XXXII CONFERENZA ITALIANA DI SCIENZE REGIONALI

UN MODELLO NON PARAMETRICO PER LA DINAMICA SPAZIALE DELL'OCCUPAZIONE GIOVANILE

Rocco Vincenzo SANTANDREA[1], Iary Ilario Paolo GOFFREDO[2]

SOMMARIO

Il presente lavoro costituisce un approfondimento nell'ambito del progetto di ricerca intrapreso dall'IPRES (Istituto Pugliese di Ricerche Economiche e Sociali) e denominato "*inserimento lavorativo dei giovani diplomati nel mercato del lavoro*".
Dopo aver fornito una sintetica analisi delle principali questioni riguardanti i giovani (15-29 anni di età) ed il loro inserimento nel mercato del lavoro, comparando situazioni e dinamiche a livello di alcuni Paesi dell'Unione Europea e nazionale, si passa alla formulazione di un approccio metodologico innovativo dello stesso fenomeno, basato sulla costruzione di un modello di classificazione ad albero. Tale modello, infatti, si ritiene che possa fornire gli strumenti idonei ad individuare le caratteristiche che comportano un maggiore rischio di non occupazione del collettivo di riferimento. L'individuazione delle caratteristiche di rischio più significative può consentire di sviluppare utili suggerimenti di policy, finalizzati a ridurre e/o correggere i fattori di rischio individuati.

[1] Ricercatore IPRES, Piazza Garibaldi 13, 70122, Bari, e-mail: vincenzo.santandrea@ipres.it
[2] Dottore di ricerca in Statistica – Università "Aldo Moro"di Bari, e-mail: goffredo@dss.uniba.it

Introduzione[3]

La questione dei giovani e del mercato del lavoro è divenuta uno snodo fondamentale del dibattito in corso sia a livello internazionale che nazionale nell'ambito delle politiche di crescita dell'occupazione.

Sotto il profilo internazionale, recenti ricerche ed analisi[4] hanno evidenziato come la crisi economica globale degli ultimi tre anni abbia avuto un impatto significativamente maggiore sui giovani tra i 15 e i 29 anni rispetto alle altre classi di età.

Inoltre, la fase attuale di incerta ripresa economica a livello internazionale, sembra penalizzare maggiormente questa classe di età rispetto alle altre, tanto che le maggiori organizzazioni internazionali (ILO, OECD, UE) sollecitano specifiche attenzioni e misure di intervento per questa tipologia di soggetto sociale[5].

A livello nazionale, negli ultimi anni si è sviluppata un'ampia letteratura scientifica che ha cercato di analizzare e mettere a fuoco il sistema bloccato del mercato del lavoro con riferimento ai giovani[6].

La questione dei giovani ed il lavoro è stata oggetto di una riflessione specifica del Governatore della Banca d'Italia nelle considerazioni finali del 2010 e del 2011, che ha evidenziato l'ampliamento delle differenze nelle condizioni lavorative tra i giovani e i lavoratori più anziani a favore di questi ultimi[7]. Nelle ultime considerazioni finali il Governatore della Banca d'Italia[8] ha ripreso l'argomento che: "*La diffusione nell'ultimo quindicennio dei contratti di lavoro a tempo determinato e parziale ha contribuito a innalzare il tasso di occupazione, ma a costo di introdurre nel mercato un pronunciato dualismo: da un lato i lavoratori in attività a tempo indeterminato, maggiormente tutelati; dall'altro una vasta sacca di precariato, soprattutto giovanile, con scarse tutele e retribuzioni*".

Da ultimo, il tema considerato di questo lavoro è stato oggetto di un contributo da parte del Direttore Generale della Banca d'Italia[9] che ha evidenziato una specificità del caso italiano delle difficoltà di inserimento lavorativo dei giovani nel mercato del lavoro, con un dualismo

[3] Pur se il lavoro è frutto di una comune condivisione, il paragrafo 1 è da attribuirsi a Rocco Vincenzo Santandrea, mentre i paragrafi da 2 a 4 sono da attribuirsi a Iary I.P. Goffredo; il paragrafo 5 è da attribuirsi ai due autori.

[4] Crf. OECD (2010*) Off a good start? Jobs for Youth*, Paris; EUROSTAT (2009) *Youth in Europe*, Bruxelles; ILO (2010) *Global employment trends for Youth*, Geneva

[5] EC Commission *An EU Strategy for Youth – Investing and Empowering,* COM (2009) 200 final; Schindler M. (2009) *The Italian Labor Market: Recent Trends, Institutions and Reform Options*, IMF, WP 09/47

[6] Boeri T., Galasso V. (2007) Conto i Giovani, Mondadori, Milano; Rosina A., Ambrosi E. (2009) *Non è un Paese per giovani*, Marsilio; Leombruni R., Taddei F. (2009) *Giovani precari in una Paese per vecchi*, Il Mulino, 6, 912-920

[7] Draghi M (2010) *Considerazioni Finali*, Banca D'Italia, 2010, 13

[8] Draghi M (2010) *Considerazioni Finali*, Banca D'Italia, 2011, 13

[9] Saccomanni F. (2011) *La generazione esclusa: il contributo dei giovani alla crescita economica*, 41° Convegno dei Giovani Imprenditori di Confindustria, Santa Margherita Ligure, 11 giugno, (mimeo).

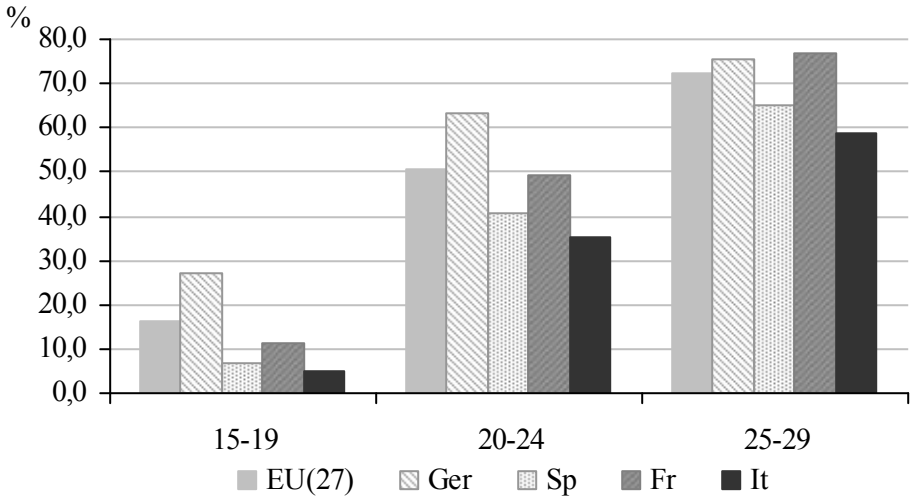
accentuato tra Centro-Nord e Mezzogiorno. Con riferimento a quest'ultimo aspetto, se da un lato esiste una specificità italiana, dall'altro nel Mezzogiorno questo fenomeno è significativamente accentuato[10].

1 La dinamica dell'occupazione giovanile

1.1 Un confronto con alcuni Paesi dell' Unione Europea

L'analisi comparata del tasso di occupazione dei giovani ha considerato tre Paesi membri dell'Unione Europea di dimensioni simili all'Italia e il dato medio dell'UE a 27 Paesi membri, e tre classi di età distinte: 15-19, 20-24, 25-29 anni.
Come si può osservare, l'Italia evidenzia situazioni nettamente differenti nei tassi di occupazione giovanili, in modo particolare in rapporto al valore medio di EU 27 e della Germania sin dalla fascia dei giovani 15-19 anni. Le maggiori differenze si manifestano nelle due prime classi di età, ma sono significative anche rispetto alla terza fascia[11].

Figura 1 - Tassi di occupazione giovanili in alcuni Paesi UE - 2010



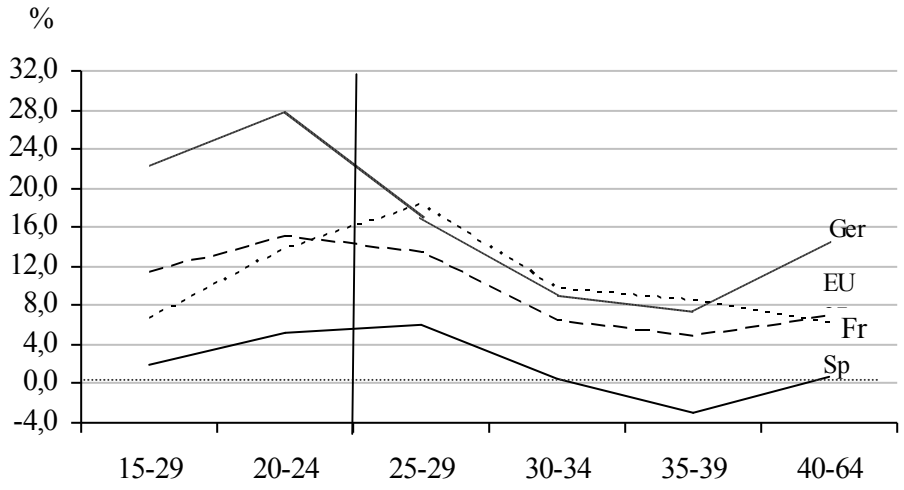
Fonte: EUROSTAT, Elaborazioni IPRES

La Spagna è molto simile all'Italia per la fascia 15-19 anni, mentre ha una posizione nettamente migliore per le due classi successive. Prendendo in considerazione anche i tassi di occupazione di altre classi di età, emerge la specificità del caso italiano in riferimento al mercato del lavoro: una difficoltà strutturale ad offrire opportunità occupazionali alle giovani generazioni.

[10] Bianchi L., Provenzano G., (2010) *Ma il cielo è sempre più su?*, Castelvecchio Editore, Tazebao,Roma
[11] IPRES (2010) *Capitale umano qualificato, mercato del lavoro e mobilità territoriale*, Quaderni *IPRES*, 2, Cacucci Editore, Bari

Infatti, misurando le distanze in termini di punti percentuali tra i diversi Paesi considerati, emerge come le differenze di gran lunga maggiori nei confronti dell'Italia riguardano i tassi di occupazione delle classi di età 15-19 e 20-24; diminuiscono leggermente per la classe di età 25-29; mentre distanze nettamente inferiori si riscontrano per i tassi di occupazione delle classi di età 30-64 anni.

Figura 2 - Distanza in termini di punti percentuali nei tassi di occupazione tra Italia e alti Paesi considerati – 2010



Fonte: EUROSTAT, Elaborazioni IPRES

La crisi economica degli ultimi tre anni ha colpito in modo particolare i giovani[12]. Questa situazione non ha riguardato tutte le economie, ma si sono manifestate situazioni differenziate tra le diverse macroaree economiche del mondo. Con riferimento ai tassi di occupazione 15-24 anni nel periodo 2008-2009 si osserva la seguente situazione[13]:

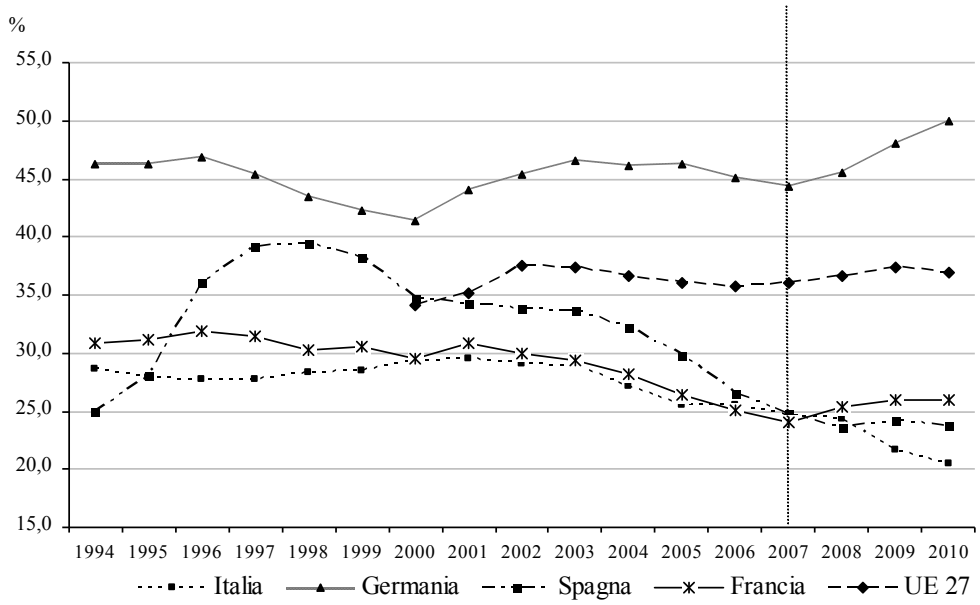
Aree in cui il tasso di occupazione è diminuito	Aree in cui il tasso di occupazione è aumentato	Aree in cui il tasso di occupazione è rimasto stazionario
Economie sviluppate ed Unione Europea (- 2,7 punti %);	Est Asia (+ 0,5 punti %);	Sud Est Asia e Pacifico;
America Latina, (- 1,3 punti %);	Medio Oriente (+ 0,5 punti %);	Sud Asia;
Europa Centrale e Sud Est (non EU) (- 0,7 punti %);	Nord Africa (+ 0,5 punti %)	Africa Sub-Sahariana

È da osservare che nelle aree Asiatiche ed Africane la dinamica demografica delle giovani generazioni è di gran lunga superiore a quella delle economie avanzate e dell'Europa Centrale e del Sud-Est Europa.

[12] Il rapporto dell'OECD del 2010 (OECD 2010), evidenzia come la crisi economica abbia colpito in modo particolarmente forte "particularly hard" i giovani ed e le persone con contratti di natura temporanea, che si concentrano in questa categoria di lavoratori.
[13] ILO (2010), op.cit.

Come è evidente, vi è stata una accentuazione particolare per le economie sviluppate e l'Unione Europea. Tuttavia, anche all'interno di quest'ultima area le situazioni si presentano molto differenziate tra i Paesi membri. Infatti, prendendo in considerazione la dinamica del tasso di occupazione dei giovani di 15-24 anni nei Paesi UE considerati negli ultimi quindici anni, si possono avanzare alcune considerazioni. In primo luogo, a partire dagli anni 2000 mentre per la Germania si osserva un leggero trend crescente, si verifica un forte declino per la Spagna, e un declino simile per la Francia e l'Italia, mentre a livello UE 27 si ha un andamento costatante attorno ad un tasso del 35-37%. In secondo luogo, a partire dal 2008 si osserva una sorta di "rottura" tra l'Italia e gli altri Paesi considerati nella dinamica dei tassi di occupazione giovanile 15-24 anni: nella prima diminuiscono costantemente, in Spagna si mantengono costanti sui valori del 2007, in Francia e Germania tendono addirittura ad aumentare.

Figura 3 – Dinamica del tasso di occupazione 15-24 anni in alcuni Paesi UE



Fonte: EUROSTAT, Elaborazioni IPRES

Pertanto se la questione dell'occupazione giovanile riguarda molti Paesi, l'Italia evidenzia situazioni particolarmente critiche[14].

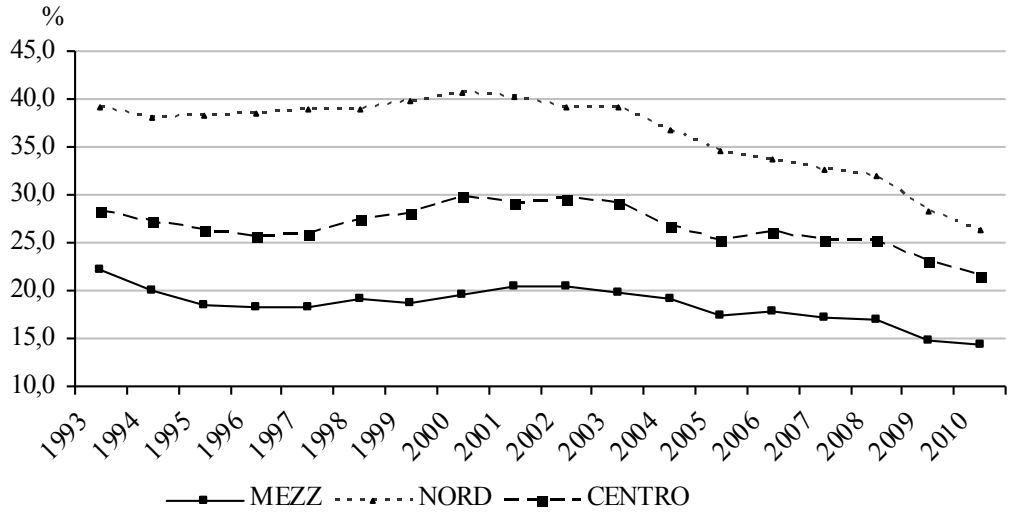
1.2 L'occupazione giovanile in Italia

Per quanto riguarda i giovani, il mercato del lavoro tra le tre grandi macro aree Nord, Centro e Mezzogiorno differiscono per i livelli ma non per la dinamica: il tasso di occupazione dei giovani 15-24 anni è più basso nel Mezzogiorno rispetto alle altre due macro aree, ma la dinamica temporale di circa due decenni è molto simile. Questa situazione evidenzia come la

[14] Cfr. su questo punto Saccomanni F. (2011)

questione del mercato del lavoro giovanile sia di natura strutturale e nazionale, mentre per il Mezzogiorno vi è una maggiore incidenza nei livelli.

Figura 4 - Dinamica del tasso di occupazione 15-24 anni per le tre grandi ripartizioni territoriali

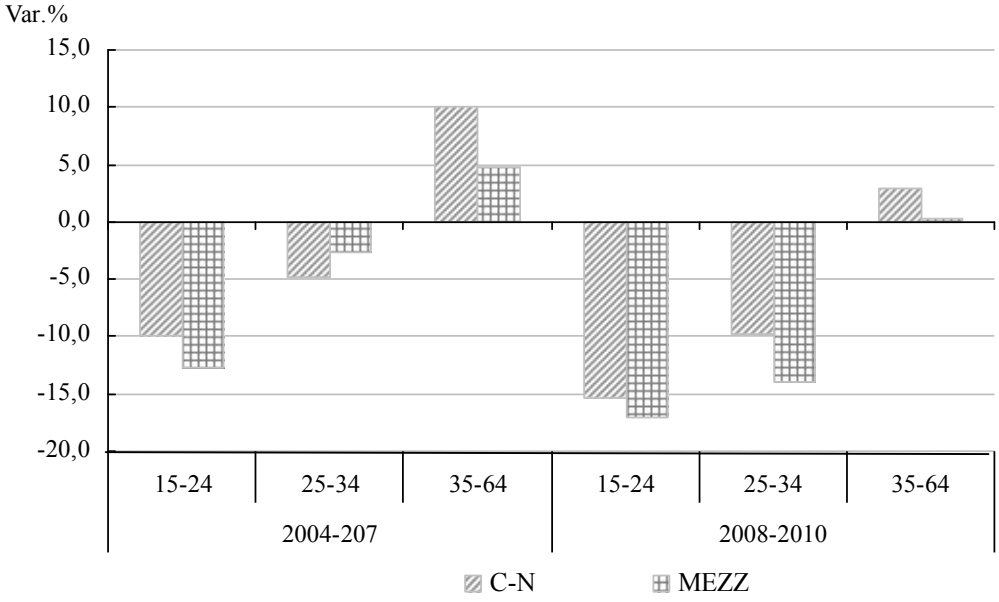


Fonte: ISTAT- RCFL, Elaborazioni IPRES

La simile dinamica temporale del tasso di occupazione giovanile si riscontra anche negli ultimi tre anni di crisi economica.

Che ci sia una diversità (significativa) nei livelli ma non nei "comportamenti" tra le diverse macro aree del Paese è dimostrata anche dal fatto che, scomponendo due periodi tra 2004-2007 e 2008-2010 (periodo pre-crisi e periodo di crisi), le dinamiche sono molto simili.

Figura 5 – Variazioni % nell'occupazione per classi di età



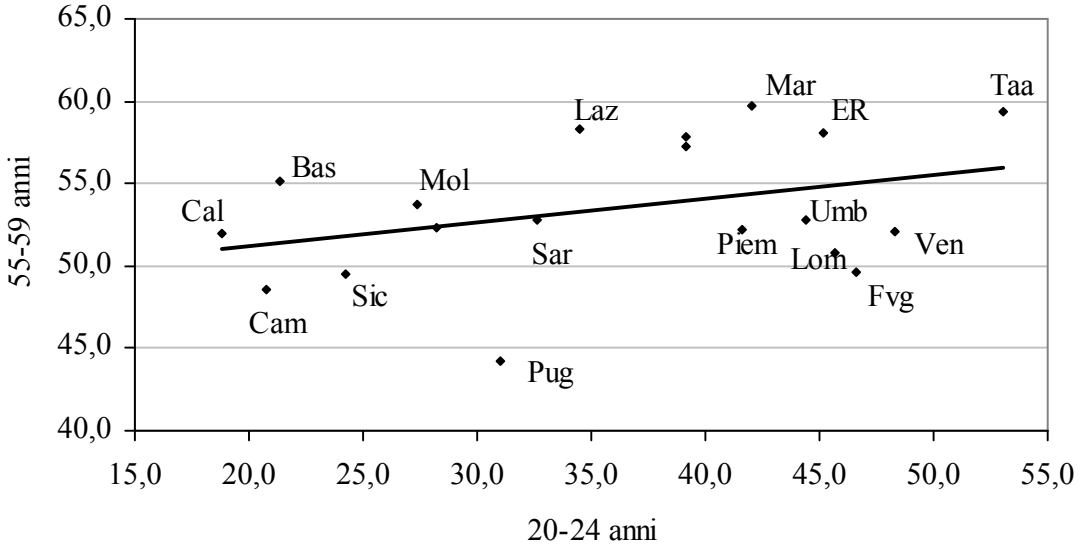
Fonte: ISTAT- RCFL, Elaborazioni IPRES

Infatti, in ambedue le aree diminuisce l'occupazione giovanile per le classi di età 15-24 e 25-34 nei due periodi considerati, mentre aumenta quella della classe 35-64, sempre nei due periodi considerati. Inoltre, è maggiore la riduzione dell'occupazione per la classe 15-24 anni rispetto a quella 25-34 in ambedue i periodi considerati. Infine, si può osservare che la difficoltà di trovare un'occupazione per i giovani viene da lontano, è un elemento strutturale; la crisi ne ha solo accentuato l'incidenza e, quindi, anche la percezione.

Per il Mezzogiorno la questione dei giovani in rapporto al mercato del lavoro assume anche un'altra valenza in quanto questa area è caratterizzata da tre ulteriori fenomeni: una riduzione significativa della componente demografica giovanile, un flusso migratorio netto Sud-Nord di significativa entità e un flusso di pendolarismo giobanile, anche esso di significativa entità[15].

Abbiamo provato anche a verificare che non ci sia una relazione inversa tra tasso di occupazione giovanile (20-24 anni) e tasso di occupazione degli anziani (55-59 anni), sulla base di una "percezione" che sostiene che la maggiore quantità di occupazione anziana sottrae lavoro alle classi più giovani.

Figura 6 – Relazione tra i tassi di occupazione dei giovani e quelli anziani- 2010



Fonte: ISTAT- RCFL, Elaborazioni IPRES

Un'analisi preliminare mostra che questa percezione è sbagliata. Prendendo in considerazione i tassi di occupazione dei giovani 20-24 anni e degli anziani 55-59 anni per le regioni italiane nel 2010, si può osservare che c'è una relazione positiva ma con valori molto modesti: sembrano due variabili poco correlate. Un medesimo esercizio effettuato recentemente

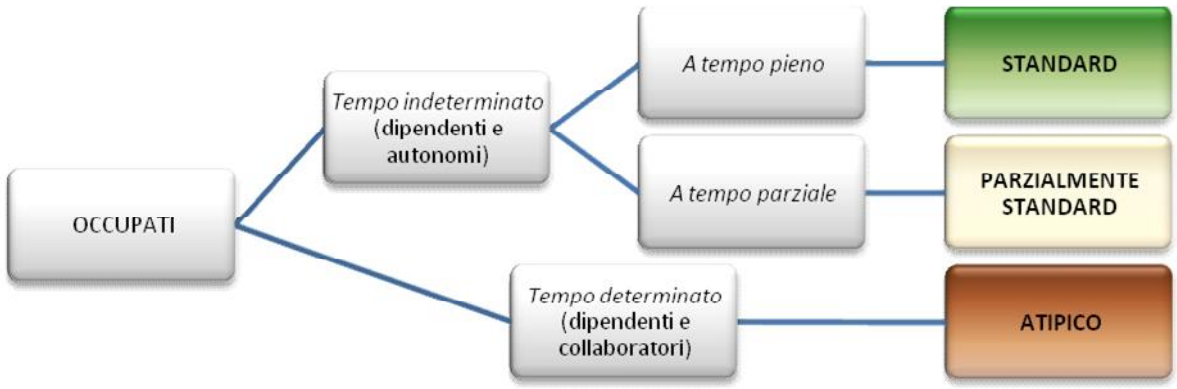
[15] SVIMEZ (2010) *Rapporto sull'economia del Mezzogiorno. 2009*, Il Mulino, Bologna; Mocetti S., Carmine P. (2010) La mobilità del lavoro in Italia: nuove evidenze sulle dinamiche migratorie, *Quaderni di Economia e Finanza*, 61, Banca D'Italia; IPRES (2010), op. cit.

dall'OCSE[16], per l'insieme dei Paesi membri, mostra una correlazione positiva nettamente maggiore di quella stimata per le regioni italiane e conferma che questa percezione è sbagliata.

La struttura delle diverse tipologie di lavoro è sostanzialmente simile nel Mezzogiorno rispetto alla ripartizione del Centro-Nord.

La crisi economica degli ultimi tre anni ha mostrato andamenti sostanzialmente simili per le diverse tipologie di lavoro tra le due aree del Paese. Per fornire una misura di questo fenomeno, abbiamo utilizzato la classificazione proposta dall'ISTAT[17], la cui rappresentazione grafica è riportata nella *Figura 7*, in base alla quale il lavoro viene classificato in "standard", "parzialmente standard" e "atipico" a seconda dell'intreccio fra posizione professionale (dipendente, collaboratore o autonomo), durata (determinata o indeterminata) e orario di lavoro (a tempo pieno o a tempo parziale).

Figura 7 – Classificazione della tipologia di lavoro in funzione della posizione professionale, della durata contrattuale e dell'orario di lavoro



Fonte: Istat – Rapporto Annuale 2008

Sulla base della classificazione appena descritta, con riferimento alla situazione dei giovani 15-29enni, non emergono significative differenze fra la situazione nel Mezzogiorno e quella del Centro-Nord: in entrambi i casi vi è una larga prevalenza del lavoro standard.

Il lavoro atipico, tuttavia, è presente in modo significativo nel mercato del lavoro giovanile, assorbendone all'incirca il 30% sia di quello meridionale che di quello centro-settentrionale e, contrariamente a quello che accade per il lavoro standard, un rilevante contributo è dato dalla componente femminile.

[16] OECD (2010). In questo studio si sostiene che non è vero che dove ci sono più occupati anziani, ci sono meno occupati giovani: *"younger workers cannot necessarily be easily substituted for older workers and the costs of subsidising early retirement can result in reduced employment opportunities for younger workers because of higher labour taxes to finance these costs*." pag. 37

[17] ISTAT (2009) *Rapporto annuale 2008*, Roma

Tabella 1 – Lavoro standard e lavoro atipico fra i giovani 15-29 anni - Anno 2010

Tipologie di lavoro	Valori assoluti (.000)		Quota %		Variazioni % 2008-2010	
	Centro-Nord	Mezzogiorno	Centro-Nord	Mezzogiorno	Centro-Nord	Mezzogiorno
Standard	1.393	525	58,6	57,1	-14,0	-19,6
Parzialmente standard	236	122	9,9	13,3	-3,5	-7,1
Atipico	747	272	31,5	29,6	-9,7	-15,0
Totale	**2.376**	**919**	**100**	**100**	**-11,7**	**-16,8**

Fonte: ISTAT- RCFL, Elaborazioni IPRES

Differenze più marcate si riscontrano nella quota di lavoro parzialmente standard, superiore nel Mezzogiorno di circa 3,5 punti percentuali.
Il confronto con la situazione al 2008, invece, fa emergere chiaramente come la crisi economica abbia colpito pesantemente l'occupazione, in modo particolare il lavoro standard e quello atipico; una maggiore incidenza di queste dinamiche si verifica nel Mezzogiorno, evidenziando una più spiccata "vulnerabilità" del mercato del lavoro nei confronti dei giovani.

L'analisi ha mostrato che vi sono delle specificità strutturali per l'Italia con riferimento al mercato del lavoro giovanile: lo status occupazionale dei giovani è molto basso, a fronte di un livello molto elevato di persone in cerca di lavoro e di inattivi in età da lavoro. Inoltre, la caratteristica duale del mercato del lavoro, sotto il profilo territoriale, si manifesta nei livelli e non nelle dinamiche.
Per indagare in modo più approfondito su queste caratteristiche strutturali abbiamo fatto ricorso all'applicazione di un modello che viene presentato nei prossimi paragrafi, con i relativi risultati.

2 Gli obiettivi di un modello empirico

Dopo aver fornito i tratti generali del contesto d'interesse per il presente lavoro, attraverso un'analisi descrittiva degli aspetti più rilevanti delle dinamiche del mercato del lavoro giovanile, si vuole ora proporre, in questa seconda parte del lavoro, un approccio metodologico per l'implementazione di un modello empirico finalizzato ad analisi interpretative e previsive del fenomeno studiato.
È verosimile pensare che il rischio, per un giovane, di non avere un inserimento lavorativo nel mercato del lavoro, consolidato dalla pesantissima crisi economica internazionale degli ultimi tre anni, possa essere sistematicamente aggravato da alcune sue caratteristiche individuali, e/o di contesto socio-economico e territoriale. L'individuazione di profili dei giovani

caratterizzati da una sistematica criticità nella collocazione sul mercato del lavoro, indispensabile per orientare opportunamente le politiche del lavoro costituisce, quindi, l'obiettivo principale di questo contributo, per il quale si propone l'implementazione di un modello di classificazione ad albero. Un modello di questo genere, infatti, permette di individuare le variabili statistiche che "tagliano" in misura maggiormente discriminante il collettivo dei giovani rispetto alla loro collocazione sul mercato del lavoro, in modo da delinearne i profili critici e quelli, invece, più virtuosi, sulla base delle caratteristiche osservate.

In particolare, dopo aver discusso, nel prossimo paragrafo, degli aspetti metodologici dei modelli di classificazione ad albero, ne verrà proposta un'applicazione concreta nel successivo e conclusivo paragrafo.

Le variabili statistiche, rappresentative delle caratteristiche individuali dei giovani presi in considerazione, nonché l'approccio metodologico utilizzato per l'implementazione del modello, non si propongono di essere esaustivi in merito alla spiegazione delle criticità delle dinamiche del mercato del lavoro giovanile, ma vogliono essere uno spunto per future ricerche nell'ambito delle quali allargare o eventualmente modificare il campo di osservazione a variabili statistiche non prese in considerazione in questo lavoro o affinare opportunamente il modello attraverso l'impiego di differenti strumenti metodologici. Tuttavia, la versatilità e le peculiarità che contraddistinguono i modelli di classificazione ad albero, che verranno compiutamente illustrate nel successivo paragrafo, ci hanno portato a ritenere che il loro impiego in tale contesto sia opportuno ed efficace ai fini interpretativi e previstivi per cui esso si rende necessario.

C'è da dire, infine, che l'impiego di questi modelli nel contesto delle dinamiche occupazionali, rappresenta una novità metodologica, visto che, finora, i modelli di classificazione ad albero sono stati impiegati prevalentemente dagli istituti di credito nell'ambito dello studio del *credit-scoring* come strumenti di valutazione del merito di credito dei potenziali fruitori dei finanziamenti, a causa delle diversificate caratteristiche delle variabili che si può verosimilmente assumere che possano influenzare la solvibilità dei richiedenti del credito.

Le peculiarità dei modelli di classificazione ad albero, tuttavia, come si è detto e come verrà ulteriormente spiegato nel successivo paragrafo, hanno portato a ritenere tale approccio adeguato anche al contesto di studio delle dinamiche occupazionali, nell'ambito delle quali valutare il rischio della non occupazione dell'individuo, con particolare riferimento all'età giovanile.

3 I modelli di classificazione ad albero

3.1 Aspetti generali[18]

I modelli di classificazione ad albero rappresentano un moderno strumento statistico metodologico per lo studio di fenomeni caratterizzati dalla presenza di una variabile di output (variabile risposta) e di una serie di variabili di input (predittori) nell'ambito dei quali si voglia ottenere la segmentazione gerarchica di un collettivo finalizzata al raggiungimento di obiettivi di carattere esplorativo o decisionale.

Le caratteristiche principali di tali modelli sono quelle di essere:

- supervisionati;
- non parametrici;
- non lineari.

Avendosi a disposizione un campione *(training set)* che contiene, accanto alle osservazioni dei predittori, anche quelle della variabile di output, queste ultime fungono da "supervisore" nel processo di apprendimento delle relazioni che legano predittori e variabile risposta finalizzato agli obiettivi per cui il modello è stato costruito, rendendo tale modello, appunto, *supervisionato*. Si fa notare, che tale aspetto costituisce anche la sostanziale differenza fra i modelli di classificazione ad albero e quelli di cluster analysis, nell'ambito dei quali, viceversa, la classificazione delle unità statistiche non è nota a priori, ma deve essere ottenuta mediante l'applicazione di opportuni algoritmi sulla base del livello di somiglianza delle unità statistiche rispetto alle caratteristiche osservate.

Le variabili di input e di output, d'altra parte, possono essere sia di tipo qualitativo che quantitativo e non devono sottostare a restrizioni di tipo distributivo affinché sia possibile impiegarle in questi modelli: per questa ragione, tali modelli possono essere definiti come *non parametrici*, che rappresenta certamente un vantaggio dal punto di vista della loro flessibilità ed applicabilità agli ambiti più diversi.

Le strutture ad albero che rappresentano le relazioni ottenute dall'applicazione di questi modelli, infine, non sono rappresentabili attraverso una semplice forma funzionale lineare e perciò si parla di modelli *non lineari*.

Gli obiettivi di carattere esplorativo della segmentazione gerarchica ottenibile mediante l'applicazione di questi modelli, consistono nell'individuazione e nella descrizione delle relazioni esistenti fra le variabili di input e l'appartenenza delle unità statistiche osservate alle

[18] Aria, M. (2009) *Alberi di Classificazione*, dispense del corso di Analisi Statistica e Sociologica per i processi economici e del lavoro nel settore turistico, Anno accademico 2009-2010, Facoltà di Economia, Università degli Studi di Napoli "Federico II", Federica e-Learning; Brieman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984) *Classification and regression trees*, Belmont C.A. Wadsworth

diverse classi della variabile di output. Quelli di carattere decisionale, invece, sono finalizzati alla costruzione di un albero che consenta di determinare una regola di classificazione per nuove unità statistiche per cui non è nota la classe di appartenenza, ma soltanto i valori assunti dalle variabili di input.

La struttura grafica mediante la quale può essere rappresentato un modello di classificazione ad albero è costituita da un elemento iniziale, detto *nodo radice*, a partire dal quale si diparte un numero finito di ulteriori nodi, ciascuno dei quali rappresentativi di uno *split*, ossia di una partizione del collettivo iniziale in funzione di una diversa modalità assunta da un predittore. Tali nodi, a loro volta, si distinguono in nodi *interni*, solitamente rappresentati tramite cerchi, e nodi *terminali* (o *foglie*), solitamente rappresentati tramite quadrati, la cui fondamentale differenza sta nel fatto che, mentre i nodi interni sono seguiti da almeno uno split, quelli terminali rappresentano una partizione ritenuta sufficientemente omogenea del collettivo, non sono seguiti da alcun ulteriore split e ciascuno di essi viene associato ad una modalità della variabile risposta in quanto ritenuta la più probabile che un item avente le caratteristiche definite da tale nodo, possa assumere. Una *branca* dell'albero (o *sottoalbero*), infine, si ottiene potando tale albero in corrispondenza di uno dei suoi nodi *terminali*. In relazione a quanto detto è evidente che le fasi di definizione degli *splits* (ossia la metodologia da utilizzare per suddividere i nodi) e la decisione su quando dichiarare terminale un nodo piuttosto che continuare a splittarlo, rappresentano gli aspetti cruciali della procedura di implementazione di un modello di classificazione ad albero, procedura che verrà di seguito analiticamente descritta.

3.2 Fasi della procedura

La procedura di implementazione di un modello di classificazione ad albero è costituita generalmente da un processo iterativo attraverso il quale, ad ogni passo, un nodo interno (detto nodo *padre*) viene tagliato in due o più nodi (detti nodi *figli*)[19]. Solitamente, la partizione dei nodi *padre* viene effettuata in modo tale da ottenere sistematicamente due nodi *figli* ad ogni taglio, ragion per cui, il modello di classificazione ad albero viene anche definito modello di *segmentazione binaria*.

Indicando, ora, con Y la variabile risposta che può assumere le modalità $y_1, y_2, \ldots, y_j, \ldots, y_J$, con $X_1, X_2, .., X_s$ il set delle variabili di input e con h un generico nodo dell'albero, le fasi della

[19] Manno A. (2002) *Imputazione di dati categoriali mancanti: il modello di classificazione ad albero,* Palermo, Italia, http://www.statistica.too.it; Aria, M. (2009) op. cit.; Mola, F., Siciliano, R. (1992) *A two-stage predictive splitting algorithm in binary segmentation*, inY. Dodge, J. Whittaker. (Eds.): Computational Statistics: COMPSTAT 92, 1, Physica Verlag, Heidelberg, 179-184; Mola, F., Siciliano, R. (1997) *A fast splitting procedure for classification trees*, Statistics and Computing 7; Brieman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984) op. cit.

procedura di implementazione di un algoritmo ad albero possono essere sintetizzate nel modo seguente:

1. **creazione dell'insieme degli splits:** devono essere individuate tutte le possibili domande dicotomiche generabili dalle variabili di input, ad ognuna delle quali verrà associato uno split binario. Il numero degli *splits* binari ottenibili da ogni predittore varia a seconda della tipologia della variabile che rappresenta tal predittore ed in particolare si avrà che:
   - per **variabili binarie** (es. *sesso:* maschio/femmina, *durata contrattuale*: a tempo determinato/indeterminato, *orario di lavoro*: tempo pieno/parziale,…) si può formare, ovviamente, un unico split binario;
   - per **variabili nominali** (es. *colore degli occhi*: azzurro, marrone, verde,… *settore occupazionale*: agricoltura, industria in senso stretto, industria delle costruzioni, commercio, altri servizi, …) posto *m* il numero di modalità che dette variabili possono assumere, si avranno $2^{m-1} - 1$ *splits*;
   - per **variabili ordinali** (es. *grado militare*, *livello di soddisfazione di un servizio*, *titolo di studio più elevato conseguito*, …) con *m* modalità, si avranno *m – 1 splits*;
   - per **variabili quantitative** (es. *numero di dipendenti di un'azienda*, *durata della disoccupazione espressa in mesi*, …) con *m* distinti valori, si avranno, infine, *m – 1 splits*;
2. **selezione degli splits:** una volta individuati tutti gli *splits* binari ottenibili da ogni nodo, viene scelto quello che, fra di essi, rende più omogenei i dati all'interno dei due nodi *figli* ottenibili. Per fare ciò si procede nel modo seguente:
   - **passo 1**: a partire dal nodo *radice*, viene definita, per ogni nodo, la probabilità che gli elementi ad esso appartenenti assumano una certa modalità y_j della variabile risposta, scelta in quanto ritenuta cruciale nell'ambito dello studio del fenomeno analizzato:

$$p(Y = y_j|h) = \frac{n_j(h)}{n(h)} \qquad \forall h$$

     dove $n_j(h)$ e $n(h)$ indicano, rispettivamente, il numero degli items complessivi appartenenti al nodo *h* e, fra questi, quelli che assumono come modalità della variabile risposta y_j ed avendosi, ovviamente, che:

$$\sum_{j=1}^{J} p(Y = y_j|h) = 1 \qquad \forall h$$

   - **passo 2**: si definisce *impurità del nodo h* una quantità che, indicata con il simbolo $i(h)$, può essere definita come una funzione $\phi(\cdot)$ non negativa di $p(y_j|h)$, tale che:

$$i(h) = \phi[p(y_1|h), p(y_2|h),..., p(y_J|h)]$$

tale che:

- $\phi\left(\frac{1}{J},\frac{1}{J},...,\frac{1}{J}\right)=max$;
- $\phi(1,0,...,0)=\phi(0,1,...,0)=...=\phi(0,0,...,1)=0$;
- $\phi(\cdot)$ funzione simmetrica di $p(y_1|h),\ p(y_2|h),\ \ldots,\ p(y_J|h)$.

L'impurità di un nodo, dunque, è espressione della sua omogeneità interna rispetto alle modalità della variabile risposta: essa è infatti massima quando le modalità della variabile risposta sono equamente ripartite fra gli items presenti nel nodo, ed è nulla quando, invece, detti items presentano tutti la stessa modalità.

Uno degli indici di impurità più noti ed anche più frequentemente utilizzati, è quello elaborato da Gini, dato da:

$$i(h)=\sum_{j\neq i}p(y_j|h)p(y_i|h)=\left(\sum_j p(y_j|h)\right)^2-\sum_j p(y_j|h)^2=1-\sum_j p(y_j|h)^2$$

A tal proposito, posto $p(h)=\sum_{j=1}^{J}p(y_j,h)=n(h)/n$ come la proporzione delle unità del collettivo appartenenti al nodo h, è possibile definire il concetto di impurità totale di un albero H contente un insieme $\tilde{H}$ di nodi terminali come quella quantità *I(H)* tale che:

$$I(H)=\sum_{h\in\tilde{H}}i(h)p(h)=\sum_{h\in\tilde{H}}I(h)$$

in cui $I(h)$ rappresenta l'impurità del nodo h-esimo.

- **passo 3**: per ogni split binario del nodo *h*, l'insieme delle unità ad esso appartenenti viene suddiviso in due gruppi che assumeranno le generiche modalità h_L ed h_R in proporzioni pari a p_L e p_R rispettivamente. In ogni nodo, viene selezionato lo *split* che da luogo al maggior decremento d'impurità, ossia quello *split* che rende massima la quantità:

$$\Delta i(s,h)=i(h)-p_L i(h_L)-p_R i(h_L)$$

3. **definizione della regola di arresto della procedura:** fondamentale per il controllo della dimensione dell'albero finale, viene definita in modo tale che vi sia una sufficiente omogeneità interna fra le unità appartenenti ai nodi terminali dell'albero.

   Alcune delle regole di arresto più diffuse sono quelle per cui un generico nodo *h* diventa terminale se:

   - *la numerosità dello stesso è inferiore ad una certa soglia prefissata*: si fissa una soglia minima per il numero di osservazioni contenute in un nodo *padre* o, eventualmente, nei nodi *figli* da esso generati. Questa regola risponde all'esigenza di ottenere alberi i cui nodi finali non siano espressione di un numero

eccessivamente limitato di osservazioni, o addirittura di un'unica osservazione, risultando, in questo modo, poco informativi;

- *la sua impurità è inferiore ad una certa soglia prefissata*: in tal caso ulteriori partizioni dello stesso nodo non produrranno alcun miglioramento nell'accuratezza della struttura, ma solo una maggiore complessità dell'albero;
- *il massimo incremento di impurità (ottenibile dal migliore split) è inferiore ad una certa soglia prefissata*: in questo modo si vuole porre un freno alle partizioni dell'albero il cui contributo, in termini di purezza (riduzione di impurità), è praticamente nullo;
- *la complessità dell'albero ha superato una certa soglia prefissata*: tale complessità può essere definita in termini di numero di nodi terminali che è anche pari al numero degli *splits* più uno, ovvero in riferimento al numero di livelli dell'albero che ne da una misura della profondità;

4. **assegnazione della modalità di risposta ad ogni nodo:** i nodi terminali ottenuti mediante le procedure di classificazione ad albero possono essere definiti come una partizione del collettivo originario pura al suo interno. Perciò ognuno di tali nodi può essere etichettato attribuendo, alle unità che esso contiene, una classe o un valore della variabile risposta. In tal modo sarà possibile definire, ad esempio, i diversi percorsi che conducono alla medesima classe di risposta.
5. **potatura dell'albero:** operazione che serve, come si è detto, ad ottenere un sottoalbero ottimale che possa essere impiegato a fini decisionali. A tale scopo si definisce una misura che tenga conto del trade-off fra l'aumento dell'impurità dell'albero e la semplificazione della sua struttura. Tale misura, detta *parametro di costo complessità*, è calcolata per ognuno dei nodi interni ed, indicata con α_h, può essere definita come:

$$\alpha_h = \frac{aumento\ \ di\ impurità}{riduzione\ \ della\ \ complessit\ à} = \frac{R(h) - R(H_h)}{\left|\tilde{H}\right| - 1}$$

dove *R* rappresenta il tasso di errata classificazione, ovvero il rapporto fra le unità erroneamente classificate ed il numero totale di quelle considerate, con riferimento sia ad un singolo nodo che ad un intero albero. Ad ogni passo, sostanzialmente, verrà potato il ramo il cui nodo di partenza presenta l'*alfa* minimo dell'intera struttura ad albero. In questo modo sarà possibile, ripetendo iterativamente la procedura, ottenere una sequenza di alberi, via via più piccoli, tutti potenzialmente alberi decisionali, fra i quali, applicando a ciascuno di essi un campione di dati (detto campione test), si potrà scegliere il migliore come quello che ne minimizza il tasso di errata classificazione.

4 Una possibile applicazione dei modelli di classificazione ad albero allo studio della dinamica occupazionale giovanile

Nel presente paragrafo, verrà proposta un'applicazione dei modelli di classificazione ad albero nell'ambito delle dinamiche occupazionali, con particolare riferimento alla situazione giovanile.
Le principali ragioni per cui si propone questo genere di approccio nel presente contesto di studio risiedono nella adattabilità delle variabili che vengono solitamente impiegate nello studio di questa tipologia di fenomeni nonché alle peculiarità dei modelli ad albero così come sono state descritte nel paragrafo precedente.

4.1 Descrizione dei dati e delle variabili

Il database analizzato è stato costruito unificando i 4 database ciascuno relativo ad uno dei trimestri del 2010 della rilevazione continua ISTAT sulle forze di lavoro, nell'ambito del quale sono stati, poi, estrapolati le informazioni relative ai giovani di età compresa fra 15 e 29 anni.
Le variabili incluse nel modello, coerentemente con quanto descritto in fase teorica nel precedente paragrafo, sono:
Variabile dipendente:

Y = *Condizione occupazionale* ("Occupato", "Non occupato")

Variabili dipendenti:

$\mathbf{X_1}$ = *Sesso* ("Maschio", "Femmina")

$\mathbf{X_2}$ = *Titolo di studio* ("Fino alla licenza media", Diploma", "Laurea e oltre")

$\mathbf{X_3}$ = *Zona geografica di residenza* ("Nord-ovest", "Nord-est", "Centro", "Sud", "Isole")

$\mathbf{X_4}$ = *Stato civile* ("Libero", "Sposato", "Separato o vedovo")

Il fenomeno dello scoraggiamento lavorativo, che porta un numero crescente di giovani ad abbandonare le azioni di ricerca di un lavoro qualora esse si rivelino infruttuose per un periodo di tempo ritenuto eccessivamente prolungato[20], cela una effettiva disponibilità a lavorare, manifestandosi, in effetti, come inattività. Per tale ragione si è ritenuto opportuno considerare la dicotomia *occupato – non occupato* quale variabile dipendente accorpando, nella modalità *non occupato*, gli status di disoccupato e di inattivo in età da lavoro, entrambi possibili sintomi, a vario titolo, di una condizione di disagio.
Fra le variabili dipendenti la scelta di escludere l'età è da imputare al fatto che il collettivo oggetto di studio è interamente concentrato all'interno di un intervallo di età già sufficientemente ristretto e tale da rendere irrilevanti, oltre che fuorvianti, ulteriori

[20] L'entità di questo fenomeno è confermata, fra gli altri, anche dagli studi condotti nell'ambito del progetto di ricerca Ipres, *Il mercato del lavoro e i diplomati*, da cui il presente contributo ha origine

disaggregazioni. Il persistere di significative differenze di genere nella situazione occupazionale[21] e la radicata disparità territoriale che contraddistingue il nostro Paese giustificano l'inclusione delle rispettive variabili *sesso* e *zona geografica di residenza*, mentre appare ovvio, oltre che auspicabile, che un più elevato *titolo di studio* muti significativamente il rischio di non occupazione di un giovane, rendendo fondamentale l 'osservazione in uno studio conoscitivo come questo. Le scelte di vita più importanti, infine, rappresentate qui dallo *stato civile*, possono assumere il duplice ruolo di causa-effetto nella determinazione dello status occupazionale di un giovane: se, da una parte, quelli che hanno già formato un proprio nucleo familiare sono maggiormente pressati dall'esigenza di avere un lavoro, di contro, quelli che non riescono a trovarne uno, si trovano nella condizione di non potersi distaccare dal proprio nucleo familiare di origine.

4.2 Descrizione del modello

Il modello è stato implementato impiegando la versione 16 del software statistico SPSS che, al suo interno, contiene la procedura di generazione di modelli di classificazione ad albero.
Volendo, il presente modello, essere utile non solo a fini descrittivi, ma anche e soprattutto a fini previstivi, si è ritenuto opportuno impiegare l'algoritmo di classificazione di tipo CRT (Classification and Regression Tree) che in output fornisce, fra l'altro, il valore di *improvement*, ovvero il decremento di impurità ottenuto ad ogni split. Grazie a ciò è possibile ottenere, quindi, una valutazione del livello discriminante delle caratteristiche implicate in ciascuno split, rispetto al fenomeno analizzato ovvero, in questo caso, il rischio di non occupazione. Sarà possibile, così, individuare le caratteristiche che maggiormente sono in grado di delineare i profili di rischiosità giovanile rispetto allo status di non occupazione, ma anche, vice-versa, quelle che sono in grado di definirne i profili maggiormente virtuosi.

4.3 Presentazione dei risultati

Il modello di classificazione finale, costruito ed opportunamente potato, e la cui rappresentazione grafica è riportata alla pagina seguente in figura 1, è composto complessivamente da 18 nodi figli, oltre al nodo radice (rappresentato con il colore blu), di cui 9 sono nodi terminali (rappresentati con il colore verde o rosso a seconda che la categoria di assegnazione del nodo sia, rispettivamente, "Occupato" o "Non occupato"). In tal senso si può affermare che vengono individuati, sulla base dei predittori impiegati, 9 profili significativi in termini di rischiosità di non occupazione.

[21] IPRES (2010) *1° rapporto sulla condizione femminile in Puglia*, Cacucci Editore, Bari

La variabile che discrimina in maniera più significativa le unità del collettivo preso in esame rispetto al rischio di non occupazione è il titolo di studio. Lo split dell'intero collettivo di riferimento fra "giovani almeno diplomati" e "giovani non diplomati", è infatti quello che produce il maggiore decremento di impurità (*improvement* 0,028) fra quelli possibili con le variabili prese in considerazione, essendo anche, questo split, quello che produce il maggior *improvement* dell'intero albero di classificazione.
Sebbene i giovani appartenenti ad entrambi i gruppi ottenuti vengono classificati, a fini previstivi, come "Non occupati", i primi hanno il 78,3% di possibilità di ricadere nello status di "Non occupazione", mentre i secondi solo il 55,4%.
A questo punto, splittando progressivamente il collettivo è possibile trarre alcune conclusioni interessanti. Ad esempio, mentre per i giovani almeno diplomati la maggiore discriminante per la rischiosità di non occupazione è la zona di residenza, per i non diplomati è il sesso. Il secondo split dell'albero costruito, per livello di *improvement* (0,015) è quello che distingue i diplomati residenti al centro-nord, caratterizzati da una rischiosità inferiore al 50% (46,2%) e perciò da assegnare, ai fini previstivi, alla categoria degli "Occupati", da quelli residenti nel sud e nelle isole che hanno invece ben il 70,1% di probabilità di non essere occupati.
Molto meno "differenziata" è, invece, la situazione dei non diplomati: sia per gli uomini che per le donne (split che produce un *improvement* di 0,004) il rischio di non occupazione è elevatissimo (72,1% e 85,5%, rispettivamente) tuttavia, se se il nodo relativo alla condizione femminile può essere considerato terminale in quanto ulteriori classificazioni non producono variazioni significative, per quel che concerne la condizione maschile, invece, il rischio di non occupazione si riduce drasticamente per coloro che mettono su famiglia (sposati, separati o vedovi, 24,1% di rischio di non occupazione) rispetto a coloro che, invece, restano liberi (75,2%).
Inoltre, mentre per i giovani maschi diplomati residenti nel centro-nord, che hanno un rischio di non occupazione del 41,4% e che sono, quindi, classificabili, come "Occupati", può essere considerata terminale; per le donne, invece, per poter beneficiare di un rischio di non occupazione inferiore al 50% ed essere classificate anch'esse come "Occupate", devono risiedere al nord (47,6%) oppure conseguire un titolo di studio superiore al diploma (48,7%). L'investimento in formazione compiuto per conseguire almeno il diploma produce, invece, un effetto largamente inferiore al sud e nelle isole e in particolare fra le donne: senza particolari distinzioni esse hanno, infatti, il 75,9% di probabilità di non essere occupate, mentre per i maschi la situazione si presenta molto diversificata fra coloro che hanno messo su famiglia (12,2% di rischio di non occupazione) e coloro che sono, invece, liberi (66%).

In sintesi, sulla base delle variabili analizzate nel modello, i profili "virtuosi" individuati, ossia quegli incroci di caratteristiche individuali che permettono, ai giovani che le

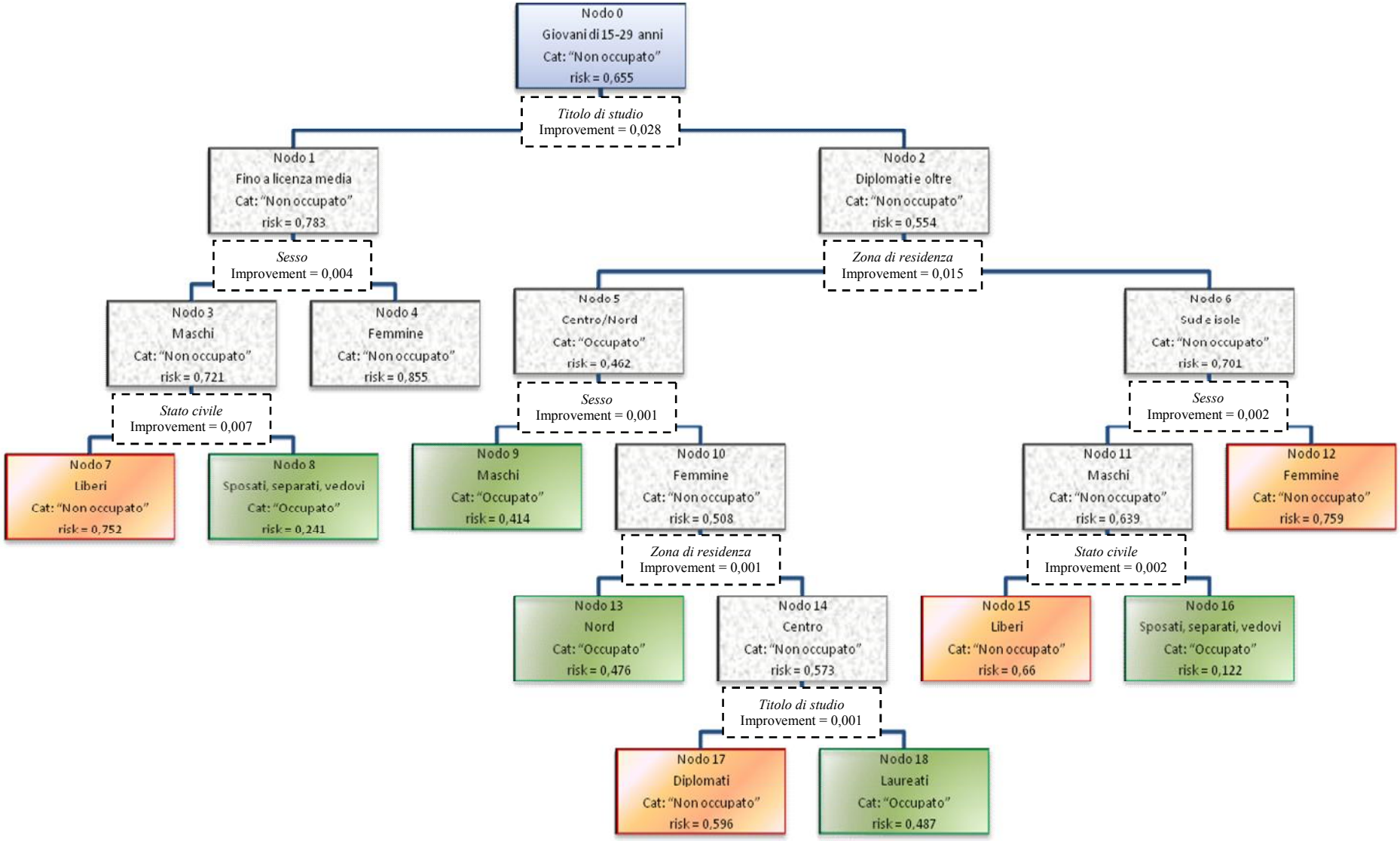
posseggono, di avere una probabilità di essere occupato superiore a quella di non esserlo, possono essere delineati nel modo seguente:

- ✓ maschi, non diplomati, che hanno costituito un proprio nucleo familiare;
- ✓ maschi, almeno diplomati, residenti nel centro-nord;
- ✓ femmine, almeno diplomate, residenti nel nord Italia;
- ✓ femmine, laureate, residenti nel centro Italia.

Vice-versa i profili "critici", ossia quelli per cui, per un giovane, è più probabile essere non occupato piuttosto che occupato, sono i seguenti:

- ✓ maschi, non diplomati, che vivono ancora nel nucleo familiare d'origine;
- ✓ femmine, diplomate, residenti nel centro Italia;
- ✓ maschi, almeno diplomati, residenti nel Mezzogiorno, che vivono ancora nel nucleo familiare d'origine;
- ✓ femmine, almeno diplomate, residenti nel Mezzogiorno.

Figura 8 – Albero di classificazione



4.4 Validazione del modello

Per effettuare una valutazione in merito alla bontà del modello costruito si ricorre al metodo di validazione della *cross-validation*. Ogni unità del collettivo osservato viene riassegnata ad una delle due categorie della variabile risposta del modello ("Occupato" o "Non occupato") impiegando il modello costruito e, conoscendo a priori la reale categoria di appartenenza delle unità analizzate, una misura della bontà del modello in termini di affidabilità ai fini previstivi sarà data dalla quota di unità correttamente riclassificate. Il risultato di tale operazione è riportato nella tabella che segue.

Tabella 2 – Cross-validation sul modello ad albero costruito

Categoria osservata	**Categoria di previsione**		**% di casi osservati**
	Occupato	Non occupato	
Occupato	1.729.150	1.566.387	34,5
Non occupato	1.299.152	4.963.196	65,5
% di casi correttamente classificati	52,5	79,3	**70,0**

Il 70% delle unità del collettivo analizzato (quasi 6,7 milioni di unità) vengono correttamente riclassificate usando il modello di classificazione ad albero costruito. Tale percentuale, inoltre, si dimostra essere più elevata fra i non occupati (79,3%) che non fra gli occupati (52,5%). Tale peculiarità può essere interpretata positivamente per una duplice ragione: non solo, infatti, i non occupati sono molto più numerosi degli occupati nel collettivo osservato (ne costituiscono il 65,5%) e quindi una maggiore capacità di prevederne lo status implica una migliore capacità previsiva del modello nel suo complesso, ma ne rappresentano anche la parte più critica e che maggiormente necessiterebbe di essere correttamente individuata. In altre parole è certamente più importante individuare le cause che contribuiscono a determinare un maggiore "rischio di non occupazione" (rischio elevato) piuttosto che non quelle che contribuiscono a influenzare un "rischio basso".

5 Conclusioni e sviluppi futuri

L'analisi ha mostrato che vi sono delle specificità strutturali per l'Italia con riferimento al mercato del lavoro giovanile: lo status occupazionale dei giovani è molto basso, a fronte di un livello molto elevato di persone in cerca di lavoro e di inattivi in età da lavoro. Per quanto riguarda i giovani, il mercato del lavoro tra le grandi macro aree del Paese differiscono per i

livelli ma non per la dinamica: il tasso di occupazione dei giovani 15-24 anni è più basso nel Mezzogiorno rispetto alle altre due macro aree, ma la dinamica temporale di circa due decenni è molto simile. Inoltre, e la difficoltà di trovare un'occupazione per i giovani viene da lontano, è un elemento strutturale di almeno degli ultimi due decenni; la crisi ne ha solo accentuato l'incidenza e, quindi anche la percezione. Per indagare su queste caratteristiche strutturali abbiamo fatto ricorso all'applicazione di un modello di classificazione e previsione ad albero. L'impiego di questi modelli nel contesto delle dinamiche occupazionali, rappresenta una novità metodologica, visto che, finora, i modelli di classificazione ad albero sono stati impiegati prevalentemente dagli istituti di credito nell'ambito dello studio del *credit-scoring* come strumenti di valutazione del merito di credito dei potenziali fruitori dei finanziamenti.

Notevoli sono gli sviluppi e gli affinamenti metodologici che sarebbe possibile apportare sulla falsa riga del modello proposto in questo lavoro.

La scelta dei predittori, anzitutto, potrebbe essere orientata alla individuazione delle più opportune decisioni di policy del mercato del lavoro da mettere in atto da parte delle autorità competenti. La classificazione delle unità di un collettivo in esame sulla base delle esigenze individuali di servizi pubblici o della percezione del livello qualitativo degli stessi può portare all'individuazione di profili critici per il rischio di non occupazione che possono a loro volta, suggerire l'orientamento delle politiche del mercato del lavoro più opportune da mettere in atto, sulla base delle caratteristiche individuali osservate.

Lo stesso modello, inoltre, potrebbe venire applicato ripetutamente sullo stesso collettivo ad intervalli di tempo costanti, in modo da mettere in atto una sorta di indagine panel volta ad individuare le dinamiche strutturali che definiscono le caratteristiche rilevanti per la determinazione del rischio di non occupazione. E così la "scomparsa" di una certa caratteristica individuale fra le cause maggiormente discriminanti per il rischio di non occupazione, potrebbe rappresentare la "risoluzione di un problema" da parte del sistema economico che produceva un rischio di non occupazione sistematicamente più elevato le persone aventi quella caratteristica, così come, allo stesso modo, però, l'individuazione di una nuova caratteristica discriminante, può segnalare la "nascita di un nuovo problema" suggendo, di conseguenza, la messa in atto delle opportune contromisure.

6 Bibliografia

Aria, M. (2009) *Alberi di Classificazione*, dispense del corso di Analisi Statistica e Sociologica per i processi economici e del lavoro nel settore turistico, Anno accademico 2009-2010, Facoltà di Economia, Università degli Studi di Napoli "Federico II", Federica e-Learning

Bianchi, L., Provenzano G., (2010) *Ma il cielo è sempre più su?*, Castelvecchio Editore, Tazebao, Roma

Boeri, T., Galasso V. (2007) *Conto i Giovani*, Mondadori, Milano

Brieman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984) *Classification and regression trees*, Belmont C.A. Wadsworth

Draghi, M. (2010) *Considerazioni Finali*, Banca D'Italia,

Draghi, M. (2011) *Considerazioni Finali*, Banca D'Italia,

EC *Commission An EU Strategy for Youth – Investing and Empowering,* COM (2009) 200 final;

EUROSTAT (2009) *Youth in Europe*, Bruxelles

ILO (2010) *Global employment trends for Youth*, Geneva

IPRES (2010) *Capitale umano qualificato, mercato del lavoro e mobilità territoriale*, Quaderni IPRES, 2, Cacucci Editore, Bari

IPRES (2010) *1° rapporto sulla condizione femminile in Puglia*, Cacucci Editore, Bari

ISTAT (2009) *Rapporto annuale 2008*, Roma

Leombruni R., Taddei F. (2009) *Giovani precari in una Paese per vecchi*, Il Mulino, 6, 912-920

Manno A. (2002) *Imputazione di dati categoriali mancanti: il modello di classificazione ad albero,* Palermo, Italia, http://www.statistica.too.it

Mola, F., Siciliano, R. (1992) *A two-stage predictive splitting algorithm in binary segmentation*, inY. Dodge, J. Whittaker. (Eds.): Computational Statistics: COMPSTAT 92, 1, Physica Verlag, Heidelberg, 179-184

Mola, F., Siciliano, R. (1997) *A fast splitting procedure for classification trees*, Statistics and Computing 7

Moretti, S., Carmine, P. (2010) *La mobilità del lavoro in Italia: nuove evidenze sulle dinamiche migratorie*, Quaderni di Economia e Finanza, 61, Banca D'Italia

OECD (2010*)* Off *a good start? Jobs for Youth*, Paris;

Rosina, A., Ambrosi, E. (2009) Non è un Paese per giovani, Marsilio;

Saccomanni, F. (2011) *La generazione esclusa: il contributo dei giovani alla crescita economica*, 41° Convegno dei Giovani Imprenditori di Confindustria, Santa Margherita Ligure, 11 giugno, (mimeo)

Schindler M. (2009) *The Italian Labor Market: Recent Trends, Institutions and Reform Options*, IMF, WP 09/47

SVIMEZ (2010) *Rapporto sull'economia del Mezzogiorno. 2009*, Il Mulino, Bologna

ABSTRACT

This work comes from a research project of the Research Institute IPRES (Istituto Pugliese di Ricerche Economiche e Sociali) concerning the inclusion in labour market of not graduated young people (aged from 15 to 29 years old).
First of all, a summary about the matter is provided: we talk about young people and most problems of their inclusion in labour market. Then, in the second part of the paper, we consider the hypothesis to carry out a classification and regression tree model in order to analyze the unemployment trend behind young people. Even if this kind of models has been usually used in evaluating of credit scoring, in fact, we think that an opportune classification of a certain population (for example, in this case, the younger one) may be useful to draft some critical profiles respect of the risk to be not employed.
Such an identification of critical profiles, besides, should allow to carry out the opportune policy instruments in order to reduce, as possible, the respective risks factor.
In the specification of the tree model we propose, four independent variables have been considered (age, gender, marital status and geographic zone of residence) in order to estimate the status of a dichotomous independent variable (employed or not employed). However, obviously, a different set of variables should be used, depending on own respective requirements.