

## **L'utilizzo del *machine learning* per la profilazione statistica dei centri per l'impiego**

**di Maria Cristina Maurizio**

Nel contesto dei Sistemi pubblici per l'impiego, la gestione efficiente delle risorse umane e finanziarie gioca un ruolo fondamentale nel determinare l'efficacia e la reattività delle politiche per l'occupazione. Per questo motivo, i centri per l'impiego si sono da sempre confrontati con la sfida di classificare in modo efficace le persone in base alla loro difficoltà di reinserimento lavorativo. Tale classificazione, nota come profilazione (o in inglese *profiling*), serve a un duplice scopo: aiutare le persone a trovare nuove opportunità di lavoro e migliorare l'accesso alle politiche attive del lavoro.

Già negli anni Novanta, alcuni Paesi come Australia, Canada e Stati Uniti avevano introdotto metodi e strumenti per identificare i soggetti a rischio di disoccupazione di lungo termine (OECD, 1998). Tuttavia, solo a partire dagli anni 2000, la profilazione ha guadagnato notevole popolarità anche in Europa, e molti Paesi hanno iniziato a sviluppare procedure più precise nei propri Sistemi pubblici per l'impiego.

Nella letteratura si possono distinguere tre principali categorie di profilazione:

- la profilazione basata su regole: metodo in cui si utilizzano criteri amministrativi di ammissibilità e soglie per classificare i disoccupati in base a determinate caratteristiche (come l'età, la durata del periodo di disoccupazione, ecc.);
- la profilazione basata esclusivamente sull'operatore: in questo caso, invece, i centri per l'impiego hanno piena discrezionalità nella classificazione dei soggetti che si rivolgono a loro;
- la profilazione statistica: quest'ultimo approccio si basa su metodi di analisi statistica che sfruttano correlazioni o schemi nei dati per classificare i soggetti in base al loro rischio di disoccupazione.

Concentrandosi in particolar modo sulla profilazione statistica, un evidente vantaggio di questo tipo di procedura è la possibilità di effettuare previsioni che tengano conto di un numero maggiore di informazioni rispetto ai semplici metodi basati su regole, esplorando più a fondo la complessità dei meccanismi della disoccupazione. Inoltre, essa è caratterizzata da una minore discrezionalità rispetto al metodo basato solamente sugli operatori, che, in alcuni casi, potrebbero essere portati a concentrarsi maggiormente sui soggetti più facilmente collocabili. Grazie all'impiego dell'analisi statistica, la profilazione può consentire agli operatori di ottenere una comprensione più approfondita delle dinamiche occupazionali, identificare i fattori associati ai gruppi a rischio e migliorare l'utilizzo delle risorse.

Negli ultimi anni, a questa diffusione della profilazione statistica, si è aggiunta anche una trasformazione metodologica: la crescente disponibilità di big data e l'aumento delle capacità computazionali hanno reso sempre più realizzabile l'uso di tecniche di *machine learning* nei Sistemi pubblici per l'impiego. Per *machine learning* si intende quella branca dell'intelligenza artificiale (IA) che comprende una famiglia di algoritmi e tecniche

matematiche progettate per permettere ai computer di apprendere caratteristiche e meccanismi presenti nei dati e, tramite questa conoscenza, generare previsioni. L'idea alla base è quella di sfruttare questa capacità dei sistemi di IA di individuare schemi e usarli per fare previsioni riguardo ad osservazioni non ancora acquisite. Rispetto ai modelli tradizionali gli approcci di *machine learning* offrono diversi vantaggi, fra cui la capacità di gestire un più ampio numero di caratteristiche predittive con una maggiore flessibilità. Questo li rende capaci di scoprire relazioni complesse tra caratteristiche e l'obiettivo della previsione.

Molti settori applicano già metodi di *machine learning* per scopi analitici, ma il suo impiego a supporto delle decisioni operative quotidiane è ancora limitato. Nonostante questo, alcuni Paesi hanno già iniziato a utilizzare queste tecniche nei loro Sistemi pubblici per l'impiego come, ad esempio, la regione belga delle Fiandre.

Una delle principali preoccupazioni che invita a riflettere sull'uso di questi algoritmi in contesti sociali, riguarda il rischio di iniquità e discriminazione. Un possibile caso di discriminazione potrebbe verificarsi qualora un gruppo di soggetti venisse considerato ad alto rischio di disoccupazione semplicemente a causa di una specifica caratteristica che possiede. Infatti, per costruzione, gli algoritmi imparano dai dati a loro mostrati. Tuttavia, è proprio in questi che spesso si riflettono le disuguaglianze già presenti nella società. Si pensi, per esempio, ad un mercato caratterizzato da una forte predominanza maschile: in questa situazione l'algoritmo potrebbe finire per penalizzare gli individui di sesso femminile, indipendentemente dalle altre caratteristiche, a causa della loro minore rappresentanza. Di conseguenza, un sistema automatizzato finirebbe per ripetere o amplificare queste disparità, in particolare trattando in modo diverso persone con caratteristiche simili solo perché appartengono a gruppi differenti (per esempio per genere, età o provenienza). Questo rende la loro applicazione in ambiti come il lavoro, la formazione o le politiche sociali particolarmente delicata: è necessario garantire che le decisioni restino giuste, trasparenti e controllabili, evitando che l'efficienza tecnica si traduca in nuove forme di ingiustizia. Rimane comunque evidente che, anche dopo aver informato gli operatori sulle criticità e analizzato i risultati, le discriminazioni continuano a essere presenti nel mercato del lavoro e non possono essere ignorate. È dunque necessario tenerne conto nella progettazione di politiche e strumenti volti a favorire un inserimento lavorativo soddisfacente nel minor tempo possibile. La presenza di tali discriminazioni, quindi, deve essere riconosciuta, quantificata e affrontata a livello sistemico, al fine di ridurla quanto più possibile.

Un'ulteriore criticità riguarda il fatto che i metodi *machine learning* sono spesso considerati "scatole nere", non solo per il pubblico in generale, a cui appaiono come meccanismi oscuri, ma anche per i ricercatori più esperti, poiché la loro flessibilità nello stimare le relazioni fra caratteristiche e l'oggetto della previsione, rende complessa l'interpretazione dei risultati. A tal fine, negli ultimi anni, sono stati sviluppati diversi strumenti per "aprire" queste "scatole nere" e garantire maggiore trasparenza e responsabilità, come le SHapley Additive exPlanations (SHAP) e le Local Interpretable Model-Agnostic Explanations (LIME). Questi strumenti permettono non solo di identificare le variabili più rilevanti nel modello, ma anche di comprendere, per ciascun soggetto, se tali fattori contribuiscano ad aumentare o

diminuire la probabilità dell'esito previsto. Questi progressi sono fondamentali per promuovere l'uso dei metodi machine learning nelle politiche pubbliche, poiché nell'Unione europea il Regolamento generale sulla protezione dei dati (GDPR) del 2016 impone che i sistemi decisionali automatizzati siano spiegabili alle persone interessate.

In Italia, la prima esperienza con l'uso del profiling statistico nella fornitura di servizi per il lavoro a livello nazionale si è verificata nel 2015, durante l'attuazione del Programma Garanzia Giovani, con l'obiettivo di valutare il livello di occupabilità dei giovani *NEET* (*Not in Education, Employment, or Training* – cioè, non impegnati in percorsi di istruzione, lavoro o formazione). In seguito, l'estensione di questo approccio a tutti i beneficiari dei servizi per il lavoro è avvenuta con la delibera n. 6/2016 del Consiglio di amministrazione dell'Agenzia Nazionale per le Politiche Attive del Lavoro (ANPAL), accompagnata dall'introduzione della Dichiarazione di Immediata Disponibilità al lavoro (DID) nel 2017. La DID è costituita da un questionario principalmente volto a raccogliere informazioni di tipo anagrafico e la sua firma sancisce il momento dell'inizio ufficiale dello stato di disoccupazione. In questo contesto, l'algoritmo alla base della profilazione statistica era costruito con un modello più semplice, utilizzando i dati dell'Indagine Continua sulle Forze di Lavoro dell'Istituto Nazionale di Statistica (ISTAT). Tale tecnica statistica fornisce coefficienti relativi alle caratteristiche individuali (sesso, età, livello di istruzione, esperienze lavorative precedenti, domicilio, ecc.) che indicano l'associazione con la probabilità di rimanere disoccupati dopo 12 mesi. Questi coefficienti vengono poi applicati ai dati con le caratteristiche individuali raccolti tramite la DID online per stimare la probabilità di disoccupazione.

Il successivo cambiamento significativo si è verificato nel novembre 2021, quando, in linea con gli obiettivi del Piano Nazionale di Ripresa e Resilienza (PNRR), è stato adottato il Programma Garanzia di Occupabilità dei Lavoratori (GOL). Il programma mirava a riqualificare l'offerta di politiche attive da parte dei servizi pubblici per l'impiego, progettando percorsi personalizzati di inserimento e reinserimento lavorativo. Dalla prospettiva della profilazione statistica, esso ha modificato la base di dati su cui veniva allenato il modello, passando a un database più ricco, formato dall'archivio amministrativo delle Comunicazioni Obbligatorie (riguardanti assunzioni, cessazioni, trasformazioni e proroghe) inviate agli enti territoriali e nazionali competenti dai datori di lavoro. Questo ha permesso l'utilizzo di molte più informazioni legate alla storia lavorativa, che hanno notevolmente aumentato l'efficacia del modello. Nonostante il miglioramento messo in atto dal programma GOL, rimangono alcuni accorgimenti da considerare nell'analizzare il modello di profilazione sopra menzionato. Un punto critico riguarda la mancanza di una stima periodica del modello. Infatti, la profilazione statistica attualmente è stata costruita utilizzando dati del 2018-2019 e, di conseguenza, i coefficienti stimati non sono stati aggiornati fino a oggi. Un aggiornamento del modello consentirebbe probabilmente di migliorarne le prestazioni, tenendo conto delle nuove tendenze del mercato del lavoro, ottenendo previsioni più accurate e una classificazione più efficace. Inoltre, il modello utilizzato è stimato a livello nazionale, il che consente solo un ruolo limitato alla dimensione

geografica, nonostante essa costituisca sicuramente un rilevante fattore di eterogeneità del mercato del lavoro italiano.

Da un'analisi condotta sui dati della Toscana, utilizzando tecniche sia tradizionali sia più innovative, è emerso che queste ultime forniscono previsioni più precise nel distinguere i soggetti che troveranno lavoro da quelli che risulteranno disoccupati. Tuttavia, poiché un aspetto fondamentale per l'applicabilità degli algoritmi di *machine learning* in contesti pubblici è la necessità di trasparenza e interpretabilità dei modelli predittivi, si è scelto di utilizzare la metodologia SHAP (*SHapley Additive exPlanations*). Questo approccio permette di capire come ogni variabile contribuisca alle previsioni del modello, sia in positivo sia in negativo. I risultati di questa analisi mostrano che, coerentemente con la letteratura precedente, le variabili relative alla storia lavorativa passata, e nello specifico all'ultima occupazione, sono molto rilevanti. In particolare, la durata e la tipologia dell'ultimo contratto, il numero di datori di lavoro negli ultimi anni e il tempo trascorso dall'ultima occupazione influenzano fortemente la probabilità di trovare un nuovo impiego. Le persone con storie lavorative più frammentate risultano avere maggiore probabilità di essere occupate a distanza di un anno dalla sottoscrizione della DID, così come chi ha avuto esperienze recenti in settori come il turismo o la ristorazione. Questo particolare risultato richiede necessariamente un approfondimento sulla stabilità e la qualità delle occupazioni trovate dagli individui con queste specifiche caratteristiche. La maggiore probabilità di reimpiego osservata, infatti, potrebbe celare una situazione patologica di spostamenti frequenti tra lavori di breve durata, caratterizzati da una maggiore instabilità. Un possibile esempio potrebbe essere un/a cameriere/a che viene impiegato/a frequentemente ma per pochi giorni, in base alle necessità di un locale, risultando quindi occupato per il sistema, ma soffrendo una situazione lavorativa fortemente instabile. Tale dinamica riflette una condizione di fragilità che meriterebbe certamente di essere non solo ulteriormente analizzata, ma anche tenuta in considerazione dagli operatori dei centri per l'impiego, in quanto potrebbe costituire uno dei motivi principali alla base del contatto con il servizio.

In conclusione, questi risultati dimostrano che un uso accorto delle tecniche di *machine learning*, insieme a una gestione consapevole degli squilibri nei dati, può migliorare sensibilmente la capacità predittiva dei Sistemi pubblici per l'impiego. Inoltre, le analisi condotte hanno permesso di spiegare i risultati, "aprendo la scatola nera" per capire quali caratteristiche incidano di più sul ritorno al lavoro. Tra queste vengono evidenziate in particolar modo la durata e la tipologia dell'ultimo contratto, la frammentazione della carriera e il settore economico di provenienza. Tuttavia, non è da trascurarsi la necessità di utilizzare un approccio quanto più etico e trasparente, informando gli operatori di prestare sempre attenzione a potenziali discriminazioni esistenti nei dati e replicate dal modello, e garantendo sempre la comprensibilità dei risultati agli operatori e agli utenti.