto l'induzione diretta dei comportamenti distorsivi attraverso la somministrazione del PPIR con istruzioni manipolate. Più in particolare, i sotto-gruppi *Gen-Ho* e *Clin-Ho* hanno ricevuto le

consegne standard previste dal manuale, mentre i partecipanti dei sottogruppi *Gen-Fg* e *Gen-Fb* hanno completato il questionario dopo essere stati sollecitati a rispondere in modo da promuovere un'immagine di sé rispettivamente positiva e negativa. Dall'analisi dei punteggi medi delle scale di contenuto e di controllo è emerso che i sotto-gruppi *Gen-Fg* e *Gen-Fb* hanno prodotto profili significativamente diversi — con dimensioni dell'effetto medio-grandi — rispetto ai partecipanti nella condizione *Uni-Ho* e *Clin-Ho*. Ciò nonostante, le distribuzioni dei punteggi di scala sono risultate ampiamente sovrapposte, incluso le scale di controllo VR e DR.

Nella seconda parte delle analisi, sono stati implementati 3 algoritmi ML — una regressione logistica (LR), una foresta casuale (RF) e una macchina XGBoost (XGB) — al fine di generare 12 classificatori ML, 6 dei quali deputati alla rilevazione dei comportamenti di *faking good* e i rimanenti al *faking bad*. Considerando entrambi gli scenari distorsivi (i.e., positivo e negativo), 6 classificatori (LR-FG-S, LR-FB-S, LR-FG-S, LR-FB-S, XGB-FG-S, XGB-FB-S) sono stati addestrati sui punteggi di scala e altri 6 (LR-FG-I, LR-FB-I, RF-FG-I, RF-FB-i, XGB-FG-I, XGB-FB-I) sui *pattern* di risposta dei profili PPIR. I risultati hanno messo in luce la superiorità dei classificatori ML basati sui *pattern* di risposta rispetto agli altri. Intuitivamente, le sottili relazioni intercorrenti tra gli item contengono informazioni sul *faking* che non sono sfruttate dalle strategie di detezione basate esclusivamente su punteggi di scala.

Tra i classificatori ML implementati, XGB-FG-I e XGB-FB-I si sono rivelati i migliori in termini prestazionali. La ricognizione dell'elenco dei predittori maggiormente influenti ha rivelato che in entrambi i casi sono risultati decisivi alcuni item delle scale CN Mancanza di pianificazione, RN Anticonformismo ribelle, C Freddezza emotiva, F mancanza di paura, BE Esternalizzazione della colpa. Va sottolineato che le relazioni tra tali predittori e i due classificatori ML si sono orientate in direzioni divergenti: nel caso di XGB-FG-I lungo una direttiva di significati tale da veicolare un'immagine di salute mentale, benessere affettivo e pro-socialità; nel caso di XFG-FB-I lungo la direttrice opposta e cioè verso la rappresentazione di un disagio psichico contraddistinto da impulsività, diffidenza/risentimento, senso di abbandono, comportamenti devianti e dis-regolazione emotiva.

Successivamente, sono state confrontate le prestazioni di XGB-FG-I e XFG-FB-I con i classificatori di riferimento CBC-FG (basato sulla scala VR Risposte vistuose) e CBC-FB

(basato sulla scala DR Risposte devianti). I risultati hanno mostrato che i modelli ML erano costantemente più precisi di CBC-FG e CBC-FB nell'individuare la presenza di contraffazione delle risposte. L'indagine dei casi misclassificati ha confermato la superiorità di XGB-FG-I e XGB-FB-I. Complessivamente, gli individui classificati erroneamente da tali modelli avevano un profilo di personalità più distante da quello della classe di appartenenza, rispetto al profilo dei casi erroneamente classificati da CBC-FG e CBC-FB. Quest'evidenza suggerisce che il tipo di errori commessi da XGB-FG-I e XGB-FB-I sono stati più sottili, e forse più scusabili, di quelli di CBC-FG e CBC-FB.

Nell'ultima parte delle analisi, l'attenzione è stata rivolta al gruppo clinico *Clin-Ho*, ricavato da una banca dati olandese (van Dongen et al., 2017). Non è stato possibile ottenere informazioni sull'attendibilità dei profili PPIR dei pazienti psichiatrici del gruppo. In mancanza di una simile informazione, l'ipotesi di lavoro più conservativa è che tali pazienti abbiano risposto in modo onesto. Si è dunque proceduto a calcolare il tasso di falsi positivi (FPR) dei classificatori ML, di CBC-FG e CBC-FB. Nel caso del *faking good*, XGB-FG-I ha esibito un basso FPR (7%) pari a quello di CBC-FG (7%), mentre nel caso del *faking bad*, XGB-FB-I ha ottenuto un FPR superiore a quello di CBC-FB (13% vs 6%), producendo più falsi positivi. Fermo restando che tale risultato merita ulteriori approfondimenti empirici, nella valutazione del *faking* è preferibile disporre di strategie detettive ad elevata sensibilità, anche a discapito della specificità (i.e., alti valori di FPR). Infatti, un profilo attendibile classificato come inattendibile (i.e., falso positivo) spingerebbe il clinico a un supplemento di indagine a scopo dirimente (il cui esito potrebbe anche disconfermare l'*early warning* della strategia detettiva). Di converso, un profilo inattendibile classificato come attendibile (i.e., falso negativo) non riceverebbe ulteriori attenzioni da parte del clinico in merito alla sua presunta veridicità.

Entrambi gli studi presentano due importanti limitazioni. In primo luogo, la manipolazione diretta dei comportamenti di *faking* è stata criticata da alcuni autori. Riprendendo le parole di Smith e McDaniel (2012): *"La manipolazione diretta del faking negli studi sperimentali [...] non può dare un'indicazione precisa di come le persone si comporteranno nelle situazioni reali. Al limite, fornisce informazioni sull'estensione massima delle condotte distorsive quando gli individui decidono di mentire. [...] È proprio il problema della motivazione che viene rimosso in questo genere di studi e ciò limita la nostra comprensione del problema."* (p. 55); in altri termini, la manipolazione diretta del *faking*

chiarisce quella che potrebbe essere la performance di picco (capacità massima) e non la performance effettivamente riscontrata nei *setting* naturali (capacità tipica). Studi futuri dovranno affrontare questo problema misurando il *faking* in scenari realistici.

In secondo luogo, i *dataset* utilizzati per il confronto tra i classificatori ML e i modelli di riferimento basati sulle scale di controllo erano sbilanciati dal punto di vista del genere e dell'età, con una componente di giovani studentesse sovra-rappresentata (in entrambi gli studi). Non è stato possibile valutare l'accuratezza dei classificatori ML con dati più eterogenei. Relativamente allo studio 2, inoltre, è mancata la possibilità di valutare appieno le potenzialità dell'approccio ML con il campione clinico *Clin-Ho*, poiché non si disponeva dei dati relativi alle condotte di *faking* dei pazienti psichiatrici.

Nonostante queste limitazioni, l'approccio qui proposto appare molto interessante poiché consentirebbe di ridurre la lunghezza dei questionari di personalità eliminando gli item delle scale di controllo. Forse anche in modo più interessante, tale approccio potrebbe essere usato per aggiungere un meccanismo di rilevazione del *faking* ai *self-report* che sono sprovvisti di strategie per la detezione degli stili di risposta distorsivi. In sintesi, sono state raccolte prove sufficienti — almeno rispetto agli strumenti usati nei due studi — per affermare che i classificatori ML basati sule matrici di risposta agli item rappresentano una valida alternativa alle scale *Lie*.

Da ultimo, vale la pena di sottolineare che gli algoritmi di *machine learning* potrebbero essere impiegati — si presume con esiti ugualmente interessanti — per ulteriori problemi di natura classificatoria come, ad esempio, (i) la diagnosi clinica, o (ii) la specificazione di sotto-tipi diagnostici, ipotizzando la presenza di schemi di risposta agli item distintivi per ogni classe nosografica.

Bibliografia

Alpaydin, E. (2009). *Introduction to machine learning*. Cambridge, MA, USA: MIT press.

American Psychiatric Association. (1952). *Diagnostic and statistical manual of mental disorders* (1st ed). Washington, DC, USA: Amer Psychiatric Pub Incorporated.

American Psychiatric Association. (1968). *Diagnostic and statistical manual of mental disorders* (2st ed). Washington, DC, USA: Amer Psychiatric Pub Incorporated.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed). Washington, DC, USA: Amer Psychiatric Pub Incorporated.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed). Washington, DC, USA: Amer Psychiatric Pub Incorporated.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed). Washington, DC, USA: Amer Psychiatric Pub Incorporated.

Anderson, J. L., Sellbom, M., Wygant, D. B. & Edens, J. F. (2013). Examining the necessity for and utility of the Psychopathic Personality Inventory—Revised (PPI–R) validity scales. *Law and Human Behavior*, *37*(5), 312.

Andrade, J. T. (2008). The inclusion of antisocial behavior in the construct of psychopathy: A review of the research. *Aggression and Violent Behavior*, *13*(4), 328–335.

Ayodele, T. O. (2010). Types of machine learning algorithms. In Z. Yagang (Cur.), *New advances in machine learning* (Cap. 3, pp. 20–48). IntechOpen.

Baer, R. & Miller, J. (2002). Underreporting of psychopathology on the MMPI-2: A meta-analytic review. *Psychological Assessment*, *14*(1), 16–26.

Baer, R. A., Wetter, M. W., Nichols, D. S., Greene, R. & Berry, D. T. (1995). Sensitivity of MMPI-2 validity scales to underreporting of symptoms. *Psychological Assessment*, *7*(4), 419.

Barrick, M. R. & Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of applied psychology*, *81*(3), 261.

Biderman, M. (2014). Against all odds: bifactors in EFAs of Big Five Data. In *Soc Ind Org Psychol (29th annual conference)*, Honolulu, Hawaii.

Biderman, M. D. & Nguyen, N. T. (2004). Structural equation models of faking ability in repeated measures designs. In *19th annual society for industrial and organizational psychology conference, Chicago, IL*.

Binet, A. (1905). Binet-Simon intelligence scale, New Methods for the Diagnosis of the Intellectual Level of Subnormals. *L'Année Psychologique*, *12*, 191–244.

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T. & Smith, M. A. (2006, novembre). A Meta-Analytic Investigation of Job Applicant Faking on Personality Measures: Job Applicant Faking on Personality Measures. *International Journal of Selection and Assessment*, *14*(4), 317–335.

Blackburn, R. (2007). Personality disorder and antisocial deviance: Comments on the debate on the structure of the psychopathy checklist-revised. *Journal of personality disorders*, *21*(2), 142–159.

Bodholdt, R. H., Richards, H. R. & Gacono, C. B. (2000). Assessing psychopathy in adults: The Psychopathy Checklist-Revised and screening version. *The clinical and forensic assessment of psychopathy: A practitioner's guide*, 55–86.

Book, A., Holden, R., Starzyk, K., Wasylkiw, L. & Edwards, M. (2006). Psychopathic traits and experimentally induced deception in self-report assessment. *Personality and Individual Differences*, *41*(601–608), 10–1016.

Boone, K., Savodnik, I., Ghaffarian, S., Lee, A., Freeman, D. & Berman, N. (1995). Rey 15-item memorization and dot counting scores in a "stress" compensation population: Relationship to personality (MCMI) scores. *Journal of Clinical Psychology*, *51*, 457–463.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984). Classification and Regression Trees. Republished by CRC Press. Wadsworth, Belmont, CA.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, *16*(3), 199–231.

Brislin, S. J., Drislane, L. E., Smith, S. T., Edens, J. F. & Patrick, C. J. (2015). Development and validation of triarchic psychopathy scales from the Multidimensional Personality Questionnaire. *Psychological assessment*, *27*(3), 838.

Brown, R. D. & Harvey, R. J. (2003). Detecting personality test faking with appropriateness measurement: Fact or fantasy. In *Annual Conference of the Society for Industrial and Organizational Psychology*. Citeseer, Orlando.

Brown, R. & Barrett, P. (1999). Differences between applicant and non-applicant personality questionnaire data: Some implications for the creation and use of norm tables. In *BPS Test User Conference Proceedings* (pp. 76–86).

Brunetti, D. G., Schlottmann, R. S., Scott, A. B. & Hollrah, J. L. (1998). Instructed faking and MMPI-2 response latencies: The potential for assessing response validity. *Journal of Clinical Psychology*, *54*(2), 143–153.

Butcher, J., Dalstrom, W., Graham, J., Tellegen, A. & Kraemmer, B. (1989). *Manual for administering and scoring the MMPI-2*. Minneapolis, MN, USA: University of Minnesota Press.

Butcher, J. N., Morfitt, R. C., Rouse, S. V. & Holden, R. R. (1997). Reducing MMPI-2 defensiveness: The effect of specialized instructions on retest validity in a job applicant sample. *Journal of Personality Assessment*, *68*(2), 385–401.

Caprara, G., Barbaranelli, C., Borgogni, L. & Vecchione, M. (2007). *Big Five Questionnaire: Manual*. Firenze, FI, ITA: Organizzazioni Speciali.

Cawley, G. C. & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*(Jul), 2079–2107.

Chen, T. & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). ACM.

Cherkassky, V. & Mulier, F. M. (2007). *Learning from data: concepts, theory, and methods*. Hoboken, NJ, USA: John Wiley & Sons.

Christiansen, N. D., Goffin, R. D., Johnston, N. G. & Rothstein, M. G. (1994). Correcting the 16PF for faking: Effects on criterion-related validity and individual hiring decisions. *Personnel Psychology*, *47*(4), 847–860.

Cima, M., Merckelbach, H., Hollnack, S., Butt, C., Kremer, K. & Schellbach-Matties, R. (2003). The other side of malingering: Supernormality. *The Clinical Neuropsychologist*, *17*, 235–243.

Cima, M. & van Oorsouw, K. (2013). The relationship between psychopathy and crime related amnesia. *International Journal of Law and Psychiatry*, *36*, 23–29.

Ciulla, S. (2010). *Valutazione psicologica della psicopatia* (Tesi di dottorato, Università degli studi di Palermo, Facoltà di scienze della formazione, Dipartimento di Psicologia).

Cleckley, H. (1976). *The mask of sanity* (5th ed). St. Louis, MO, USA: Mosby.

Cooke, D. J., Hart, S. D., Logan, C. & Michie, C. (2012). Explicating the construct of psychopathy: Development and validation of a conceptual model, the Comprehensive Assessment of Psychopathic Personality (CAPP). *International Journal of Forensic Mental Health*, *11*(4), 242–252.

Cooke, D. J., Kosson, D. S. & Michie, C. (2001). Psychopathy and ethnicity: Structural, item, and test generalizability of the Psychopathy Checklist—Revised (PCL-R) in Caucasian and African American participants. *Psychological assessment*, *13*(4), 531.

Cooke, D. J. & Michie, C. (1997). An item response theory analysis of the Hare Psychopathy Checklist–Revised. *Psychological assessment*, *9*(1), 3.

Copestake, S., Gray, N. S. & Snowden, R. J. (2011). A comparison of a self-report measure of psychopathy with the psychopathy checklist-revised in a UK sample of offenders. *Journal of Forensic Psychiatry & Psychology*, *22*(2), 169–182.

Crowhurst, B. & Coles, E. (1989). Kurt Schneider's concepts of psychopathy and schizophrenia: a review of the English literature. *The Canadian Journal of Psychiatry*, *34*(3), 238–243.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, *2*(4), 303–314.

De Maesschalck, R., Jouan-Rimbaud, D. & Massart, D. L. (2000). The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, *50*(1), 1–18.

Delain, S., Stafford, K. & Ben-Porath, Y. (2003). Use of the TOMM in a criminal court forensic assessment setting. *Assessment*, *10*, 370–381.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K. & Cooper, H. (2003). Cues to deception. *Psychological bulletin*, *129*(1), 74.

Domingos, P. (2000). A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning* (pp. 231–238).

Douglas, E. F., McDaniel, M. A. & Snell, A. F. (1996). The Validity of Non-cognitive Measures Decays When Applicants Fake. In *Academy of Management Proceedings* (Vol. 1996, *1*, pp. 127–131). Academy of Management Briarcliff Manor, NY 10510.

Dunnette, M. D., McCartney, J., Carlson, H. C. & Kirchner, W. K. (1962). A study of faking behavior on a forced-choice self-description checklist. *Personnel Psychology*, *15*(1), 13–24.

Dwyer, D. B., Falkai, P. & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual review of clinical psychology*, *14*, 91–118.

Edens, J., Buffington, J. & Tomicic, T. (2000). An investigation of the relationship between psychopathic traits and malingering on the psychopathic personality inventory. *Assessment*, *7*, 281–296.

Ellingson, J. E., Smith, D. B. & Sackett, P. R. (2001). Investigating the influence of social desirability on personality factor structure. *Journal of Applied Psychology*, *86*(1), 122.

Ferrando, P. J. & Chico, E. (2001). Detecting dissimulation in personality test scores: A comparison between person-fit indices and detection scales. *Educational and Psychological Measurement*, *61*(6), 997–1012.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (4th). Newbury Park, CA: Sage Publications Ltd.

First, M., Gibbon, M., Spitzer, R., Williams, J. & Benjamin, L. (1997). *Structured interview for DSM-IV axis II disorders (SCID-II)*. Washington, DC, USA: American Psychiatric Press.

Forth, A. E., Kosson, D. S. & Hare, R. D. (2003). *Hare psychopathy checklist: Youth version*. Toronto, ON, CAN: Multi-Health Systems.

Freeman, J. & Samson, F. (2012). Are you telling the truth? Psychopathy assessment and impression management in a community sample. *The Open Criminology Journal*, *5*, 16–23.

Friedman, J., Hastie, T. & Tibshirani, R. (2001). *The elements of statistical learning*. New York, NY, USA: Springer series in statistics.

Grice, H. P., Cole, P., Morgan, J. et al. (1975). Logic and conversation. *1975*, 41–58.

Grillo, J., Brown, R., Hilsabeck, R., Price, J. & Lees-Haley, P. (1994). Raising doubts about claims of malingering: Implications of relationships between MCMI-II and MMPI-2 performances. *Journal of Clinical Psychology*, *50*(651–655), 4–651.

Gupta, B., Rawat, A., Jain, A., Arora, A. & Dhami, N. (2017). Analysis of various decision tree algorithms for classification in data mining. *International Journal of Computer Applications*, *163*(8), 15–19.

Gurley, J. R. (2009). A history of changes to the criminal personality in the DSM. *History of Psychology*, *12*(4), 285.

Häkkänen-Nyholm, H. & Nyholm, J.-O. (2012). *Psychopathy and law: A practitioner's guide*. New York, NY, USA: John Wiley & Sons.

Hall, R. C. & Hall, R. C. (2012). Plaintiffs who malinger: Impact of litigation on fake testimony. In M. Ziegler, C. MacCann & R. Roberts (Cur.), *New perspectives on faking in personality assessment* (Cap. 16, pp. 255–281). New York, NY, US: Oxford University Press.

Hanin, B. (2017). Universal function approximation by deep neural nets with bounded width and relu activations. *arXiv preprint arXiv:1708.02691*.

Hare, R. D. (1980). A research scale for the assessment of psychopathy in criminal populations. *Personality and individual differences*, *1*(2), 111–119.

Hare, R. D. (1991). *Manual for the Hare Psychopathy Checklist* (Revised). Toronto, ON, CAN: Multi-Health Systems.

Hare, R. D. (2003). *The Hare Psychopathy Checklist-Revised* (2ª ed.). Toronto, ON, CAN: Multi Health Systems.

Hare, R. D. (2009). *La psicopatia: valutazione diagnostica e ricerca empirica*. Roma, RM, ITA: Astrolabio.

Hare, R. D., Harpur, T. J., Hakstian, A. R., Forth, A. E., Hart, S. D. & Newman, J. P. (1990). The revised psychopathy checklist: reliability and factor structure. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *2*(3), 338.

Hare, R. D., Hart, S. D. & Harpur, T. J. (1991). Psychopathy and the DSM-IV criteria for antisocial personality disorder. *Journal of abnormal psychology*, *100*(3), 391.

Hart, S., Cox, D. & Hare, R. D. (1995). *Manual for the Hare psychopathy checklist: Screening version*. Toronto, ON, CAN: Multi-Health Systems.

Hathaway, S. & McKinley, J. (1943). *Manual for administering and scoring the MMPI*. Minneapolis, MN, USA: National Computer Systems.

Heggestad, E. D. (2012). A conceptual representation of faking: Putting the horse back in front of the cart. In M. Ziegler, C. MacCann & R. Roberts (Cur.), *New perspectives on faking in personality assessment* (Cap. 6, pp. 87–102). New York, NY, US: Oxford University Press.

Heinze, M. & Vess, J. (2005). The relationship among malingering, psychopathy, and the MMPI-2 validity scales in maximum security forensic psychiatric inpatients. *Journal of Forensic Psychology Practice*, *5*, 35–53.

Henry, M. S. & Raju, N. S. (2006). The effects of traited and situational impression management on a personality test: An empirical analysis. *Psychology Science*, *48*(3), 247.

Hildebrand, M. & de Ruiter, C. (2004). PCL-R psychopathy and its relation to DSM-IV Axis I and II disorders in a sample of male forensic psychiatric patients in the Netherlands. *International journal of law and psychiatry*, *27*(3), 233–248.

Holden, R. (1996). *Holden psychological screening inventory manual*. Tonawanda, NY, USA: Multi-Health Systems.

Holden, R. R. (2007, luglio). Socially desirable responding does moderate personality scale validity both in experimental and in nonexperimental contexts. *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement*, *39*(3), 184–201.

Holden, R. R., Fekken, G. C. & Cotton, D. H. (1991). Assessing psychopathology using structured test-item response latencies. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *3*(1), 111.

Holden, R. R. & Jackson, D. N. (1985). Disguise and the structured self-report assessment of psychopathology: I. An analogue investigation. *Journal of Consulting and Clinical Psychology*, *53*(2), 211.

Holden, R. R. & Kroner, D. G. (1992). Relative efficacy of differential response latencies for detecting faking on a self-report measure of psychopathology. *Psychological Assessment*, *4*(2), 170.

Holden, R. R., Wood, L. L. & Tomashewski, L. (2001). Do response time limitations counteract the effect of faking on personality inventory validity? *Journal of Personality and Social Psychology*, *81*(1), 160.

Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin*, *30*(2), 161–172.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D. & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of applied psychology*, *75*(5), 581.

Hsu, L. M., Santelli, J. & Hsu, J. R. (1989). Faking detection validity and incremental validity of response latencies to MMPI subtle and obvious items. *Journal of Personality assessment*, *53*(2), 278–295.

Huber, C. H. (2017). *Faking and the Validity of Personality Tests: Using New Faking-Resistant Measures to Study Some Old Questions* (Tesi di dottorato, University of Minnesota).

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M. & Jordan, M. I. (2017). How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1724–1732). JMLR.

Kohlberg, L. (1971). Stages of moral development. *Moral education*, *1*(51), 23–92.

Krahn, L. E., Bostwick, J. M. & Stonnington, C. M. (2008). Looking toward DSM–V: should factitious disorder become a subtype of somatoform disorder? *Psychosomatics*, *49*(4), 277–282.

Kucharski, L., Duncan, S., Egan, S. & Falkenbach, D. (2006). Psychopathy and malingering of psychiatric disorder in criminal defendants. *Behavioral Sciences and the Law*, *24*, 633–644.

Kuncel, N. & Tellegen, A. (2001). The Social Desirability of Personality Items and Endorsed Trait Level: A reconceptualization of the measurement of item desirability. In *Poster Session Presented at the Annual Meeting of the Society for Industrial and Organizational Psychology.*

Kuncel, N. R. & Borneman, M. J. (2007, giugno). Toward a New Method of Detecting Deliberately Faked Personality Tests: The use of idiosyncratic item responses. *International Journal of Selection and Assessment*, *15*(2), 220–231.

Kuncel, N. R. & Tellegen, A. (2009, giugno). A Conceptual And Empirical Reexamination Of The Measurement Of The Social Desirability Of Items: Implications For Detecting Desirable Response Style And Scale Development. *Personnel Psychology*, *62*(2), 201–228.

Lazarus, R. S. (1998). The costs and benefits of denial. *Fifty years of the research and theory of RS Lazarus: An analysis of historical and perennial issues*, *1*, 227–251.

Leary, M. R. & Kowalski, R. M. (1990). Impression management: A literature review and two-component model. *Psychological bulletin*, *107*(1), 34.

Levenson, M. R., Kiehl, K. A. & Fitzpatrick, C. M. (1995). Assessing psychopathic attributes in a noninstitutionalized population. *Journal of personality and social psychology*, *68*(1), 151.

Lewis, M. (2010). *Understanding the Affective and Cognitive Components of Psychopathy: Developing a New Assessment* (Tesi di dottorato, University of Central Lancashire, Inghilterra).

Lilienfeld, S. O. (1994). Conceptual problems in the assessment of psychopathy. *Clinical Psychology Review*, *14*(1), 17–38.

Lilienfeld & Andrews. (1996). Development and preliminary validation of a self-report measure of psychopathic personality traits in noncriminal population. *Journal of personality assessment*, *66*(3), 488–524.

Lilienfeld, Fowler, Katherine & Patrick. (2006). The self-report assessment of psychopathy. *Handbook of psychopathy*, 107–132.

Lilienfeld & Widows. (2005). *Professional manual for the psychopathic personality inventory-revised (PPI-R)*. Lutz, FL, USA: Psychological Assessment Resources.

Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., ... Berk, M. (2016, dicembre). Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *Journal of medical Internet research*, *18*(12), e323.

Lynam, D. R., Gaughan, E. T., Miller, J. D., Miller, D. J., Mullins-Sweatt, S. & Widiger, T. A. (2011). Assessing the basic traits associated with psychopathy: Development and validation of the Elemental Psychopathy Assessment. *Psychological Assessment*, *23*(1), 108.

MacNeil, B. & Holden, R. (2006). Psychopathy and the detection of faking self-report inventories of personality. *Personality and Individual Differences*, *41*, 641–651.

Malterer, M. B., Lilienfeld, S. O., Neumann, C. S. & Newman, J. P. (2010). Concurrent validity of the Psychopathic Personality Inventory with offender and community samples. *Assessment*, *17*(1), 3–15.

Marion, B., Sellbom, M., Salekin, R., Toomey, J., Kucharski, L. & Duncan, S. (2012). An examination of the association between psychopathy and dissimulation using the MMPI-2-RF validity scales. *Law and human behavior*, *37(4)*, 219.

Marshall, M. B., De Fruyt, F., Rolland, J.-P. & Bagby, R. M. (2005). Socially desirable responding and the factorial stability of the NEO PI-R. *Psychological assessment*, *17*(3), 379.

Marsland, S. (2011). *Machine Learning: An Algorithmic Perspective*. Boca Raton, FL, USA: Chapman e Hall/CRC.

McDaniel, M. A. & Timm, H. (1990). Lying takes time: Predicting deception in biodata using response latency. In *98th Annual Convention of the American Psychological Association*, Boston.

McFarland, L. A. & Ryan, A. M. (2006). Toward an Integrated Model of Applicant Faking Behavior. *Journal of Applied Social Psychology*, *36*(4), 979–1016.

Meijer, R. & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied psychological measurement, 25(2)*, 107–135.

Millon, T., Simonsen, E. & Birket-Smith, M. (1998). Historical conceptions of psychopathy in the United States and Europe. *Psychopathy: Antisocial, criminal, and violent behavior*, 3–31.

Mitchell, T. M. (1997). *Machine Learning* (1ª ed.). New York, NY, USA: McGraw-Hill, Inc.

Morey, L. (1996). *An interpretive guide to the personality assessment inventory (PAI).* Tampa, FL, USA: Psychological Assessment Resources.

Mueller-Hanson, R. A., Heggestad, E. D., Thornton, G. et al. (2006). Individual differences in impression management: An exploration of the psychological processes underlying faking. *Psychology Science*, *48*(3), 288.

Natali, E. (2008). *La costruzione di una batteria psico-attitudinale multipla (BPM)* (Tesi di dottorato, Dipartimento di Psicologia Generale, Università degli studi di Pdova).

Neal, T. M. & Sellbom, M. (2012). Examining the factor structure of the hare self-report psychopathy scale. *Journal of personality assessment*, *94*(3), 244–253.

Nelson, N., Hoelzle, J., Sweet, J., Arbisi, P. & Demakis, G. (2010). Updated meta-analysis of the MMPI-2 symptom validity scale (FBS): Verified utility in forensic practice. *The Clinical Neuropsychologist*, *24*(4), 701–724.

Neumann, C. S., Hare, R. D. & Pardini, D. A. (2015). Antisociality and the construct of psychopathy: Data from across the globe. *Journal of personality*, *83*(6), 678–692.

Ones, D., Viswesvaran, C. & Reiss, A. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, *81*(6), 660–679.

Ones, D. S. & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human performance*, *11*(2-3), 245–269.

Patrick, C. (2006). Back to the future: Cleckley as a guide to the next generation. Christopher Patrick, Handbook of Psychopathy. New York: Guilford Press.

Paulhus, D. L. (1991). Measurement and Control of Response Bias. In J. P. Robinson, P. R. Shaver & L. S. Wrightsman (Cur.), *Measures of Personality and Social Psychological Attitudes* (pp. 17–59).

Paulhus, D. L., Harms, P. D., Bruce, M. N. & Lysy, D. C. (2003). The over-claiming technique: Measuring self-enhancement independent of ability. *Journal of personality and social psychology*, *84*(4), 890.

Paulhus, D. (1988). Balanced inventory of desirable responding (BIDR). *Acceptance and Commitment Therapy. Measures Package*, *41*, 79586–7.

Paulhus, D. & Neumann, C. (2013). Manual for the Hare self-report psychopathy scale. *Toronto, ON, Canada: Multi-Health Systems*.

Pauly, O. (2012). *Random Forests for Medical Applications* (Tesi di dottorato, Technische Universität München).

Piedmont, R. L., McCrae, R. R., Riemann, R. & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology*, *78*(3), 582.

Pierson, A., Rosenfeld, B., Green, D. & Belfi, B. (2011). Investigating the relationship between antisocial personality disorder and malingering. *Criminal Justice and Behavior*, *38*, 146–156.

Porter, S., ten Brinke, L. & Wallace, B. (2012). Secrets and lies: Involuntary leakage in deceptive facial expressions as a function of emotional intensity. *Journal of Nonverbal Behavior*, *36*, 23–37.

Poythress, N. G., Edens, J. F. & Watkins, M. M. (2001). The relationship between psychopathic personality features and malingering symptoms of major mental illness. *Law and Human Behavior*, *25*(6), 567–582.

Rogers, R., Gillis, J., Dickens, S. & Bagby, R. (1991). Standardized assessment of malingering: Validation of the structured interview of reported symptoms. *Psychological Assessment*, *3*, 89–96.

Rogers, R., Sewell, K., Martin, M. & Vitacco, M. (2003). Detection of feigned mental disorders: A meta-analysis of the MMPI-2 and malingering. *Assessment*, *10*(2), 160–177.

Rosse, J. G., Stecher, M. D., Miller, J. L. & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology*, *83*(4), 634.

Sackett, P. R., Schmitt, N., Ellingson, J. E. & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist*, *56*(4), 302.

Salekin, R. T., Rogers, R. & Sewell, K. W. (1996). A review and meta-analysis of the Psychopathy Checklist and Psychopathy Checklist-Revised: Predictive validity of dangerousness. *Clinical psychology: Science and practice*, *3*(3), 203–215.

Schmit, M. J. & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology*, *78*(6), 966.

Schneider, K. (1933). Psychopatische Persönlichkeiten. *DMW-Deutsche Medizinische Wochenschrift*, *59*(30), 1156–1160.

Sellbom, M., Cooke, D. & Shou, Y. (2019). Development and Initial Validation of the Comprehensive Assessment of Psychopathic Personality-Self-Report (CAPP-SR). *Psychological Assessment*, *7*(31), 878–894.

Skeem, J. L. & Cooke, D. J. (2010). Is criminal behavior a central component of psychopathy? Conceptual directions for resolving the debate. *Psychological assessment*, *22*(2), 433.

Smialowski, P., Frishman, D. & Kramer, S. (2009). Pitfalls of supervised feature selection. *Bioinformatics*, *26*(3), 440–443.

Smith, D. B. & Ellingson, J. E. (2002). Substance versus style: A new look at social desirability in motivating contexts. *Journal of Applied Psychology*, *87*(2), 211.

Smith, D. & McDaniel, M. (2012). Questioning old assumptions: Faking and the personality-performance relationship. In M. Ziegler, C. MacCann & R. Roberts (Cur.), *New perspectives on faking in personality assessment* (Cap. 4, pp. 53–70). New York, NY, US: Oxford University Press.

Smith, G. P. & Burger, G. K. (1997). Detection of malingering: validation of the Structured Inventory of Malingered Symptomatology (SIMS). *Journal of the American Academy of Psychiatry and the Law Online*, *25*(2), 183–189.

Snell, A. F., Sydell, E. J. & Lueke, S. B. (1999). Towards a theory of applicant faking: Integrating studies of deception. *Human Resource Management Review*, *9*(2), 219–242.

Sokolova, M., Japkowicz, N. & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence* (pp. 1015–1021). Springer.

Sokolova, M. & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, *45*(4), 427–437.

Stark, S., Chernyshenko, O. S., Chan, K.-Y., Lee, W. C. & Drasgow, F. (2001). Effects of the testing situation on item responding: Cause for concern. *Journal of Applied psychology*, *86*(5), 943.

Strickland, C. M., Drislane, L. E., Lucy, M., Krueger, R. F. & Patrick, C. J. (2013). Characterizing psychopathy using DSM-5 personality traits. *Assessment*, *20*(3), 327–338.

Topping, G. D. & O'Gorman, J. (1997). Effects of faking set on validity of the NEO-FFI. *Personality and Individual Differences*, *23*(1), 117–124.

Tourangeau, R. & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological bulletin*, *103*(3), 299.

van Dongen, J. D., Drislane, L. E., Nijman, H., Soe-Agnie, S. E. & van Marle, H. J. (2017). Further evidence for reliability and validity of the triarchic psychopathy measure in a forensic sample and a community sample. *Journal of psychopathology and behavioral assessment*, *39*(1), 58–66.

Vasilopoulos, N. L., Reilly, R. R. & Leaman, J. A. (2000). The influence of job familiarity and impression management on self-report measure scale scores and response latencies. *Journal of Applied Psychology*, *85*(1), 50.

Viswesvaran, C. & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and psychological measurement*, *59*(2), 197–210.

Vrij, A., Edward, K. & Bull, R. (2001). People's insight into their own behaviour and speech content while lying. *British Journal of Psychology*, *92*(2), 373–389.

Vroom, V. H. (1964). *Work and motivation*. New York, NY, USA: John Wiley & Sons.

Weiner, J. A. & Gibson, W. M. (2000). Practical effects of faking on job applicant attitude test scores. In *15th annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA*.

White, L. A., Young, M. C. & Rumsey, M. G. (2001). ABLE implementation issues and related research. In J. P. Campbell & D. J. Knapp (Cur.), *Exploring the limits in personnel selection and classification* (pp. 525–558). Lawrence Erlbaum Associates Publishers.

Woodworth, R. S. (1920). *Personal data sheet*. Chicago, IL, USA: Stoelting.

Zickar, M. J., Gibby, R. E. & Robie, C. (2004). Uncovering Faking Samples in Applicant, Incumbent, and Experimental Data Sets: An Application of Mixed-Model Item Response Theory. *Organizational Research Methods*, *7*(2), 168–190.

Zickar, M. J. & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied psychological measurement*, *20*(1), 71–87.

Zickar, M. & Sliter, K. (2012). Searching for unicorns: Item response theory-based solutions to the faking problem. In M. Ziegler, C. MacCann & R. Roberts (Cur.), *New perspectives on faking in personality assessment* (Cap. 8, pp. 113–130). New York, NY, US: Oxford University Press.

Ziegler, M., MacCann, C. & Roberts, R. (2012). *New Perspectives on Faking in Personality Assessment*. New York, NY, USA: Oxford University Press.

Zuckerman, M., DePaulo, B. M. & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. *Advances in experimental social psychology*, *14*, 1–59.