Osservatorio del Mercato del Lavoro Puglia

NOTA METODOLOGICA SUL FORECASTING REALIZZATO

Approccio e strumenti per la previsione occupazionale
delle filiere della S3 in Puglia

Finanziato dall'Unione europea NextGenerationEU
Italiadomani
REGIONE PUGLIA
ARTI PUGLIA
intellera Part of Accenture Group

Indice

1 INTRODUZIONE

La presente Nota metodologica documenta il percorso di analisi e sperimentazione condotto nell'ambito dell'Osservatorio del Mercato del Lavoro in Puglia al fine di sviluppare, valutare e contestualizzare modelli previsionali applicati al mercato del lavoro regionale, con un focus specifico sulle filiere Tessile–Abbigliamento–Calzaturiero (TAC), Automotive e Industrie Culturali e Creative (ICC).

L'iniziativa si colloca in un quadro di crescente attenzione, a livello europeo e nazionale, verso strumenti di Labour Market Intelligence in grado di supportare le politiche pubbliche con informazioni tempestive, granulari e, ove possibile, previsionali. In particolare, la programmazione di interventi in materia di formazione, orientamento, politiche attive del lavoro e sostegno alla transizione industriale richiede:

- una lettura robusta delle dinamiche occupazionali recenti, disaggregate per settore, territorio e profilo professionale;
- la capacità di anticipare, su un orizzonte di breve-medio periodo, l'evoluzione della domanda di lavoro, almeno in termini di segnali di crescita, contrazione o trasformazione dei fabbisogni;
- una progressiva integrazione tra fonti amministrative, dati statistici e nuove fonti digitali (job posting online, banche dati settoriali, ecc.), sfruttando le opportunità offerte dalle tecniche di intelligenza artificiale (IA) e machine learning (ML).

Tradizionalmente, il forecasting del mercato del lavoro è stato affidato a sistemi macroeconomici complessi, come quelli sviluppati da INAPP (1) (2), a livello nazionale, e da Cedefop (3), a livello europeo. Tali sistemi, basati su modelli macroeconometrici multisettoriali, garantiscono un'elevata coerenza con la contabilità nazionale e con gli scenari macro-ufficiali, ma presentano limiti di granularità e di reattività rispetto a segnali emergenti "dal basso".

Parallelamente, negli ultimi anni si è consolidata una vasta letteratura sull'uso di metodi di ML e, più in generale, di tecniche data-driven per la previsione di grandezze economiche e del mercato del lavoro. In questo filone, i job posting online sono stati spesso utilizzati come proxy tempestivi della domanda di lavoro, soprattutto per specifici comparti e profili professionali. Tali approcci, pur meno strutturati sul piano macroeconomico, consentono una maggiore adattabilità e una migliore capacità di sfruttare l'informazione contenuta in grandi basi dati micro.

Il presente lavoro si colloca all'intersezione di questi due approcci: da un lato, riconosce il ruolo imprescindibile dei modelli "classici" (INAPP, Cedefop) come quadro di riferimento per la comprensione delle tendenze strutturali e per la coerenza con le politiche nazionali ed europee; dall'altro, esplora in modo pragmatico la possibilità di utilizzare tecniche IA/ML, applicate a dati di job posting online, per costruire strumenti previsionali a breve termine specificamente tarati sul contesto pugliese e sulle filiere di interesse.

Più in dettaglio, la Nota persegue i seguenti obiettivi:

1. documentare le fonti dati utilizzate, con particolare attenzione a LMI (Labour Market Intelligence) (4) e ai job posting Lightcast (5), esplicitando le scelte di integrazione, i vincoli di qualità e le implicazioni per il disegno dei modelli (Capitolo 2);
2. contestualizzare il lavoro rispetto ai modelli previsionali consolidati, presentando in modo sintetico ma rigoroso la struttura dei sistemi di previsione INAPP e Cedefop, nonché i loro punti di forza e di debolezza dal punto di vista di un ente regionale (Capitolo 3);
3. descrivere la metodologia IA/ML adottata, distinguendo tra:
    - soluzioni ipoteticamente considerate ma limitate dai vincoli informativi;
    - soluzioni effettivamente implementate nel progetto (Capitolo 4);
4. illustrare il processo di validazione e confronto tra modelli, basato su metriche standard di forecasting e su procedure di backtesting coerenti con la natura temporale dei dati (Capitolo 5);
5. esplicitare i limiti e le assunzioni del sistema previsionale, in un'ottica di trasparenza metodologica e di corretta interpretazione dei risultati (Capitolo 6);
6. proporre linee di sviluppo future, sia sul piano dei dati (ad esempio integrazione con le Comunicazioni Obbligatorie – COB, arricchimento informativo su competenze), sia sul piano metodologico (ad esempio forecasting gerarchico, modelli multivariati, scenari probabilistici), con l'intento di delineare un percorso graduale di rafforzamento della capacità previsiva regionale (Capitolo 7).

La logica di fondo è quella di interpretare la sperimentazione IA/ML non come un esercizio isolato, ma come un tassello di un più ampio ecosistema di Labour Market Intelligence regionale, integrato con le informazioni e le proiezioni prodotte a livello nazionale ed europeo. Il presente documento è da intendersi come una Nota metodologica "di frontiera": sufficientemente solida e documentata per essere riutilizzabile, ma al tempo stesso esplicitamente aperta a revisioni e miglioramenti man mano che nuove basi dati e nuove esigenze di policy verranno alla luce.

2 DATI E FONTI DISPONIBILI

L'analisi integra fonti istituzionali e dati digitali, con l'obiettivo di bilanciare robustezza statistica, tempestività informativa e aderenza al contesto regionale. La scelta delle fonti non risponde a una logica di esaustività, ma a quella di "massimo utilizzo realistico" delle informazioni attualmente accessibili a livello regionale.

Le fonti principali sono:

- Labour Market Intelligence (LMI), integrata nel sito web dell'Osservatorio del Mercato del Lavoro sviluppato da ARTI;
- job posting online forniti da Lightcast;
- classificazioni e tassonomie: CP2021, ESCO, Atlante del Lavoro;
- informazioni di contesto sulle filiere TAC, Automotive e ICC derivanti da precedenti studi regionali.

2.1 Labour Market Intelligence (LMI)

LMI (4) costituisce la fonte istituzionale di riferimento sui flussi del mercato del lavoro regionale. La piattaforma permette di osservare, con cadenza trimestrale, i flussi di:

- attivazioni e cessazioni di rapporti di lavoro;
- trasformazioni contrattuali;
- distribuzione per settore economico, territorio e caratteristiche dei lavoratori.

Nel presente lavoro:

- è stato utilizzato l'ultimo biennio per cui il dato era disponibile (2023–2024);
- l'unità temporale di osservazione è il trimestre;
- le informazioni professionali sono state ricondotte alla classificazione CP2021, a partire dal lavoro pregresso di mappatura delle filiere.

LMI è stato impiegato con tre funzioni principali:

1. definizione del perimetro di analisi. Le informazioni sulle filiere regionali, sulle specializzazioni produttive e sulle professioni più rappresentate hanno guidato la selezione delle serie da analizzare con i modelli previsionali;
2. contestualizzazione dei risultati. Le tendenze osservate nelle attivazioni contrattuali (crescita, stabilità, stagionalità) sono state utilizzate come riferimento qualitativo per interpretare le previsioni basate sui job posting;
3. verifica preliminare di fattibilità modellistica. La limitata profondità storica (due anni) e la frequenza trimestrale hanno suggerito di non forzare l'uso di LMI come base diretta per modelli di serie temporali avanzati, confermando la necessità di utilizzare, nella fase implementativa, basi dati con maggiore densità temporale.

2.2 Annunci di lavoro online (Lightcast)

La seconda fonte è rappresentata dai job posting online, forniti da Lightcast (5). Tali dati presentano alcune caratteristiche particolarmente rilevanti dal punto di vista previsivo:

- tempestività: i job posting riflettono in modo rapido le variazioni della domanda di lavoro espressa formalmente dalle imprese sui canali digitali;
- granularità: gli annunci sono etichettati con informazioni dettagliate su professioni, competenze, settore, localizzazione, tipologia contrattuale, ecc;
- frequenza: la disponibilità di dati mensili consente di identificare pattern di breve periodo e di stimare modelli con un numero di osservazioni superiore a quello delle serie trimestrali LMI.

Nel presente progetto:

- sono stati estratti i job posting relativi alla Puglia;
- il periodo di osservazione va da gennaio 2021 ad agosto 2025;
- la granularità temporale utilizzata per il forecasting è il mese;
- le professioni sono classificate secondo ESCO;
- il perimetro settoriale è stato definito selezionando insiemi di codici ESCO coerenti con le filiere TAC, Automotive e ICC (Appendice).

2.3 Considerazioni su qualità, granularità e limiti delle fonti

L'integrazione tra LMI e job posting Lightcast offre un quadro informativo più ricco di quello ottenibile con una sola fonte, ma introduce anche una serie di questioni metodologiche che è opportuno esplicitare:

- diverse nature statistiche. LMI è una fonte amministrativa ufficiale; i job posting sono una fonte "osservazionale" di natura digitale. Ne derivano:
    - differenze nei meccanismi di produzione del dato;
    - differenti coperture per settore e dimensione d'impresa;
    - possibili bias nell'uso dei canali digitali di reclutamento.
- disallineamento di classificazioni. Le categorie professionali e settoriali adottate dalle diverse fonti non coincidono perfettamente (CP2021 vs ESCO vs ATECO). La costruzione di mapping tra queste classificazioni è un passaggio metodologicamente delicato, che può introdurre margini di errore, soprattutto per le categorie meno frequenti o ambigue;
- copertura parziale della domanda di lavoro. I job posting digitali non rappresentano l'insieme della domanda di lavoro, ma solo la parte veicolata tramite piattaforme online. Alcuni settori – come, ad esempio, alcuni segmenti del TAC, possono fare un uso più limitato di tali canali, facendo prevalere forme di reclutamento informale o relazionale;
- profondità storica limitata (LMI) e intermedia (job posting). Il biennio 2023–2024 di LMI non è sufficiente per elaborare modelli econometrici su base trimestrale con orizzonti previsivi di

medio periodo. I job posting, disponibili dal 2021/2022, offrono maggiore profondità, ma restano comunque su un orizzonte temporale relativamente breve rispetto alla dinamica strutturale del mercato del lavoro;

- assenza di informazione sugli esiti occupazionali. Le fonti considerate descrivono l'offerta di posti (job posting) e i flussi di attivazioni/cessazioni. Non forniscono, nella configurazione qui utilizzata, informazioni sulla durata effettiva dei contratti, sulle transizioni nel ciclo di vita lavorativa o sulla stabilizzazione. **Il forecasting prodotto è dunque incentrato sulla dinamica delle opportunità (annunci) più che sull'occupazione effettivamente realizzata**.

Questi vincoli hanno guidato le scelte metodologiche successive: i dati LMI sono stati impiegati a fini descrittivi e per la definizione del perimetro, mentre i modelli previsionali sono stati addestrati sulle serie mensili di job posting, dove la densità temporale consente applicazioni più rigorose delle tecniche di forecasting.

3 MODELLI PREVISIONALI CLASSICI: INAPP E CEDEFOP

In questa sezione si presentano, a fini comparativi, i principali elementi metodologici dei modelli previsionali sviluppati da INAPP a livello nazionale e da Cedefop a livello europeo. L'obiettivo non è replicare in dettaglio tali modelli – che sono altamente complessi – ma evidenziarne l'architettura generale, i punti di forza e i limiti, così da comprendere in quale misura e con quali avvertenze un esercizio di forecasting regionale basato su IA/ML può dialogare con essi.

3.1 Il modello INAPP

INAPP, spesso in collaborazione con il Ministero del Lavoro e altri soggetti istituzionali, realizza periodicamente esercizi previsionali di medio periodo (tipicamente 3–5 anni) sulla domanda e offerta di lavoro in Italia. Queste previsioni si basano su un sistema integrato di modelli macroeconomici, multisettoriali e regionali, con disaggregazione finale per qualifiche e professioni.

3.1.1 Architettura del modello

L'architettura del modello può essere descritta per blocchi.

a) Modello macroeconomico trimestrale

Il primo modulo è un modello macroeconomico che proietta le principali grandezze (PIL, consumi, investimenti, inflazione, export, salari, variabili del settore pubblico e finanziario). Le equazioni dei diversi blocchi sono stimate in un framework di tipo VECM (Vector Error Correction Model), che consente di rappresentare congiuntamente:

- relazioni di lungo periodo tra le variabili (cointegrazione);
- dinamiche di breve periodo e meccanismi di aggiustamento verso l'equilibrio.

Da questo modello si derivano scenari macro coerenti, inclusi i livelli aggregati di occupazione e le dinamiche del mercato del lavoro. Le proiezioni costituiscono gli input di vincolo per gli stadi successivi.

b) Modello multisettoriale input–output

A partire dallo scenario macro complessivo, un modulo multisettoriale disaggrega le previsioni per settore di attività economica, utilizzando:

- tavole input–output e tavole delle risorse e degli impieghi;
- relazioni econometriche specifiche per domanda finale, valore aggiunto, prezzi e mercato del lavoro.

In tal modo si ottiene:

- la distribuzione prevista del valore aggiunto e dell'occupazione tra i diversi settori (agricoltura, manifattura per branche, costruzioni, servizi);
- la possibilità di analizzare gli effetti intersettoriali (traino o freno) fra comparti, inclusi quelli legati a politiche specifiche come il PNRR.

c) Modello econometrico regionale

Le dinamiche settoriali nazionali vengono poi ripartite a livello regionale mediante modelli panel (settore × regione × tempo) con effetti fissi, che stimano valore aggiunto e occupazione per singola regione.

Per garantire la coerenza tra il totale nazionale e le disaggregazioni regionali e settoriali viene adottata una procedura iterativa di riconciliazione, spesso basata su algoritmi di tipo RAS (scaling delle matrici). Il risultato è un set di previsioni di occupati per settore e regione, coerente con lo scenario macro originario.

d) Modello di domanda di lavoro per qualifica/professione

A valle della disaggregazione regionale, un ulteriore modulo traduce la crescita (o riduzione) degli occupati settoriali in fabbisogni per qualifica e professione, utilizzando distribuzioni storiche delle figure professionali all'interno dei settori e delle aree geografiche.

Un elemento centrale è la distinzione fra:

- domanda aggiuntiva (variazione netta degli occupati in una professione);
- domanda sostitutiva (posti liberati da pensionamenti, mobilità, altri motivi di uscita).

La somma delle due componenti fornisce il fabbisogno totale di assunzioni per ciascuna professione, anche in settori con occupazione stabile o in lieve contrazione.

e) Modello di offerta di lavoro

In parallelo, viene modellata l'offerta di lavoro futura per età, genere, qualifica e titolo di studio, partendo da proiezioni demografiche e tassi di partecipazione al mercato del lavoro. L'incontro tra domanda e offerta previste consente di individuare possibili squilibri quantitativi e qualitativi.

3.1.2 Caratteristiche dell'approccio e punti di forza

Complessivamente, l'approccio INAPP è di tipo top–down:

- dalle variabili macro si discende progressivamente a settori, regioni, professioni;
- ad ogni livello si mantengono i vincoli di coerenza con i totali sovraordinati;
- la conoscenza esperta entra sia nella definizione degli scenari macro, sia nella taratura di alcune elasticità chiave e nella validazione dei risultati.

I punti di forza sono i seguenti:

- coerenza con gli scenari macro ufficiali e con la contabilità nazionale;
- capacità di distinguere esplicitamente tra domanda aggiuntiva e sostitutiva;
- possibilità di analisi strutturali (specializzazioni territoriali, effetti intersettoriali, impatto di politiche pubbliche).

3.1.3 Limiti dal punto di vista regionale

I principali limiti del modello sono, invece:

- aggiornamento tipicamente annuale, con limitata reattività a shock improvvisi;
- affidabilità piena solo a livelli relativamente aggregati (macrosettori, grandi gruppi professionali);
- assenza di una modellizzazione diretta delle competenze (hard, soft, digitali), che restano implicite nelle classificazioni;
- natura prevalentemente deterministica, che non sfrutta in modo diretto grandi moli di dati micro e non cattura facilmente pattern non lineari o emergenti a livello micro.

3.2 Il modello Cedefop

Il sistema di previsione sviluppato da Cedefop (Skills Forecast) fornisce, a livello europeo, una rappresentazione coerente e armonizzata delle tendenze future della domanda e dell'offerta di lavoro in tutti gli Stati membri. Dal punto di vista metodologico, esso può essere considerato un "gemello europeo" – più ampio e comparativo – degli approcci nazionali come quello INAPP, con alcune peculiarità rilevanti.

3.2.1 Architettura del modello

Il modello si articola in tre componenti principali:

a) Modulo macroeconomico ed occupazionale europeo

Basato su modelli come E3ME, che integrano aspetti macroeconomici, energetici e ambientali, il modulo genera proiezioni per:

- PIL e altre grandezze macro per ciascun Paese;
- occupazione per settore economico (classificazioni compatibili con NACE/ATECO);
- principali indicatori del mercato del lavoro (tassi di disoccupazione, partecipazione, ecc.).

b) Moduli di domanda di lavoro per occupazione e qualifica

A valle delle proiezioni settoriali, un complesso sistema di relazioni empiriche e matrici di struttura occupazionale permette di:

- disaggregare l'occupazione prevista per settore in occupazioni (classificazione ISCO);
- tradurre l'occupazione per occupazione in fabbisogni per livello di istruzione (ISCED), stimando quante persone con titoli di studio bassi, medi o alti saranno richieste in ciascun Paese;
- stimare la domanda aggiuntiva (crescita netta di occupazione) e la domanda sostitutiva (posti da rimpiazzare per pensionamenti, mobilità, ecc.), ottenendo così il fabbisogno totale di ingressi futuri.

c) Modulo di offerta di lavoro e di confronto domanda–offerta

Utilizzando proiezioni demografiche, dati sulla partecipazione al mercato del lavoro e informazioni sui sistemi educativi, viene stimata l'offerta futura di lavoro:

- per età, genere e livello di istruzione;
- per status economico (occupati, disoccupati, inattivi).

Il confronto sistematico tra domanda e offerta consente di individuare:

- potenziali carenze (shortages) di determinate qualifiche;
- eccessi di offerta (surplus) in altre aree;
- possibili aree di mismatch strutturale.

3.2.2 Processo di previsione e validazione

Una caratteristica fondamentale del modello Cedefop è la combinazione tra:

- approccio modellistico strutturato, basato su relazioni econometriche e matrici di struttura;
- validazione esperta, affidata a gruppi di esperti nazionali (Institutional Country Experts – ICEs).

In pratica:

- le previsioni prodotte dal modello vengono sottoposte a un processo di revisione qualitativa da parte di esperti nei singoli Paesi;
- gli esperti verificano che gli andamenti previsti siano coerenti con informazioni di contesto recente, piani di politica economica e del lavoro e altri esercizi previsionali nazionali;
- eventuali discrepanze marcate possono portare a revisioni delle ipotesi di scenario, ad affinamenti delle relazioni strutturali o ad aggiustamenti mirati nei parametri del modello.

3.2.3 Tipologie di output

Cedefop rende disponibili, per ciascun Paese, diverse tipologie di output:

- serie storiche e proiezioni di occupati per settore economico, su orizzonti pluriennali;
- serie analoghe per grandi gruppi professionali (ISCO);
- stime del fabbisogno totale di lavoratori per livello di istruzione (ISCED), distinguendo la componente aggiuntiva e quella sostitutiva;
- indicatori sintetici di mismatch tra domanda e offerta di qualifiche;
- analisi di scenario su settori chiave (ad esempio digitale, green, sanità), talvolta accompagnate da approfondimenti tematici.

3.2.4 Rilevanza e limiti per un ente regionale

Dal punto di vista di una regione come la Puglia, il modello Cedefop presenta una duplice valenza:

- **Valore informativo**:
  - offre un quadro comparativo che consente di collocare la Puglia, tramite l'Italia, all'interno del contesto europeo;
  - permette di comprendere se le tendenze osservate (ad esempio crescita di determinate professioni o qualifiche) sono specifiche del territorio o coerenti con una dinamica più generale;
  - supporta scelte di policy che mirano a posizionare il territorio rispetto alle traiettorie di trasformazione europee (transizione verde e digitale).
- **Limiti operativi per l'uso diretto a livello regionale**:
  - le previsioni Cedefop sono prodotte solo a livello nazionale;
  - una loro applicazione diretta alla Puglia richiederebbe ipotesi ulteriori (mantenimento delle quote regionali, regole di ripartizione basate su indicatori di specializzazione produttiva, ecc.);

- tali ipotesi, pur utilizzabili in esercizi esplorativi, introducono margini di errore non trascurabili, soprattutto in un contesto con specificità produttive e demografiche come quello pugliese.

In termini metodologici, il modello Cedefop rappresenta dunque un riferimento essenziale per allineare le analisi regionali alle traiettorie europee e per garantire coerenza concettuale, ma non è uno strumento direttamente utilizzabile per generare previsioni a grana fine su settori e professioni in Puglia. In questi ambiti, il ricorso a dati e modelli specifici, come quelli IA/ML sperimentati in questo progetto, appare maggiormente appropriato.

4 METODOLOGIA E MODELLI IA/ML PROPOSTI

4.1 Obiettivo e approccio

L'attività modellistica ha perseguito un obiettivo operativo chiaro: sviluppare e valutare approcci previsionali data-driven, basati su tecniche di IA/ML, per stimare l'evoluzione a breve termine (6 mesi) del numero di annunci di lavoro digitali nei settori TAC, Automotive e ICC in Puglia.

Il disegno del lavoro ha cercato di conciliare:

- rigore metodologico, applicando tecniche coerenti con le caratteristiche dei dati (serie temporali, lunghezza limitata, presenza di stagionalità);
- pragmatismo, tenendo conto dei vincoli informativi e operativi e concentrandosi su soluzioni implementabili con i dati effettivamente disponibili.

La sequenza seguita è stata:

1. **valutazione di fattibilità su LMI (attivazioni contrattuali).** Sono stati sperimentati modelli di serie temporali (ARIMA/SARIMA, ETS) e modelli di ML (Random Forest, GBM) sulle serie trimestrali di attivazioni per CP2021 nei settori selezionati. La scarsità di osservazioni e l'aggregazione trimestrale hanno reso i risultati poco stabili, confermando l'inadeguatezza di tali serie per un forecasting robusto a livello disaggregato;
2. **passaggio a un livello settoriale più aggregato.** Si è tentato di prevedere il totale degli occupati per settore, sempre su base trimestrale. Anche in questo caso, la combinazione di profondità storica limitata e complessità della dinamica occupazionale ha prodotto previsioni troppo sensibili alle specificazioni del modello e ai singoli shock recenti;
3. **concentrazione su job posting mensili (Lightcast).** Alla luce di tali evidenze, la fase implementativa si è focalizzata sui job posting:
    - serie mensili dal 2021;
    - variabile target: numero di annunci per filiera (TAC, Automotive, ICC);

- orizzonte di previsione: 6 mesi.

Questa scelta è coerente con la letteratura che indica come i job posting offrano, per orizzonti relativamente brevi, un segnale tempestivo della domanda di lavoro. Sul piano metodologico, l'uso di serie mensili consente di applicare in modo più rigoroso sia modelli statistici classici sia tecniche di ML orientate alle serie temporali.

I modelli considerati sono stati:

- ARIMA e SARIMA;
- ETS (Error–Trend–Seasonality);
- Gradient Boosting Machine (GBM e LightGBM);
- Random Forest;
- Ensemble stacking (ETS + ARIMA + GBM).

L'attenzione si è progressivamente concentrata sui modelli che, in letteratura, mostrano migliori prestazioni nel forecasting di serie temporali economiche con struttura non lineare e segnali rumorosi: ETS come benchmark interpretabile, ARIMA come riferimento autoregressivo, GBM/LightGBM come tecniche di ML per catturare pattern complessi.

4.2 Descrizione dei modelli

4.2.1 Gradient Boosting Machine (GBM e LightGBM)

Il Gradient Boosting costruisce un modello forte come somma di molti modelli deboli (alberi decisionali a bassa profondità), addestrati in sequenza per minimizzare una funzione di perdita, ad esempio l'errore quadratico medio. Ad ogni iterazione, il modello successivo impara a correggere gli errori residui del modello corrente.

Da un punto di vista qualitativo, questo approccio è particolarmente adatto quando si sospetta la presenza di relazioni non lineari tra le variabili esplicative (nel nostro caso, i lag della serie) e la variabile target (numero di annunci). Inoltre, consente di modellare interazioni implicite tra lag diversi (combinazioni di pattern di breve termine e stagionali) e offre buone capacità di generalizzazione se accompagnato da adeguate tecniche di regolarizzazione (controllo della profondità degli alberi, learning rate, numero di iterazioni) (6).

LightGBM rappresenta un'implementazione avanzata di Gradient Boosting, progettata per:

- ridurre il tempo di addestramento tramite algoritmi di crescita degli alberi più efficienti;
- gestire in modo efficace dataset con un numero consistente di variabili e di osservazioni;
- includere meccanismi di regularization ed early stopping che aiutano a evitare l'overfitting (7).

Nel progetto, i modelli GBM/LightGBM sono stati formulati come modelli di forecasting univariato, in cui le feature sono lag di 1, 6 e 12 mesi ed eventuali feature derivate dalla data (mese, trimestre) e/o dummy stagionali.

Questa scelta è metodologicamente coerente con l'idea di rappresentare la dipendenza immediata dai valori più recenti (lag 1), pattern stagionali intra-annuali (lag 12) e possibili effetti a medio termine (lag 6), ad esempio cicli produttivi semestrali.

LightGBM è risultato il modello con il miglior compromesso tra:

- accuratezza previsionale (errori ridotti rispetto ad altre tecniche);
- stabilità al variare delle finestre di addestramento e validazione;
- complessità computazionale e gestione operativa.

4.2.2 ARIMA e SARIMA

I modelli ARIMA/SARIMA si basano sull'idea che il valore futuro di una serie possa essere espresso come combinazione lineare di valori passati (componente autoregressiva, AR), errori passati (componente di media mobile, MA) e differenze della serie (componente integrata, I) necessarie per rendere la serie stazionaria (8).

La variante ARIMA estende questa logica includendo componenti stagionali AR, I, MA ad una data periodicità, ad esempio 12 mesi per dati mensili. È un approccio adatto a serie con stagionalità regolari e relativamente stabili, per le quali è ragionevole ipotizzare relazioni lineari tra passato e futuro (8).

Nel contesto del progetto, ARIMA ha svolto principalmente il ruolo di benchmark statistico per valutare quanto della dinamica dei job posting potesse essere spiegata da un modello lineare autoregressivo. Inoltre, è stato utilizzato come componente dell'ensemble, contribuendo con una previsione centrata sulla dipendenza temporale interna.

SARIMA, invece, è stato testato ma non utilizzato operativamente a causa della limitata lunghezza delle serie, della difficoltà a identificare in modo stabile i parametri stagionali e della maggiore complessità, non giustificata dai benefici marginali in termini di accuratezza.

4.2.3 ETS – Error, Trend, Seasonality

Il modello ETS assume che la serie possa essere scomposta in tre componenti: livello, trend e stagionalità, più un termine di errore. Nel caso in esame si è utilizzata una specificazione additiva, in cui il livello cattura il valore medio di base, il trend rappresenta la tendenza di crescita o decrescita e la stagionalità rappresenta pattern ricorrenti con periodicità predefinita.

ETS è apprezzato per la sua interpretabilità, poiché consente di visualizzare e comprendere separatamente le componenti, per la flessibilità nel trattamento di trend e stagionalità e per la robustezza come baseline in molti contesti di previsione. Nel progetto è stato impiegato sia come riferimento per valutare il valore aggiunto dei modelli di machine learning, sia per verificare la presenza e la stabilità di pattern stagionali nei job posting (8).

4.2.4 *Random Forest*

La Random Forest costruisce numerosi alberi decisionali su sotto-campioni bootstrap del dataset e aggrega le previsioni mediandole. Questo metodo è robusto al rumore, capace di catturare relazioni non lineari e relativamente semplice da addestrare, con la possibilità di interpretare l'importanza delle variabili (9).

Nel progetto è stata utilizzata con feature analoghe a quelle di GBM/LightGBM, come lag e variabili temporali, principalmente per confrontare i risultati con quelli ottenuti dai modelli di boosting. Le performance inferiori rispetto a LightGBM ne hanno limitato l'uso alle fasi sperimentali.

4.3 Modello ensemble e "pool of experts"

Per sfruttare la complementarità dei diversi modelli, è stato sviluppato un ensemble di tipo stacking che combina ETS (componenti di livello, trend, stagionalità), ARIMA (dinamica autoregressiva lineare) e GBM (non linearità e interazioni tra lag).

La logica alla base del "pool of experts" è che ciascun modello contribuisca catturando la parte della dinamica che gestisce meglio. L'ensemble è stato costruito stimando separatamente i modelli su una porzione della serie, calcolando le previsioni su una finestra di validazione e pesando ciascun modello in funzione della sua performance.

In teoria, un ensemble ben bilanciato riduce il rischio di errori sistematici dei singoli modelli, stabilizza le previsioni in presenza di cambi strutturali e migliora la robustezza rispetto a shock locali (10) (11).

Nella pratica del progetto, l'ensemble ha fornito risultati generalmente buoni, ma LightGBM – da solo – ha raggiunto performance pari o superiori, con un grado di complessità gestionale minore. Per questa ragione, l'ensemble è stato mantenuto come prototipo metodologico e non come modello operativo principale.

4.4 Approccio gerarchico e riconciliazione tra livelli

La letteratura recente sul forecasting gerarchico suggerisce di modellare congiuntamente le serie a diversi livelli di aggregazione (ad esempio nazionale/regionale/provinciale, settore/sottosettore) e di riconciliare le previsioni per garantire coerenza aggregativa. Nel contesto del progetto, la gerarchia naturale sarebbe composta da:

- livello 0: totale annunci Puglia;
- livello 1: annunci per filiera (TAC, Automotive, ICC);
- livello 2: annunci per provincia e filiera;
- livello 3: annunci per cluster professionali ESCO.

Un sistema gerarchico completo potrebbero adottare approcci bottom–up (somma delle previsioni dei livelli inferiori), top–down (ripartizione della previsione aggregata su livelli inferiori) o metodi di

riconciliazione ottimale che aggiustano congiuntamente le previsioni per minimizzare l'errore complessivo. Nel progetto, per ragioni di numerosità e complessità, questa prospettiva è stata solo abbozzata: sono stati condotti alcuni test tra livello regionale e provinciale, ma non è stato implementato un sistema completo fino al dettaglio dei cluster professionali. Le previsioni operative riguardano principalmente livelli aggregati (Puglia × filiera), dove i volumi sono sufficienti e il problema di coerenza gerarchica è meno critico.

La formalizzazione di un sistema gerarchico strutturato è eventualmente rinviata a fasi successive, in cui l'accesso a basi dati più dettagliate (ad esempio COB) e una maggiore profondità storica renderanno tale approccio più robusto e giustificato.

5 VALIDAZIONE E CONFRONTO

La fase di validazione è cruciale per garantire che i modelli non si limitino a spiegare il passato, ma siano effettivamente in grado di prevedere il futuro con un livello di errore accettabile per le finalità di policy.

5.1 Metodologia di validazione

La validazione è stata condotta mediante backtesting, applicando il modello a finestre temporali storiche e generando previsioni out-of-sample rispetto ai dati successivi. Le stime prodotte sono state confrontate con i valori osservati, al fine di valutare la capacità predittiva del modello sugli orizzonti considerati. In particolare:

- la serie completa di job posting è stata suddivisa in periodo di addestramento e periodo di test/validazione, collocato verso la fine della serie;
- sono stati effettuati esercizi di previsione "a finestra scorrevole" (rolling-origin): il modello viene addestrato fino a un certo mese, prevede i 3 o 6 mesi successivi, confronta le previsioni con i valori effettivamente osservati e ripete il processo spostando in avanti la finestra temporale.

Questo metodo è coerente con la natura temporale dei dati, evita di mescolare passato e futuro nel training, consente di valutare la stabilità delle performance nel tempo e permette di stimare l'errore atteso in condizioni simili all'uso operativo del modello.

Le metriche utilizzate sono:

- **MAE** (Mean Absolute Error), che misura l'errore medio in valore assoluto fra osservati e previsti, fornendo una misura immediatamente interpretabile dell'errore medio in termini di numero di annunci;
- **MAPE** (Mean Absolute Percentage Error), che esprime l'errore medio in termini percentuali rispetto al valore osservato, consentendo confronti su serie con scale diverse. È utile ma può risultare instabile in presenza di valori molto bassi;

- **RMSE** (Root Mean Squared Error), che penalizza maggiormente gli errori di grande entità, risultando utile per valutare la capacità del modello di evitare scostamenti rilevanti;
- **R^2** (coefficiente di determinazione), che indica la quota di variabilità della serie osservata, spiegata dal modello. Valori prossimi a 1 suggeriscono un buon adattamento, sebbene R^2 non sia una metrica concepita specificamente per le serie temporali.

5.2 Confronto dei modelli

Sulla base delle metriche sopra descritte, il confronto ha evidenziato che:

- ETS e ARIMA forniscono una base solida, catturando bene trend e stagionalità, ma tendono a mostrare limiti nell'adattarsi a fasi di brusca crescita o contrazione;
- Random Forest risulta meno performante di GBM/LightGBM, confermando l'aspettativa che, per serie temporali con feature lag ben definite, il boosting offra vantaggi superiori;
- l'ensemble ETS + ARIMA + GBM migliora in diversi casi le performance rispetto ai singoli ETS e ARIMA, ma non supera in modo sistematico LightGBM;
- LightGBM raggiunge, in media, valori inferiori di MAE e RMSE, con MAPE più contenuti e un buon equilibrio tra adattamento e generalizzazione.

In termini operativi, ciò ha giustificato la scelta di adottare LightGBM come modello principale per il forecasting a 6 mesi del numero di job posting in Puglia, mantenendo ETS e ARIMA come termini di confronto interpretativi e l'ensemble come possibile evoluzione futura in presenza di maggiori dati e risorse computazionali.

6 LIMITI E ASSUNZIONI

L'intero impianto previsionale si fonda su un insieme di assunzioni, che è opportuno esplicitare per delimitare il corretto utilizzo dei risultati:

- **Proxy di domanda di lavoro**. I job posting digitali sono utilizzati come proxy della domanda di lavoro. Si assume che, sebbene non esaustivi, essi riflettano in modo sistematico tendenze e variazioni della domanda di lavoro formale in determinati segmenti del mercato.
- **Orizzonte previsivo breve (6 mesi)**. L' orizzonte così limitato è adatto a previsioni a brevissimo termine (nowcasting) sulla domanda di lavoro. Si assume che le relazioni osservate nel passato recente rimangano relativamente stabili nei successivi 6 mesi, pur riconoscendo che shock imprevisti possono alterare significativamente le dinamiche.
- **Modellizzazione univariata**. L'uso prevalente di modelli univariati, basati su lag della serie, implica l'assunzione che l'informazione contenuta nella storia dei job posting sia sufficiente a produrre previsioni utili a breve termine. Si riconosce che modelli multivariati, che includano

indicatori macro o di contesto, potrebbero migliorare ulteriormente le prestazioni, ma richiedono basi dati armonizzate non ancora disponibili in modo sistematico.

- **Stazionarietà locale**. Si assume che, pur in presenza di trend e stagionalità, le relazioni tra i lag della serie e i valori futuri siano relativamente stabili all'interno dell'orizzonte considerato. Cambiamenti strutturali improvvisi (ad esempio innovazioni tecnologiche dirompenti, crisi impreviste) possono compromettere questa assunzione.
- **Limiti di granularità**. Le previsioni a livello molto disaggregato (singole professioni in specifiche province) sono soggette a elevata incertezza a causa delle basse numerosità osservate. Si raccomanda l'uso prevalente dei risultati a livelli aggregati (nel caso del progetto per filiera).

Questi limiti non annullano l'utilità del sistema, ma devono essere tenuti presenti nella lettura dei risultati e nella loro integrazione con altri strumenti di analisi e programmazione.

7 CONCLUSIONI E LINEE DI MIGLIORAMENTO

La sperimentazione condotta nell'ambito dell'Osservatorio del Mercato del Lavoro in Puglia ha permesso di consolidare un quadro metodologico per l'uso di tecniche IA/ML nel forecasting a breve termine del mercato del lavoro regionale. Ha inoltre consentito di verificare, su dati reali, la capacità di modelli come LightGBM di prevedere la dinamica dei job posting e ha messo in evidenza limiti e potenzialità di un approccio data-driven basato su fonti digitali, in relazione ai modelli previsionali tradizionali (INAPP, Cedefop).

Il valore aggiunto principale per l'Osservatorio può essere sintetizzato in tre elementi:

1. **Chiarezza metodologica**. Il percorso è stato strutturato in modo trasparente, documentando fonti e manipolazioni dei dati, modelli considerati e criteri di selezione ed infine limiti e assunzioni. Questo approccio facilita la riproducibilità, l'aggiornamento e la comunicazione interna ed esterna dei risultati.
2. **Capacità prognostica a breve termine**. Il modello LightGBM offre una base operativa per aggiornamenti periodici (ad esempio semestrali) delle previsioni di job posting e fornisce supporto informativo a decisioni di breve periodo in materia di orientamento, formazione e politiche attive nelle filiere selezionate.
3. **Fondamento per un'evoluzione futura**. La Nota non si limita a fotografare uno stato dell'arte, ma delinea traiettorie di sviluppo future, in coerenza con le esigenze emergenti.

Le principali linee di miglioramento possono essere ricondotte a tre assi.

a) Dati e infrastruttura informativa:

- potenziamento dell'accesso a dati COB a livello micro e loro integrazione sistematica con i dati utilizzati;

- arricchimento delle informazioni su competenze e percorsi formativi, attraverso un uso più integrato di Atlante del Lavoro, ESCO e dati regionali sull'offerta formativa;
- miglioramento continuo dei processi di data quality per i job posting per esempio monitorando la copertura e gestendo eventuali cambiamenti nelle fonti.

b) Evoluzione metodologica:

- passaggio progressivo verso modelli multivariati, includendo indicatori macro e di contesto, per rendere le previsioni più sensibili alle dinamiche economiche generali;
- sviluppo di un sistema di forecasting gerarchico con riconciliazione tra livelli, in modo da ottenere previsioni coerenti per regione, province, settori e cluster professionali;
- introduzione di previsioni probabilistiche e per scenari, affiancando ai valori puntuali intervalli di confidenza e scenari alternativi;
- eventuale sperimentazione di modelli deep learning per serie temporali, laddove la crescita delle basi dati lo giustifichi.

c) Integrazione con la governance e le politiche regionali:

- allineamento sistematico tra le previsioni prodotte e gli strumenti di pianificazione regionale (Piani del Lavoro, Programmazione FSE+, PNRR, ecc.);
- utilizzo dei risultati come input di sistemi informativi e cruscotti per i servizi per il lavoro, i decisori regionali e gli attori di filiera;
- definizione di un quadro di AI governance per i modelli previsionali, che includa procedure di monitoraggio, revisione periodica, documentazione e gestione dei rischi (bias nei dati, degradazione delle performance nel tempo).

In conclusione, **il lavoro svolto rappresenta un primo passo verso la costruzione di una capacità previsionale regionale basata sull'integrazione tra fonti tradizionali e nuove fonti digitali, supportata da tecniche di IA/ML e orientata alle esigenze concrete di policy**. La prospettiva è quella di un percorso incrementale, nel quale l'accrescimento delle basi dati e la maturazione di competenze metodologiche consentano, nel tempo, di passare da un sistema sperimentale a un'infrastruttura stabile a supporto delle decisioni della Regione Puglia.

Appendice

Tabella 1: Descrizione fonte dati lightcast

COLONNE DISPONIBILI	DESCRIZIONE DATO
general_id	Chiave numerica univoca per annuncio
year_grab_date	Anno di pubblicazione
month_grab_date	Mese di pubblicazione
day_grab_date	Giorno di pubblicazione
year_expire_date	Anno di scadenza
month_expire_date	Mese di scadenza
day_expire_date	Giorno di scadenza
idesco_level_4	Codice ESCO
idcity	Identificativo città di riferimento
idprovince	Identificativo provincia di riferimento
idregion	Identificativo regione di riferimento
idmacro_region	Identificativo macro-regione di riferimento (nord-ovest, nord-est, sud, isole, centro)
idcountry	Identificativo stato di riferimento
idcontract	Identificativo tipologia contratto proposta
ideducational_level	Identificativo livello educativo richiesto
idsector	Identificativo settore di riferimento
idmacro_sector	Identificativo macro-settore di riferimento
idexperience	Identificativo anni di esperienza richiesti
idworking_hours	Identificativo tipologia contratto (PT o FT)
source	Sito fonte dato

Tabella 2: Colonne presenti nei dataset con attivazioni/cessazioni

COLONNE	DESCRIZIONE DATO
Codice CP 2021	Codice della classificazione delle professioni CP2021
Descrizione CP 2021	Descrizione derivante dalla classificazione delle professioni CP2021
Codice ESCO	Codice della classificazione ESCO delle professioni
Descrizione ESCO	Descrizione derivante dalla classificazione ESCO delle professioni
23q1	Rilevazioni anno 2023 primo trimestre
23q2	Rilevazioni anno 2023 secondo trimestre
23q3	Rilevazioni anno 2023 terzo trimestre
23q4	Rilevazioni anno 2023 quarto trimestre
24q1	Rilevazioni anno 2024 primo trimestre
24q2	Rilevazioni anno 2024 secondo trimestre
24q3	Rilevazioni anno 2024 terzo trimestre
24q4	Rilevazioni anno 2024 quarto trimestre
25q1	Rilevazioni anno 2025 primo trimestre

Bibliografia

1. INAPP. Scenari di medio termine per l'economia e l'occupazione. INAPP. [Online] 2024. https://oa.inapp.gov.it/server/api/core/bitstreams/7f05ae45-ed15-448b-a2ab-68df95282aaa/content.
2. Mereu, Maria Grazia. Scenari di medio termine per l'economia e l'occupazione. INAPP. [Online] INAPP, 2024. https://oa.inapp.org/xmlui/handle/20.500.12916/4208.
3. Cedefop – European Centre for the Development of Vocational Training. Skills Forecast Methodological Framework. Cedefop. [Online] 2023. https://www.cedefop.europa.eu/files/skills_forecast_methodological_framework.pdf.
4. Sviluppo Lavoro Italia. LMI – Labour Market Intelligence 2025. Consultabile su: Osservatorio del Mercato del Lavoro ARTI Puglia. [Online] https://osservatoriolavoro.arti.puglia.it/lmi-labour-market-intelligence-2025.
5. LIGHTCAST. Italian Labour Market Dataset. 2025.
6. Leogrande, Angelo. Clustering and Prediction of the Employment Rate in the Italian Regions in the Period 2004-2019.
7. Corporation, Microsoft. LightGBM. Read the docs. [Online] https://lightgbm.readthedocs.io/en/stable/Features.html.
8. Kyungsu, Kim. Forecasting Labor Demand: Predicting JOLT Job Openings using Deep Learning Model. Georgia Institute of Technology School of Engineering, Atlanta GA 30332, USA: s.n., arXiv preprint arXiv:2503.19048, 2025.
9. Hu J, Szymczak S. A review on longitudinal data analysis with random forest. 2023. PMID: 36653905; PMCID: PMC10025446.
10. Kyungsu, Kim. Unemployment Dynamics Forecasting with Machine-Learning Regression Models. Georgia Institute of Technology School of Engineering, Atlanta GA 30332, USA: s.n.
11. Masini, Ricardo P., Medeiros, Marcelo C., Mendes. Eduardo F. Machine Learning Advances for Time Series Forecasting. 2021.
12. ARTI – Agenzia Regionale per la Tecnologia, il Trasferimento tecnologico e l'Innovazione. Osservatorio del Mercato del Lavoro Puglia. [Online] https://osservatoriolavoro.arti.puglia.it.
13. ISTAT. Classifica delle professioni. ISTAT. [Online] ISTAT. https://www.istat.it/classificazione/classificazione-delle-professioni.
14. Commissione Europea. ESCO. [Online] European Commission. https://esco.ec.europa.eu/it/use-esco/download.